ECONOMETRIA

Ralph dos Santos Silva

http://www.ralphsilva.org/econometria.html

Instituto de Matemática
Centro de Ciências Matemáticas e da Natureza
Universidade Federal do Rio de Janeiro

Programa da disciplina

- Introdução à econometria.
- Modelo de regressão linear.
- Estimador de mínimos quadrados.
- Teste de hipóteses e seleção de modelos.
- Forma funcional e mudança de estrutura.
- Modelos de regressão não-linear, semi-paramétrico e não-paramétrico.
- Endogeneidade e estimação por variável instrumental.
- Modelo de regressão generalizado e heterocedasticidade.
- Sistema de equações.
- Modelos para dados de painel.
- Métodos de estimação em econometria.
- Método generalizado dos momentos.
- Estimação por máxima verossimilhança.
- Estimação e inferência utilizando métodos de simulação.
- Estimação e inferência bayesiana.

Referências

- Bauwens, L., Lubrano, M. e Richard, J-F. (2000). *Bayesian Inference in Dynamic Econometric Models.*
- Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*.
- Greene, W. H. (2011). *Econometric Analysis,* 7th ed.
- Gujarati, D. (2002). *Basic Econometrics*, 4th ed.
- Hayashi, F. (2000). *Econometrics*.
- Hendry, D. F. e Nielsen, B. (2007). *Econometric Modelling: A Likelihood Approach.*
- Koop, G. (2003). *Bayesian Econometrics.*
- Lancaster, T. (2004). *Introduction to Modern Bayesian Econometrics.*
- Stock, J. H. e Watson, M. W. (2006). *Introduction to Econometrics.*
- Wooldridge, J. M. (2006). *Introductory to Econometrics: A Modern Approach.*

Introdução

Econometria é a área que foca em aplicações de estatística matemática e as ferramentas da inferência estatística na mensuração empírica de relações postuladas pela teoria econômica.

Exemplo - Função de Consumo de Keynes

$$C = f(X) \quad \begin{cases} C : \text{consumo} \\ X : \text{renda} \end{cases}$$

Temos

$$0 < \frac{dC}{dX} < 1,$$

a propensão marginal a consumir.

A propensidade média a consumir (PMC), C/X, decai quando a renda sobe

$$\frac{d(C/X)}{dX} = \left(\frac{dC}{dX} - \frac{C}{X}\right)\frac{1}{X} < 0 \quad \Rightarrow \quad \frac{dC}{dX} < \frac{C}{X}.$$

O modelo de função consumo mais simples é o linear (com restrições)

$$C = \alpha + X\beta, \quad \alpha > 0 \quad \text{e} \quad 0 < \beta < 1.$$

- Podemos estudar alguns conjuntos de dados.
- Será verdade que $\alpha > 0$ e $0 < \beta < 1$?
- A relação é estável através do tempo?
- Os parâmetros mudam se considerarmos instantes de tempo diferentes? (Uma mudança na propensidade média a economizar.)
- Existem diferenças sistemáticas nas relações para diferentes países? Como poderíamos explicar estas diferenças?
- Existem outros fatores que poderiam melhorar a habilidade do modelo em explicar a relação entre consumo e renda?

Exemplo

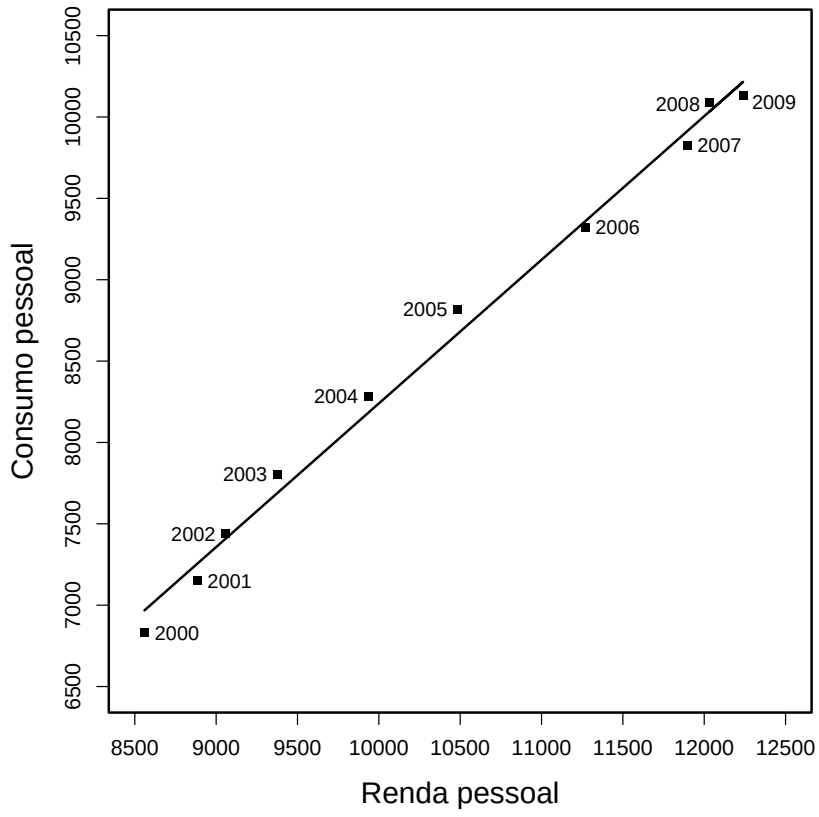


Figura 1: Consumo e renda agregados dos EUA entre 2000 e 2009.

O ajuste por mínimos quadrados resulta em $\hat{\alpha} = -587,19$ e $\hat{\beta} = 0,8827$.

Ver arquivo exemplo_01.r

Para os dados reais, temos uma estimativa negativa para a constante do modelo.

- A teoria está errada?
- Isto se deve a variabilidade dos dados?
- Será que o estimador de mínimos quadrados é adequado nesta situação?
- Como ficaria a estimação por máxima verossimilhança?
- Como ficaria a estimação bayesiana? E a distribuição a priori?

Elasticidade

Em economia, o termo elasticidade se refere a como uma variável econômica muda em função de outra variável econômica.

A elasticidade é medida em termos percentuais ao invés de absoluto.

Isto significa que medimos uma mudança na variável como uma porcentagem da quantidade original da variável.

A mudança em porcentagem da variável x é definida por

$$\text{Mudança percentual em } x = \frac{\text{Mudança em } x}{\text{Valor original de } x} = \frac{\Delta x}{x} \quad (\text{de } x \text{ para } x + \Delta x).$$

A elasticidade de y com respeito a x é dada por

$$\epsilon = \frac{\text{Mudança percentual em } y}{\text{Mudança percentual em } x} = \frac{\Delta y/y}{\Delta x/x}.$$

Em termos infinitesimais, temos que

$$\epsilon = \frac{\partial y/y}{\partial x/x} = \frac{x}{y}\frac{\partial y}{\partial x}.$$

Modelo de regressão linear múltipla

Notação

$$\begin{aligned} y &= f(x_1, x_2, \ldots, x_p) + \varepsilon \\ &= x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p + \varepsilon, \end{aligned}$$

sendo que

- y é a variável dependente;
- $(x_1, x_2, \ldots, x_p)^\top$ é o vetor de variáveis independentes (regressoras, exploratórias ou explicativas);
- $(\beta_1, \beta_2, \ldots, \beta_p)^\top$ é o vetor de coeficientes; e
- ε é um distúrbio aleatório.

Para o modelo de regressão acima, a elasticidade de y com respeito à k-ésima variável x_k é dada por

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

Equação da demanda:

$$\text{quantidade} = \beta_1 + \text{preço} \times \beta_2 + \text{renda} \times \beta_3 + \varepsilon.$$

Equação da demanda inversa:

$$\text{preço} = \gamma_1 + \text{quantidade} \times \gamma_2 + \text{renda} \times \beta_3 + \xi.$$

Estas são relações válidas de um mercado.

Ganhos e educação:

$$\begin{aligned} \text{ganhos} &= \beta_1 + \text{educação} \times \beta_2 + \text{idade} \times \beta_3 + \varepsilon, \\ \text{ganhos} &= \beta_1 + \text{educação} \times \beta_2 + \text{idade} \times \beta_3 + \text{idade}^2 \times \beta_4 + \varepsilon. \end{aligned}$$

Hipóteses do modelo de regressão linear

HP.1: Linearidade: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \varepsilon$;

HP.2: Posto completo: não existe nenhuma relação exata entre as variáveis independentes do modelo;

HP.3: Exogeneidade das variáveis independentes:
$\mathrm{E}(\varepsilon_i | x_{i1}, x_{i2}, \ldots, x_{ip}) = 0 \quad \Rightarrow \quad \mathrm{Cov}(\boldsymbol{x}_i, \varepsilon_i) = 0$
com $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$.

HP.4: Homocedasticidade e autocorrelação nula: cada distúrbio, ε_i, tem a mesma variância, σ^2, e é não correlacionado com qualquer outro distúrbio ε_j;

HP.5: Geração dos dados: os dados em $(x_{i1}, x_{i2}, \ldots, x_{ip})$ podem ser qualquer mistura de constantes e variáveis aleatórias; e

HP.6: Distribuição normal: os distúrbios são normalmente distribuídos.

Notação

Temos que

$$\begin{aligned} y_i &= x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \varepsilon_i \\ &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \text{para} \quad i = 1, 2, \ldots, n; \end{aligned}$$

- n é o número de observações;
- $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ é o vetor da i-ésima observação das variáveis independentes;
- $\boldsymbol{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kn})^\top$ é o vetor da k-ésima variável independente;
- $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ é o vetor da variável dependente;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$ é o vetor dos distúrbios;
- $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ é o vetor de coeficientes; e
- $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$ é a matriz de variáveis independentes.

Assim, o modelo de regressão linear múltipla pode ser reescrito por

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \cdots + \boldsymbol{x}_p\beta_p + \boldsymbol{\varepsilon} \quad \text{ou} \\ \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \end{aligned}$$

Exemplos

- $y = \gamma x^{\beta} e^{\varepsilon} \Rightarrow \ln y = \ln \gamma + \beta \ln x + \varepsilon$ é linear, mas
- $y = \gamma x^{\beta} + \varepsilon$ é não linear.

Os seguintes modelos são lineares:

$$
\begin{aligned}
y &= \alpha + \beta x + \varepsilon \\
y &= \alpha + \beta \cos x + \varepsilon \\
y &= \alpha + \beta \frac{1}{x} + \varepsilon \\
y &= \alpha + \beta \ln x + \varepsilon \\
\ln y &= \alpha + \beta \ln x + \varepsilon \\
z &= \ln \gamma + \beta \ln x + \varepsilon = \alpha + \beta w + \varepsilon.
\end{aligned}
$$

O modelo log-linear

$$\ln y = \beta_1 + \beta_2 \ln x_2 + \cdots + \beta_p \ln x_p + \varepsilon$$

(Elasticidade constante)

A **elasticidade de y com respeito a mudanças em x_k** é dado por

$$\epsilon = \frac{\partial y/y}{\partial x/x} = \frac{\partial y}{\partial x} \times \frac{\partial \ln y/\partial y}{\partial \ln x/\partial x} = \frac{\partial \ln y}{\partial \ln x_k} = \beta_k.$$

Exemplo - mercado de gasolina dos EUA - 1953-2004

Considere o seguinte modelo para o consumo de gasolina per capita:

$$\ln(\text{G/Pop}) = \beta_1 + \beta_2 \ln(\text{Renda/Pop}) + \beta_3 \ln \text{Preço} + \beta_4 \ln(\text{PCN}) + \beta_5 \ln(\text{PCU}) + \varepsilon,$$

sendo PCN o preço dos carros novos e PCU o preço dos carros usados.

- Elasticidade renda e preço para a gasolina; e
- Elasticidade para demanda com respeito a preço de carros novos e usados.

Em geral, utiliza-se um modelo **semi-log** para modelar taxas de crescimento:

$$\ln y_t = \boldsymbol{x}_t^\top \boldsymbol{\beta} + \gamma t + \varepsilon_t \quad \text{com} \quad \partial \ln y_t / \partial t = \gamma,$$

sendo a taxa de crescimento por período.

O modelo translog

- Se $y = g(x_1, x_2, \ldots, x_k)$, então podemos transformar para $\ln y = \ln g(x_1, x_2, \ldots, x_k) = f(\ln x_1, \ln x_2, \ldots, \ln x_k)$, pois $x_k = \exp(\ln x_k)$.
- Daí, aproximar por uma expansão de Taylor de segunda ordem em torno de $\boldsymbol{x} = \mathbf{1}$. Definimos $\ln \boldsymbol{x} = (\ln x_1, \ln x_2, \ldots, \ln x_k)^\top$. Logo,

$$\begin{aligned} \ln y &= f(\mathbf{0}) + \sum_{k=1}^{p} \frac{\partial f(\ln \boldsymbol{x})}{\partial \ln x_k}\bigg|_{\ln \boldsymbol{x}=\mathbf{0}} \ln x_k \\ &+ \frac{1}{2} \sum_{k=1}^{p} \sum_{\ell=1}^{p} \frac{\partial f(\ln \boldsymbol{x})}{\partial \ln x_k \partial \ln x_\ell}\bigg|_{\ln \boldsymbol{x}=\mathbf{0}} \ln x_k \ln x_\ell + \varepsilon. \end{aligned}$$

- Segue-se que

$$\ln y = \beta_0 + \sum_{k=1}^{p} \beta_k \ln x_k + \frac{1}{2} \sum_{k=1}^{p} \sum_{\ell=1}^{p} \gamma_{k\ell} \ln x_k \ln x_\ell + \varepsilon.$$

Distúrbios do modelo de regresssão

Temos:

- $\mathrm{E}(\varepsilon_i|\boldsymbol{X}) = 0$ por hipótese, ou seja,

$$\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \begin{bmatrix} \mathrm{E}(\varepsilon_1|\boldsymbol{X}) \\ \mathrm{E}(\varepsilon_2|\boldsymbol{X}) \\ \vdots \\ \mathrm{E}(\varepsilon_n|\boldsymbol{X}) \end{bmatrix} = \mathbf{0};$$

- $\mathrm{E}(\varepsilon_i|\varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_{i+1}, \ldots, \varepsilon_n) = 0$;
- Por hipótese, todos os distúrbios são simplesmente uma amostra aleatória de alguma população;
- $\mathrm{E}(\varepsilon_i) = \mathrm{E}_{\boldsymbol{X}}\left(\mathrm{E}(\varepsilon_i|\boldsymbol{X})\right) = \mathrm{E}_{\boldsymbol{X}}(0) = 0$;
- Para cada ε_i, $\mathrm{Cov}(\varepsilon_i, \boldsymbol{X}) = \mathrm{Cov}\left(\mathrm{E}(\varepsilon_i|\boldsymbol{X}), \boldsymbol{X}\right) = \mathrm{Cov}(0, \boldsymbol{X}) = \mathbf{0}$;

  Note que $\mathrm{E}(\varepsilon_i) = 0$ **não implica** que $\mathrm{E}(\varepsilon_i|\boldsymbol{x}_i) = 0$. Isto é, $\mathrm{E}(\varepsilon_i) = 0$ **não implica** que $\mathrm{Cov}(0, \boldsymbol{X}) = \mathbf{0}$.
- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \Rightarrow \quad \mathrm{E}(\boldsymbol{y}|\boldsymbol{X}) = \mathrm{E}(\boldsymbol{X}\boldsymbol{\beta}|\boldsymbol{X}) + \mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{0} = \boldsymbol{X}\boldsymbol{\beta}$.

Distúrbios esféricos

Temos:

- $\text{Var}(\varepsilon_i|\boldsymbol{X}) = \sigma^2$, para todo $i = 1, 2, \ldots, n$;
- $\text{Cov}(\varepsilon_i, \varepsilon_j|\boldsymbol{X}) = 0$, para todo $i \neq j$ (correlação nula), ou seja,

$$\begin{aligned}
\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top|\boldsymbol{X}) &= \begin{bmatrix} \text{E}(\varepsilon_1\varepsilon_1|\boldsymbol{X}) & \text{E}(\varepsilon_1\varepsilon_2|\boldsymbol{X}) & \cdots & \text{E}(\varepsilon_1\varepsilon_n|\boldsymbol{X}) \\ \text{E}(\varepsilon_2\varepsilon_1|\boldsymbol{X}) & \text{E}(\varepsilon_2\varepsilon_2|\boldsymbol{X}) & \cdots & \text{E}(\varepsilon_2\varepsilon_n|\boldsymbol{X}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{E}(\varepsilon_n\varepsilon_1|\boldsymbol{X}) & \text{E}(\varepsilon_n\varepsilon_2|\boldsymbol{X}) & \cdots & \text{E}(\varepsilon_n\varepsilon_n|\boldsymbol{X}) \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\boldsymbol{I};
\end{aligned}$$

- $\text{Var}(\boldsymbol{\varepsilon}) = \text{E}_{\boldsymbol{X}}\left(\text{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})\right) + \text{Var}_{\boldsymbol{X}}\left(\text{E}(\boldsymbol{\varepsilon}|\boldsymbol{X})\right) = \sigma^2\boldsymbol{I}$;
- Por hipótese, $(\boldsymbol{\varepsilon}|\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

Exemplo

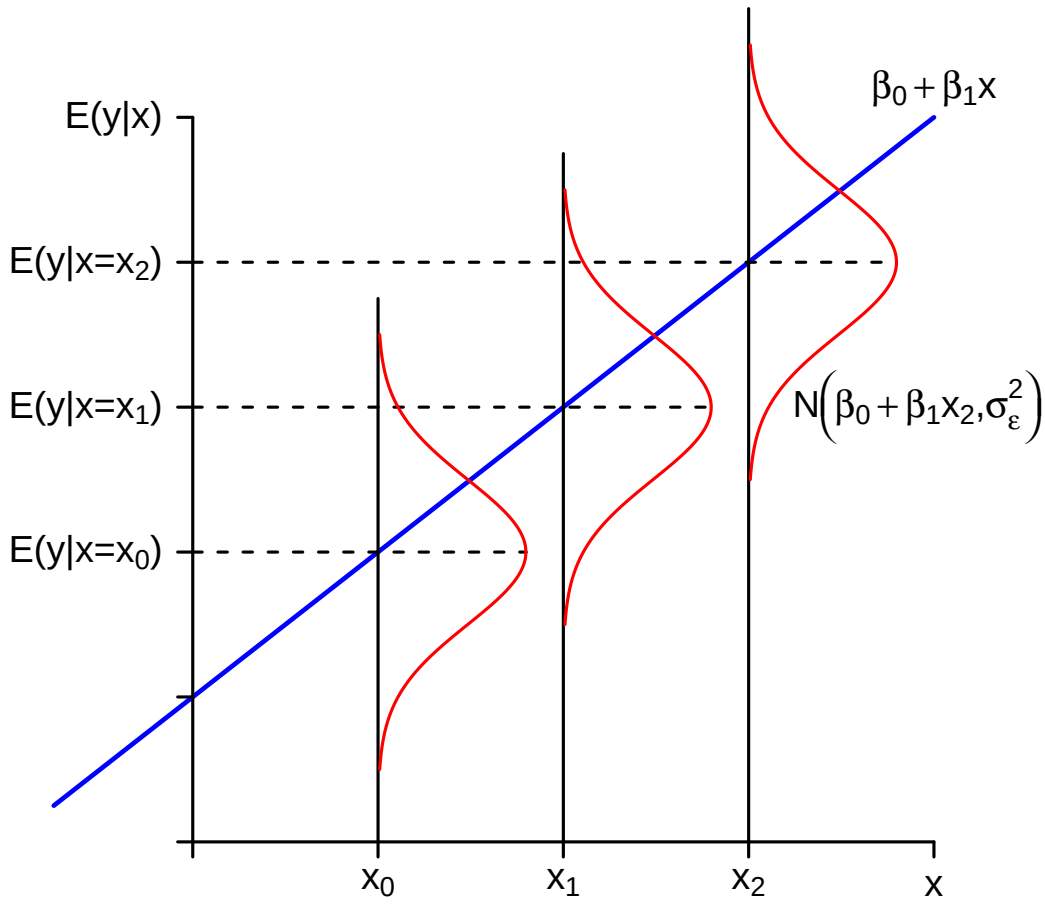


Figura 2: O modelo clássico de regressão.

Ver arquivo exemplo_02.r

Mínimos quadrados

Temos que

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i.$$

- $\boldsymbol{\beta}$ e ε_i são quantidades da população;
- A regressão populacional é $\mathrm{E}(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$;
- $\boldsymbol{b}$ e e_i são estimativas amostrais;
- A estimativa de $\mathrm{E}(y_i|\boldsymbol{x}_i)$ é $\hat{y}_i = \boldsymbol{x}_i^\top \boldsymbol{b}$;
- Seja $\varepsilon_i = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}$ o distúrbio da i-ésima observação;
- Para qualquer valor de $\boldsymbol{b}$, estimamos ε_i com o resíduo

$$e_i = y_i - \boldsymbol{x}_i^\top \boldsymbol{b};$$

- Das definições, $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i = \boldsymbol{x}_i^\top \boldsymbol{b} + e_i$;
- Podemos estimar $\boldsymbol{\beta}$ com a amostra de $(y_i, \boldsymbol{x}_i)$, $i = 1, 2, \ldots, n$;
- Na estimação de $\boldsymbol{\beta}$ por $\boldsymbol{b}$, queremos que os pontos observados fiquem perto da reta ajustada;
- Perto neste caso é definido por algum critério de ajuste;
- **Critério de mínimos quadrados.**

Temos

$$S(\boldsymbol{b}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{b})^2,$$

com $\boldsymbol{b}$ denotando a escolha do vetor de coeficientes.

Queremos minimizar $S(\boldsymbol{b})$, ou seja,

$$\min_{\boldsymbol{b}} S(\boldsymbol{b}) = \min_{\boldsymbol{b}} \boldsymbol{e}^\top \boldsymbol{e} = \min_{\boldsymbol{b}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}).$$

Mas

$$\begin{aligned} S(\boldsymbol{b}) &= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{b} + \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{b} \\ &= \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{b} + \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{b}. \end{aligned}$$

A condição necessária para um mínimo é

$$\frac{\partial S(\boldsymbol{b})}{\partial \boldsymbol{b}} = -2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}.$$

Seja $\boldsymbol{b}$ a solução. Então, $\boldsymbol{b}$ satisfaz as equações normais de mínimos quadrados:

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^\top \boldsymbol{y}.$$

Se a inversa de $\boldsymbol{X}^\top \boldsymbol{X}$ existir, que segue pela hipótese de posto completo de $\boldsymbol{X}$, então a solução é

$$\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Para esta solução minimizar a soma de quadrados,

$$\frac{\partial^2 S(\boldsymbol{b})}{\partial \boldsymbol{b} \partial \boldsymbol{b}^\top} = 2\boldsymbol{X}^\top \boldsymbol{X}$$

deve ser uma matriz positiva definida.

Seja $q = \boldsymbol{c}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{c}$ para algum vetor arbitrário $\boldsymbol{c}$ diferente de zero. Então,

$$q = \boldsymbol{v}^\top \boldsymbol{v} = \sum_{i=1}^{n} v_i^2 \quad \text{sendo} \quad \boldsymbol{v} = \boldsymbol{X}\boldsymbol{c}.$$

A menos que cada elemento de $\boldsymbol{v}$ seja zero, q é positivo.

Exemplo

$$y_i = \beta_1 + \beta_2 T_i + \beta_3 G_i + \beta_4 R_i + \beta_5 P_i + \varepsilon_i$$

Tabela 1: Observações anuais de 1968 a 1982.

Investimento Real ($\boldsymbol{Y}$)	Constante ($\mathbf{1}$)	Tendência ($\boldsymbol{T}$)	PIB Real ($\boldsymbol{G}$)	Taxa de Juros ($\boldsymbol{R}$)	Taxa de Inflação ($\boldsymbol{P}$)
0,161	1	1	1,058	5,16	4,40
0,172	1	2	1,088	5,87	5,15
0,158	1	3	1,086	5,95	5,37
0,173	1	4	1,122	4,88	4,99
0,195	1	5	1,186	4,50	4,16
0,217	1	6	1,254	6,44	5,75
0,199	1	7	1,246	7,83	8,82
0,163	1	8	1,232	6,25	9,31
0,195	1	9	1,298	5,50	5,21
0,231	1	10	1,370	5,46	5,83
0,257	1	11	1,439	7,46	7,40
0,259	1	12	1,479	10,28	8,64
0,225	1	13	1,474	11,77	9,31
0,241	1	14	1,503	13,42	9,44
0,204	1	15	1,475	11,02	5,99

Ver arquivo exemplo_03.r

Teorema (regressão ortogonal)

Se as variáveis em um modelo de regressão são não correlacionadas, isto é, são ortogonais, então os coeficientes angulares da regressão são os mesmos que os coeficientes na regressão individual simples.

Aspectos algébricos da solução de mínimos quadrados

- As equações normais são

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{b} - \boldsymbol{X}^\top \boldsymbol{y} = -\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = -\boldsymbol{X}^\top \boldsymbol{e} = \boldsymbol{0}.$$

- Para cada coluna $\boldsymbol{x}_k$ de $\boldsymbol{X}$, temos $\boldsymbol{x}_k^\top \boldsymbol{e} = 0$.
- Se a primeira coluna de $\boldsymbol{X}$ for de 1's, então
  1. os resíduos de mínimos quadrados somam zero,

$$\boldsymbol{x}_1^\top \boldsymbol{e} = \boldsymbol{1}^\top \boldsymbol{e} = \sum_{i=1}^{n} e_i = 0;$$

  2. a regressão no hiperplano passa pelo ponto das médias dos dados; a primeira equação normal implica que $\overline{y} = \overline{\boldsymbol{x}}^\top \boldsymbol{b}$;
  3. a média dos valores ajustados da regressão é igual a média dos valores atuais.

Isto só vale se a regressão tiver uma constante.

Projeção

O valor dos resíduos de mínimos quadrados é dado por

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \\ &= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \\ &= \boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} \\ &= \boldsymbol{I}\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} \\ &= (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} \\ &= \boldsymbol{M}\boldsymbol{y}, \end{aligned}$$

sendo $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ e $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P}$.

Temos que $\boldsymbol{P}$ é a matriz de projeção (leva $\boldsymbol{y}$ em $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}$):

$$\boldsymbol{P} = \boldsymbol{P}^\top, \quad \text{e} \quad \boldsymbol{P} = \boldsymbol{P}^2.$$

E $\boldsymbol{M}$ é uma matriz importante pois leva $\boldsymbol{y}$ em $\boldsymbol{e}$:

$$\boldsymbol{M} = \boldsymbol{M}^\top, \quad \boldsymbol{M} = \boldsymbol{M}^2, \quad \text{e} \quad \boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}.$$

Toda matriz simétrica e idempotente é positiva definida, em particular $\boldsymbol{P}$ e $\boldsymbol{M}$.

A matriz de projeção

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \quad \Leftrightarrow \quad \boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e} \quad \Rightarrow$$
$$\hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}.$$

$$\boldsymbol{P}\boldsymbol{M} = \boldsymbol{M}\boldsymbol{P} = \boldsymbol{0}, \quad \text{e} \quad \boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}.$$

$$\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y} = \text{projeção} + \text{resíduos}.$$

Podemos reescrever a soma de quadrados da seguinte forma:

$$\begin{aligned} \boldsymbol{y}^\top \boldsymbol{y} &= \boldsymbol{y}^\top \boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{M}^\top \boldsymbol{M}\boldsymbol{y} \\ &= \hat{\boldsymbol{y}}^\top \hat{\boldsymbol{y}} + \boldsymbol{e}^\top \boldsymbol{e}. \end{aligned}$$

Outras relações úteis:

$$\begin{aligned} \boldsymbol{e}^\top \boldsymbol{e} &= \boldsymbol{y}^\top \boldsymbol{M}^\top \boldsymbol{M}\boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{M}\boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{e} = \boldsymbol{e}^\top \boldsymbol{y} \\ \boldsymbol{e}^\top \boldsymbol{e} &= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{b} = \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{b}. \end{aligned}$$

Regressão particionada e regressão parcial

Suponha que a regressão envolva dois conjuntos de variáveis $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$. Logo,

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \varepsilon.$$

Qual é a solução algébrica para $\boldsymbol{b}_2$?

$$\begin{bmatrix} \boldsymbol{X}_1^\top \boldsymbol{X}_1 & \boldsymbol{X}_1^\top \boldsymbol{X}_2 \\ \boldsymbol{X}_2^\top \boldsymbol{X}_1 & \boldsymbol{X}_2^\top \boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1^\top \boldsymbol{y} \\ \boldsymbol{X}_2^\top \boldsymbol{y} \end{bmatrix}$$

Resolvendo para $\boldsymbol{b}_1$ (primeira equação):

$$\boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{b}_1 + \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{b}_2 = \boldsymbol{X}_1^\top \boldsymbol{y}$$
$$(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{b}_1 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{b}_2 = (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{y}$$

$$\begin{aligned} \boldsymbol{b}_1 &= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{y} - (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{b}_2 \\ &= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top (\boldsymbol{y} - \boldsymbol{X}_2 \boldsymbol{b}_2). \end{aligned}$$

$\boldsymbol{b}_1$ é o conjunto de coeficientes na regressão de $\boldsymbol{y}$ sobre $\boldsymbol{X}_1$, menos um vetor de correção. Suponha que $\boldsymbol{X}_1^\top \boldsymbol{X}_2 = \boldsymbol{0}$, então $\boldsymbol{b}_1 = (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{y}$.

Na solução de $\boldsymbol{b}_2$ temos

$$
\begin{aligned}
&(\boldsymbol{X}_2^\top \boldsymbol{X}_1)\boldsymbol{b}_1 + (\boldsymbol{X}_2^\top \boldsymbol{X}_2)\boldsymbol{b}_2 = \boldsymbol{X}_2^\top \boldsymbol{y} \\
&(\boldsymbol{X}_2^\top \boldsymbol{X}_1)(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top(\boldsymbol{y} - \boldsymbol{X}_2\boldsymbol{b}_2) + (\boldsymbol{X}_2^\top \boldsymbol{X}_2)\boldsymbol{b}_2 = \boldsymbol{X}_2^\top \boldsymbol{y} \\
&\boldsymbol{X}_2^\top \boldsymbol{P}_1\boldsymbol{y} - \boldsymbol{X}_2^\top \boldsymbol{P}_1\boldsymbol{X}_2\boldsymbol{b}_2 + (\boldsymbol{X}_2^\top \boldsymbol{X}_2)\boldsymbol{b}_2 = \boldsymbol{X}_2^\top \boldsymbol{y} \\
&\boldsymbol{X}_2^\top (\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{X}_2\boldsymbol{b}_2 = \boldsymbol{X}_2^\top (\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y} \\
&\boldsymbol{X}_2^\top \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{b}_2 = \boldsymbol{X}_2^\top \boldsymbol{M}_1\boldsymbol{y} \\
&\boldsymbol{b}_2 = (\boldsymbol{X}_2^\top \boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top \boldsymbol{M}_1\boldsymbol{y} = (\boldsymbol{X}_2^{\star\top} \boldsymbol{X}_2^\star)^{-1}\boldsymbol{X}_2^{\star\top} \boldsymbol{y}^\star,
\end{aligned}
$$

sendo $\boldsymbol{X}_2^\star = \boldsymbol{M}_1\boldsymbol{X}_2$ e $\boldsymbol{y}^\star = \boldsymbol{M}_1\boldsymbol{y}$.

$\boldsymbol{M}_1$ é a matriz construtora de resíduos definida para a regressão sobre as colunas de $\boldsymbol{X}_1$.

Logo, $\boldsymbol{M}_1\boldsymbol{X}_2$ é uma matriz de resíduos.

Cada coluna de $\boldsymbol{M}_1\boldsymbol{X}_2$ é um vetor de resíduos da regressão correspondente da coluna $\boldsymbol{X}_2$ sobre as variáveis em $\boldsymbol{X}_1$.

Teorema (Frisch-Waugh, 1933; e Lovell, 1963)

Na regressão linear de mínimos quadrados do vetor $\boldsymbol{y}$ sobre dois conjuntos de variáveis $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$, o subvetor $\boldsymbol{b}_2$ é o conjunto de coeficientes obtidos quando os resíduos da regressão de $\boldsymbol{y}$ sobre $\boldsymbol{X}_1$ sozinho são regredidos sobre o conjunto de resíduos obtidos quando cada coluna de $\boldsymbol{X}_2$ é regredida sobre $\boldsymbol{X}_1$.

Considere uma regressão de $\boldsymbol{y}$ sobre um conjunto de variáveis $\boldsymbol{X}$ e uma variável adicional $\boldsymbol{z}$, e seus respectivos coeficientes por $\boldsymbol{b}$ e c.

Corolário

O coeficiente sobre $\boldsymbol{z}$ em uma regressão múltipla de $\boldsymbol{y}$ sobre $\boldsymbol{W} = [\boldsymbol{X}, \boldsymbol{z}]$ é calculada como

$$c = (\boldsymbol{z}^\top \boldsymbol{M}\boldsymbol{z})^{-1}\boldsymbol{z}^\top \boldsymbol{M}\boldsymbol{y} = (\boldsymbol{z}^{\star\top}\boldsymbol{z}^\star)^{-1}\boldsymbol{z}^\star\boldsymbol{y}^\star,$$

sendo $\boldsymbol{z}^\star = \boldsymbol{M}\boldsymbol{z}$ e $\boldsymbol{y}^\star = \boldsymbol{M}\boldsymbol{y}$ os vetores de resíduos da regressão de mínimos quadrados de $\boldsymbol{z}$ e $\boldsymbol{y}$ sobre $\boldsymbol{X}$.

Lembrando que $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

Coeficiente de correlação parcial

Seja a regressão $y = \beta_1 + \beta_2 z + \beta_3 x + \varepsilon$. Além disso,

(a) $y_i^\star$ os resíduos da regressão de y sobre x, isto é, $y_i^\star = y_i - a - bx_i$; e

(b) $z_i^\star$ os resíduos da regressão de z sobre x, isto é, $z_i^\star = z_i - c - dx_i$,

para $i = 1, 2, \ldots, n$.

Então, a correlação parcial $r_{yz}^\star$ é a correlação simples entre $y^\star$ e $z^\star$, isto é,

$$r_{yz}^\star = \frac{\boldsymbol{z}^{\star\top}\boldsymbol{y}^\star}{\sqrt{(\boldsymbol{z}^{\star\top}\boldsymbol{z}^\star)(\boldsymbol{y}^{\star\top}\boldsymbol{y}^\star)}}.$$

pois o modelo de regressão tem a constante β_1. Daí, $\mathbf{1}^\top \boldsymbol{z}^\star = \mathbf{1}^\top \boldsymbol{y}^\star = 0$.

Notemos que $\boldsymbol{y}^\star = \boldsymbol{M}\boldsymbol{y}$ e $\boldsymbol{z}^\star = \boldsymbol{M}\boldsymbol{z}$ com $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

Soma de quadrados

Considere agora os resíduos $\boldsymbol{u}$ da regressão: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{d} + \boldsymbol{z}c + \boldsymbol{u}$.

A menos que $\boldsymbol{X}^\top\boldsymbol{z} = \mathbf{0}$, $\boldsymbol{d}$ não será igual a $\boldsymbol{b} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$.

A menos que $c = 0$, $\boldsymbol{u}$ não será igual a $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}$.

Notemos que

$$\begin{aligned} c &= (\boldsymbol{z}^{\star\top}\boldsymbol{z}^{\star})^{-1}\boldsymbol{z}^{\star\top}\boldsymbol{y}^{\star}, \quad \text{e} \\ \boldsymbol{d} &= (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}(\boldsymbol{y}-\boldsymbol{z}c) = \boldsymbol{b} - (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{z}c. \end{aligned}$$

Segue-se que

$$\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} + \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{z}c - \boldsymbol{z}c = \boldsymbol{e} - \boldsymbol{M}\boldsymbol{z}c = \boldsymbol{e} - \boldsymbol{z}^{\star}c.$$

Portanto,

$$\boldsymbol{u}^{\top}\boldsymbol{u} = \boldsymbol{e}^{\top}\boldsymbol{e} + c^2(\boldsymbol{z}^{\star\top}\boldsymbol{z}^{\star}) - 2c\boldsymbol{z}^{\star\top}\boldsymbol{e}.$$

Mas

$$\begin{aligned} &\boldsymbol{e} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{y}^{\star} \\ &\boldsymbol{z}^{\star\top}\boldsymbol{e} = \boldsymbol{z}^{\star\top}\boldsymbol{M}\boldsymbol{y} = \boldsymbol{z}^{\top}\boldsymbol{M}^{\top}\boldsymbol{M}\boldsymbol{y} = \boldsymbol{z}^{\top}\boldsymbol{M}^{\top}\boldsymbol{y} = \boldsymbol{z}^{\star\top}\boldsymbol{y}^{\star} \\ &\boldsymbol{z}^{\star\top}\boldsymbol{e} = \boldsymbol{z}^{\star\top}\boldsymbol{y}^{\star} = c(\boldsymbol{z}^{\star\top}\boldsymbol{z}^{\star}), \end{aligned}$$

pois pela equação normal $(\boldsymbol{z}^{\star\top}\boldsymbol{z}^{\star})c = \boldsymbol{z}^{\star\top}\boldsymbol{y}^{\star}$.

Assim, $\boldsymbol{u}^{\top}\boldsymbol{u} = \boldsymbol{e}^{\top}\boldsymbol{e} - c^2(\boldsymbol{z}^{\star\top}\boldsymbol{z}^{\star})$.

Teorema (mudança na soma de quadrados quando uma variável é adicionada ao modelo)

Se $\boldsymbol{e}^\top \boldsymbol{e}$ é a soma de quadrados dos resíduos quando $\boldsymbol{y}$ é regredido sobre $\boldsymbol{X}$ e $\boldsymbol{u}^\top \boldsymbol{u}$ é a soma de quadrados dos resíduos quando $\boldsymbol{y}$ é regredido sobre $\boldsymbol{X}$ e $\boldsymbol{z}$, então

$$\boldsymbol{u}^\top \boldsymbol{u} = \boldsymbol{e}^\top \boldsymbol{e} - c^2(\boldsymbol{z}^{\star\top} \boldsymbol{z}^\star) \leqslant \boldsymbol{e}^\top \boldsymbol{e},$$

sendo c o coeficiente de $\boldsymbol{z}$ na regressão completa e $\boldsymbol{z}^\star = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top]\boldsymbol{z}$ o vetor de resíduos quando $\boldsymbol{z}$ é regredido sobre $\boldsymbol{X}$.

Temos que $\boldsymbol{e}^\top \boldsymbol{e} = \boldsymbol{y}^{\star\top} \boldsymbol{y}^\star$ e pela equação normal

$$c(\boldsymbol{z}^{\star\top} \boldsymbol{z}^\star) = \boldsymbol{z}^{\star\top} \boldsymbol{y}^\star \quad \Rightarrow \quad c^2(\boldsymbol{z}^{\star\top} \boldsymbol{z}^\star) = (\boldsymbol{z}^{\star\top} \boldsymbol{y}^\star)^2/(\boldsymbol{z}^{\star\top} \boldsymbol{z}^\star).$$

Portanto,

$$\frac{\boldsymbol{u}^\top \boldsymbol{u}}{\boldsymbol{y}^{\star\top} \boldsymbol{y}^\star} = \frac{\boldsymbol{y}^{\star\top} \boldsymbol{y}^\star - (\boldsymbol{z}^{\star\top} \boldsymbol{y}^\star)^2/\boldsymbol{z}^{\star\top} \boldsymbol{z}^\star}{\boldsymbol{y}^{\star\top} \boldsymbol{y}^\star} = 1 - r^{\star 2}_{yz}.$$

Exemplo (continuação)

$$y_i = \beta_1 + \beta_2 T_i + \beta_3 G_i + \beta_3 R_i + \beta_4 P_i + \varepsilon_i$$

Seguimos a análise dos dados da equação de investimento.

Calculamos a correlação (simples) entre a variável investimento e as variáveis de tendência temporal (T), produto interno bruto (G), taxa de juros (R) e taxa de inflação (P).

Calculamos também a correlação parcial entre investimento e as quatro regressoras dado as demais regressoras.

Tabela 2: Correlação entre investimento e as demais variáveis.

	Correlação Simples	Correlação Parcial
Tendência	0,7496	-0,9360
PIB	0,8632	0,9680
Taxa de juros	0,5871	-0,5167
Taxa de inflação	0,4777	-0,0221

Ver arquivo exemplo_04.r

Qualidade de ajuste e análise de variância

Podemos perguntar se a variação em $\boldsymbol{x}$ é um bom preditor da variação em y.

A variação total é dada por:

$$SQTot = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

Já sabemos que $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} = \hat{\boldsymbol{y}} + \boldsymbol{e}$. Logo, para uma única observação, temos que

$$\begin{aligned} y_i &= \widehat{y}_i + e_i = \boldsymbol{x}_i^\top \boldsymbol{b} + e_i \\ y_i - \overline{y} &= \widehat{y}_i - \overline{y} + e_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top \boldsymbol{b} + e_i. \end{aligned}$$

Exemplo

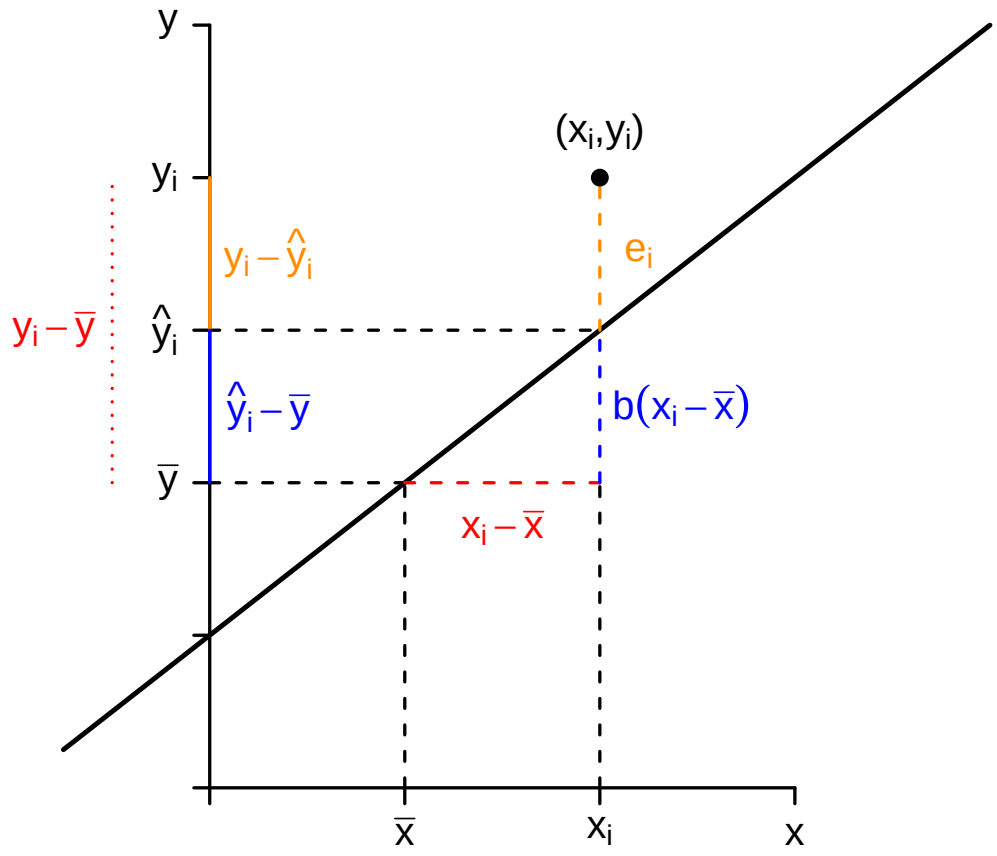


Figura 3: Decomposição de *y*.

Ver arquivo exemplo_05.r

Algumas propriedades, notações e somas de quadrados

$$\mathbf{1}\overline{x} = \mathbf{1}\frac{1}{n}\mathbf{1}^\top \boldsymbol{x} = \begin{bmatrix} \overline{x} \\ \overline{x} \\ \vdots \\ \overline{x} \end{bmatrix} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top \boldsymbol{x}$$

$$\begin{bmatrix} x_1 - \overline{x} \\ x_2 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{bmatrix} = [\boldsymbol{x} - \mathbf{1}\overline{x}] = [\boldsymbol{x} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \boldsymbol{x}] = [\boldsymbol{I}\boldsymbol{x} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \boldsymbol{x}] = [\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top]\boldsymbol{x} = \widetilde{\boldsymbol{M}}\boldsymbol{x}.$$

Definimos $\widetilde{\boldsymbol{M}} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Esta também é uma matriz idempotente.

Temos que $\widetilde{\boldsymbol{M}}\mathbf{1} = \mathbf{0} \Leftrightarrow \mathbf{1}^\top \widetilde{\boldsymbol{M}} = \mathbf{0}^\top$. Logo,

$$\sum_{i=1}^{n} (x_i - \overline{x}) = \mathbf{1}^\top [\widetilde{\boldsymbol{M}}\boldsymbol{x}] = \mathbf{0}^\top \boldsymbol{x} = 0.$$

Além disso,

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = (\boldsymbol{x} - \mathbf{1}\overline{x})^\top (\boldsymbol{x} - \mathbf{1}\overline{x}) = (\widetilde{\boldsymbol{M}}\boldsymbol{x})^\top (\widetilde{\boldsymbol{M}}\boldsymbol{x}) = \boldsymbol{x}^\top \widetilde{\boldsymbol{M}}^\top \widetilde{\boldsymbol{M}}\boldsymbol{x} = \boldsymbol{x}^\top \widetilde{\boldsymbol{M}}\boldsymbol{x}$$

$$\begin{aligned} \widetilde{\boldsymbol{M}}\boldsymbol{y} &= \widetilde{\boldsymbol{M}}\boldsymbol{X}\boldsymbol{b} + \widetilde{\boldsymbol{M}}\boldsymbol{e} \\ \widetilde{\boldsymbol{M}}\boldsymbol{e} &= \boldsymbol{e} \qquad \text{e} \qquad \boldsymbol{e}^\top \widetilde{\boldsymbol{M}}\boldsymbol{X} = \boldsymbol{e}^\top \boldsymbol{X} = \boldsymbol{0}^\top. \end{aligned}$$

Finalmente,

$$\underbrace{\boldsymbol{y}^\top \widetilde{\boldsymbol{M}}\boldsymbol{y}}_{SQTot} = \underbrace{\boldsymbol{b}^\top \boldsymbol{X}^\top \widetilde{\boldsymbol{M}}\boldsymbol{X}\boldsymbol{b}}_{SQReg} + \underbrace{\boldsymbol{e}^\top \boldsymbol{e}}_{SQRes}.$$

- SQTot é a soma de quadrados total;
- SQReg é a soma de quadrados da regressão; e
- SQRes é a soma de quadrados dos resíduos.

Coeficiente de determinação

$$R^2 = \frac{SQReg}{SQTot} = \frac{\boldsymbol{b}^\top \boldsymbol{X}^\top \widetilde{\boldsymbol{M}}\boldsymbol{X}\boldsymbol{b}}{\boldsymbol{y}^\top \widetilde{\boldsymbol{M}}\boldsymbol{y}} = 1 - \frac{\boldsymbol{e}^\top \boldsymbol{e}}{\boldsymbol{y}^\top \widetilde{\boldsymbol{M}}\boldsymbol{y}}.$$

Tabela 3: Análise de variância.

Soma de Quadrados	Fonte	Graus de Liberdade	Média Quadrática
Regressão	$\boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{y} - n\overline{y}^2$	$p-1$	
Resíduo	$\boldsymbol{e}^\top \boldsymbol{e}$	$n-p$	s^2
Total	$\boldsymbol{y}^\top \boldsymbol{y} - n\overline{y}^2$	$n-1$	$S_{yy}/(n-1) = s_y^2$

$$
\begin{aligned}
\boldsymbol{b}^\top \boldsymbol{X}^\top \widetilde{\boldsymbol{M}} \boldsymbol{X} \boldsymbol{b} &= \boldsymbol{b}^\top \boldsymbol{X}^\top \widetilde{\boldsymbol{M}} [\boldsymbol{y} - \boldsymbol{e}] = \boldsymbol{b}^\top \boldsymbol{X}^\top \widetilde{\boldsymbol{M}} \boldsymbol{y} - \boldsymbol{b}^\top \boldsymbol{X}^\top \widetilde{\boldsymbol{M}} \boldsymbol{e} \\
&= \boldsymbol{b}^\top \boldsymbol{X}^\top [\boldsymbol{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^\top] \boldsymbol{y} \qquad \text{(porque} \quad \boldsymbol{X}^\top \widetilde{\boldsymbol{M}} \boldsymbol{e} = \mathbf{0}) \\
&= \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{y} - \frac{1}{n} \boldsymbol{b}^\top \boldsymbol{X}^\top \mathbf{1}\mathbf{1}^\top \boldsymbol{y} = \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{y} - \frac{1}{n} \hat{\boldsymbol{y}}^\top \mathbf{1}\mathbf{1}^\top \boldsymbol{y} \\
&= \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{y} - \frac{1}{n} (\boldsymbol{y} - \boldsymbol{e}) \mathbf{1}\mathbf{1}^\top \boldsymbol{y} \\
&= \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{y} - \frac{1}{n} \left[\sum_{i=1}^n y_i - \sum_{i=1}^n e_i \right] \sum_{i=1}^n y_i \\
&= \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{y} - n\overline{y}^2.
\end{aligned}
$$

Exemplo

$$y_i = \beta_1 + \beta_2 T_i + \beta_3 G_i + \beta_3 R_i + \beta_4 P_i + \varepsilon_i$$

Tabela 4: Análise de variância - equação de investimento.

Soma de Quadrados	Fonte	Graus de Liberdade	Média Quadrática	R^2
Regressão	0,0159025	4	0,00397563	0,9724
Resíduo	0,0004508	10	0,00004508	
Total	0,0163533	14	0,00116810	

Ver arquivo exemplo_06.r

Teorema (Mudança no R^2 quando uma variável é adicionada ao modelo)

Seja $R^2_{\boldsymbol{Xz}}$ o coeficiente de determinação na regressão de $\boldsymbol{y}$ sobre $\boldsymbol{X}$ e uma variável adicional $\boldsymbol{z}$ ($\boldsymbol{u}$ os resíduos), seja $R^2_{\boldsymbol{X}}$ o mesmo para a regressão de $\boldsymbol{y}$ sobre $\boldsymbol{X}$ sozinha ($\boldsymbol{e}$ os resíduos), e seja $r^{\star 2}_{yz}$ o coeficiente de correlação parcial entre y e z. Então,

$$R^2_{\boldsymbol{Xz}} = R^2_{\boldsymbol{X}} + (1 - R^2_{\boldsymbol{X}}) r^{\star 2}_{yz} \geqslant R^2_{\boldsymbol{X}}.$$

Prova: Vimos que $\boldsymbol{z}^\star = \boldsymbol{Mz}$, $\boldsymbol{y}^\star = \boldsymbol{My} = \boldsymbol{e} \Rightarrow \boldsymbol{e}^\top \boldsymbol{e} = \boldsymbol{y}^{\star\top} \boldsymbol{y}^\star$, e

$$\boldsymbol{u}^\top \boldsymbol{u} = \boldsymbol{e}^\top \boldsymbol{e} - c^2(\boldsymbol{z}^{\star\top} \boldsymbol{z}^\star) = \boldsymbol{y}^{\star\top} \boldsymbol{y}^\star - \frac{(\boldsymbol{z}^{\star\top} \boldsymbol{y}^\star)^2}{\boldsymbol{z}^{\star\top} \boldsymbol{z}^\star} = \boldsymbol{e}^\top \boldsymbol{e}(1 - r^{\star 2}_{yz}).$$

Logo,

$$R^2_{\boldsymbol{Xz}} = 1 - \frac{\boldsymbol{u}^\top \boldsymbol{u}}{\boldsymbol{y}^\top \widetilde{\boldsymbol{M}} \boldsymbol{y}} = 1 - \frac{\boldsymbol{e}^\top \boldsymbol{e}}{\boldsymbol{y}^\top \widetilde{\boldsymbol{M}} \boldsymbol{y}} + \frac{\boldsymbol{e}^\top \boldsymbol{e}}{\boldsymbol{y}^\top \widetilde{\boldsymbol{M}} \boldsymbol{y}} r^{\star 2}_{yz} = R^2_{\boldsymbol{X}} + (1 - R^2_{\boldsymbol{X}}) r^{\star 2}_{yz} \geqslant R^2_{\boldsymbol{X}}.$$

Exemplo - ajuste de uma função consumo

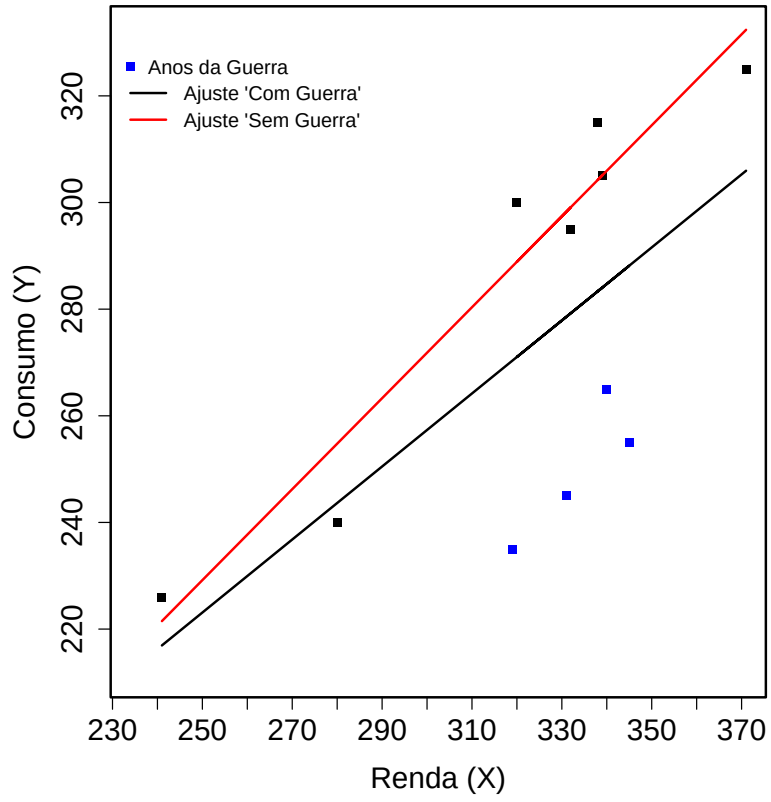


Figura 4: Dados de consumo e renda de 1940-1950 dos EUA. Fonte: *Economic Report of the President, U.S. Government Printing Office, Washington, D.C., 1983.*

Ver arquivo exemplo_07.r

Exemplo (continuação) - ajuste de uma função consumo

Tabela 5: Análise de variância: consumo e renda.

Soma de Quadrados	Fonte	Graus de Liberdade	Média Quadrática	R^2
Regressão	5768,2068	1	5768,2068	0,4571
Resíduo	6849,9751	9	761,1083	
Total	12618,1818	10	1261,8182	

Tabela 6: Análise de variância: consumo, renda e dicotômica para anos de guerra.

Soma de Quadrados	Fonte	Graus de Liberdade	Média Quadrática	R^2
Regressão x	5768,2068	1	5768,2068	
Regressão w	6173,5189	1	6173,5189	
Regressão	11941,7257	2	5970,8629	0,9464
Resíduo	676,4562	8	84,5570	
Total	12618,1818	10	1261,8,182	

O R^2 ajustado para os graus de liberdade

R^2 nunca decresce quando uma nova variável regressora é adicionado a equação de regressão. Portanto, é tentador adicionar diversas variáveis regressoras no modelo para que R^2 se aproxime de 1.

O R^2 **ajustado** para os graus de liberdade é dado por

$$\overline{R}^2 = 1 - \frac{\boldsymbol{e}^\top \boldsymbol{e}/(n-p)}{\boldsymbol{y}^\top \widetilde{\boldsymbol{M}} \boldsymbol{y}/(n-1)}.$$

Para efeitos computacionais, temos

$$\overline{R}^2 = 1 - \frac{n-1}{n-p} \times (1 - R^2).$$

- O $\overline{R}^2$ **pode decrescer** quando uma nova variável é adicionada ao conjunto de variáveis independentes.
- Na verdade, $\overline{R}^2$ **pode até ser negativo.**

- Uma alternativa para comparar modelos com uma maior penalização pela perda dos graus de liberdade é dada por

$$\overline{R}^2 = 1 - \frac{n+p}{n-p} \times (1 - R^2).$$

O modelo que tem $\overline{R}^2$ máximo é o mesmo que tem o valor mínimo do **critério de predição (CP)**:

$$\text{CP} = \frac{\boldsymbol{e}^\top \boldsymbol{e}}{n-p}\left(1 + \frac{p}{n}\right) = s^2\left(1 + \frac{p}{n}\right).$$

- Veremos outros critérios de comparação de modelos baseados no máximo da função de verossimilhança com penalização pelo número de parâmetros (AIC, BIC, HQIC, etc.).

R^2 e o termo constante do modelo

A prova que $0 \leqslant R^2 \leqslant 1$ requer que $\boldsymbol{X}$ contenha uma coluna de 1's.

Se isto não ocorrer, então $\widetilde{\boldsymbol{M}}\boldsymbol{e} \neq \boldsymbol{e}$, $\boldsymbol{e}^\top \widetilde{\boldsymbol{M}}\boldsymbol{X} \neq \boldsymbol{0}$ e o termo $2\boldsymbol{e}^\top \widetilde{\boldsymbol{M}}\boldsymbol{X}\boldsymbol{b}$ em

$$\boldsymbol{y}^\top \widetilde{\boldsymbol{M}}\boldsymbol{y} = (\widetilde{\boldsymbol{M}}\boldsymbol{X}\boldsymbol{b} + \widetilde{\boldsymbol{M}}\boldsymbol{e})^\top (\widetilde{\boldsymbol{M}}\boldsymbol{X}\boldsymbol{b} + \widetilde{\boldsymbol{M}}\boldsymbol{e})$$

não se cancela.

Daí,

$$R^2 = 1 - \frac{\boldsymbol{e}^\top \boldsymbol{e}}{\boldsymbol{y}^\top \widetilde{\boldsymbol{M}}\boldsymbol{y}}$$

pode resultar em qualquer valor.

Portanto, devemos ter cuidado ao analisar o R^2 de um modelo de regressão que não contenha o termo constante.

Este é um dos motivos que, em geral, deixa-se o termo constante no modelo de regressão linear mesmo que testes de hipóteses indiquem o contrário.

Regressão transformada linearmente

Na regressão de $\boldsymbol{y}$ sobre $\boldsymbol{X}$, suponha que as colunas de $\boldsymbol{X}$ são transformadas linearmente.

Exemplo - precificação de arte

A teoria I da determinação de preços de leilão em pinturas de Monet diz que o preço é determinado pelas dimensões (largura W e altura H) da pintura,

$$\begin{aligned} \ln \text{preço} &= \beta_1 + \beta_2 \ln W + \beta_3 \ln H + \varepsilon \\ &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon. \end{aligned}$$

A teoria II diz que os compradores estão interessados na área da pintura e a razão do aspecto,

$$\begin{aligned} \ln \text{preço} &= \gamma_1 + \gamma_2 \ln(WH) + \gamma_3 \ln(W/H) + \xi \\ &= \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \xi. \end{aligned}$$

É evidente que $z_1 = x_1$, $z_2 = x_2 + x_3$ e $z_3 = x_2 - x_3$.

Em termos matriciais, temos $\boldsymbol{Z} = \boldsymbol{XP}$ sendo

$$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & -1 \end{bmatrix}.$$

Teorema

Na regressão linear de $\boldsymbol{y}$ sobre $\boldsymbol{Z} = \boldsymbol{XP}$ sendo $\boldsymbol{P}$ uma matriz não singular que transforma as colunas de $\boldsymbol{X}$, os coeficientes são iguais a $\boldsymbol{P}^{-1}\boldsymbol{b}$ sendo $\boldsymbol{b}$ o vetor de coeficientes da regressão linear de $\boldsymbol{y}$ sobre $\boldsymbol{X}$. Além disto, os R^2 das duas regressões são idênticos.

Prova: Os coeficientes são

$$\begin{aligned} \boldsymbol{d} &= (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{y} = \left[(\boldsymbol{XP})^\top (\boldsymbol{XP})\right]^{-1} (\boldsymbol{XP})^\top \boldsymbol{y} = (\boldsymbol{P}^\top \boldsymbol{X}^\top \boldsymbol{XP})^{-1}\boldsymbol{P}^\top \boldsymbol{X}^\top \boldsymbol{y} \\ &= \boldsymbol{P}^{-1}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{P}^{\top -1}\boldsymbol{P}^\top \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{P}^{-1}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{P}^{-1}\boldsymbol{b}. \end{aligned}$$

Temos também que

$$\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{Zd} = \boldsymbol{y} - \boldsymbol{ZP}^{-1}\boldsymbol{b} = \boldsymbol{y} - \boldsymbol{XPP}^{-1}\boldsymbol{b} = \boldsymbol{y} - \boldsymbol{Xb} = \boldsymbol{e}.$$

Logo, $\boldsymbol{u}^\top \boldsymbol{u} = \boldsymbol{e}^\top \boldsymbol{e} \Rightarrow R^2_{\boldsymbol{y}|\boldsymbol{Z}} = R^2_{\boldsymbol{y}|\boldsymbol{X}}$.

Hipóteses do modelo de regressão linear

HP.1: Linearidade: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \varepsilon$;

HP.2: Posto completo: a matriz $\boldsymbol{X}$ $(n \times p)$ é de posto completo;

HP.3: Exogeneidade das variáveis independentes: $\mathrm{E}(\varepsilon_i | \boldsymbol{x}_i) = 0$;

HP.4: Homocedasticidade e autocorrelação nula: cada distúrbio, ε_i, tem a mesma variância, σ^2, e é não correlacionado com qualquer outro distúrbio ε_j;

HP.5: Geração dos dados: os dados em $(x_{i1}, x_{i2}, \ldots, x_{ip})$ podem ser qualquer mistura de constantes e variáveis aleatórias; e

HP.6: Distribuição normal: os distúrbios são normalmente distribuídos.

O estimador de mínimos quadrados

Por hipótese,

$$\mathrm{E}(\varepsilon|\boldsymbol{x}) = 0 \Rightarrow \mathrm{Cov}(\boldsymbol{x}, \varepsilon) = \mathrm{Cov}\left(\mathrm{E}(\varepsilon|\boldsymbol{x}), \boldsymbol{x}\right) = \mathbf{0} \quad \text{e} \quad \mathrm{E}(\varepsilon) = \mathrm{E}_{\boldsymbol{x}}\left(\mathrm{E}(\varepsilon|\boldsymbol{x})\right) = 0.$$

Mas,

$$\mathrm{Cov}(\boldsymbol{x}, \varepsilon) = \mathrm{E}_{\boldsymbol{x},\varepsilon}(\boldsymbol{x}\varepsilon) = \mathrm{E}_{\boldsymbol{x},y}(\boldsymbol{x}(y - \boldsymbol{x}^\top \boldsymbol{\beta})) = \mathbf{0} \quad \Rightarrow$$

$$\mathrm{E}_{\boldsymbol{x},y}(\boldsymbol{x}y) = \mathrm{E}_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{\beta} \quad \textbf{(população)}.$$

Tomemos a equação normal de mínimos quadrados $\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{b}$ que dividida por n, podemos reescrever como

$$\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right) \boldsymbol{b} \quad \textbf{(amostra)}.$$

Assim, o estimador de mínimos está parecido com sua respectiva quantidade populacional. Podemos aplicar a lei dos grandes números sob certas condições.

Preditor linear ótimo

Considere o problema de encontrar um preditor linear ótimo para y.

Critério: o erro quadrático médio (EQM) do preditor linear y.

Ignoraremos a **HP.6** e desconsideraremos **HP.1** que $\mathrm{E}(y|\boldsymbol{x})$ é linear.

Queremos minimizar

$$\mathrm{EQM} = \mathrm{E}_{\boldsymbol{x},y}([y - \boldsymbol{x}^\top \gamma]^2).$$

Isto pode ser reescrito como

$$\mathrm{EQM} = \mathrm{E}_{\boldsymbol{x},y}([y - \mathrm{E}(y|\boldsymbol{x})]^2) + \mathrm{E}_{\boldsymbol{x},y}([\mathrm{E}(y|\boldsymbol{x}) - \boldsymbol{x}^\top \gamma]^2).$$

A condição necessária é

$$\begin{aligned}
\frac{\partial \mathrm{E}_{\boldsymbol{x},y}([\mathrm{E}(y|\boldsymbol{x}) - \boldsymbol{x}^\top \gamma]^2)}{\partial \gamma} &= \mathrm{E}_{\boldsymbol{x},y}\left(\frac{\partial [\mathrm{E}(y|\boldsymbol{x}) - \boldsymbol{x}^\top \gamma]^2}{\partial \gamma}\right) \\
&= -2\mathrm{E}_{\boldsymbol{x},y}(\boldsymbol{x}[\mathrm{E}(y|\boldsymbol{x}) - \boldsymbol{x}^\top \gamma]) = \boldsymbol{0} \Rightarrow \\
\mathrm{E}_{\boldsymbol{x},y}(\boldsymbol{x}\mathrm{E}(y|\boldsymbol{x})) &= \mathrm{E}_{\boldsymbol{x},y}(\boldsymbol{x}\boldsymbol{x}^\top)\gamma.
\end{aligned}$$

Mas

$$\begin{aligned} \mathrm{E}_{\boldsymbol{x},y}(\boldsymbol{x}\mathrm{E}(y|\boldsymbol{x})) &= \mathrm{Cov}(\boldsymbol{x}, \mathrm{E}(y|\boldsymbol{x})) + \mathrm{E}_{\boldsymbol{x}}(\boldsymbol{x})\mathrm{E}_{\boldsymbol{x}}(\mathrm{E}(y|\boldsymbol{x})) \\ &= \mathrm{Cov}(\boldsymbol{x}, y) + \mathrm{E}_{\boldsymbol{x}}(\boldsymbol{x})\mathrm{E}_{y}(y) = \mathrm{E}_{\boldsymbol{x},y}(\boldsymbol{x}y). \end{aligned}$$

Portanto, a condição necessária para minimizar o preditor linear, EQM, é

$$\mathrm{E}_{\boldsymbol{x},y}(\boldsymbol{x}y) = \mathrm{E}_{\boldsymbol{x},y}(\boldsymbol{x}\boldsymbol{x}^{\top})\gamma.$$

Teorema

Se o processo de geração dos dados $(y_i, \boldsymbol{x}_i)$, $i = 1, 2, \ldots, n$, é tal que a lei dos grandes números se aplica ao estimador

$$\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\top}\right)\boldsymbol{b}$$

das matrizes

$$\mathrm{E}_{\boldsymbol{x},y}(\boldsymbol{x}y) = \mathrm{E}_{\boldsymbol{x}}(\boldsymbol{x}\boldsymbol{x}^{\top})\beta,$$

então o mínimo do preditor linear do erro quadrático esperado de y_i é estimado pela regressão de mínimos quadrados.

Estimação não tendenciosa

O estimador de mínimos quadrados é não tendencioso para cada amostra.

Temos

$$\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\beta + \varepsilon) = \beta + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \varepsilon \quad \text{e}$$

$$\mathrm{E}(\boldsymbol{b}|\boldsymbol{X}) = \beta + \mathrm{E}((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \varepsilon|\boldsymbol{X}) = \beta \quad \text{(pela HP.3)}.$$

Portanto,

$$\mathrm{E}(\boldsymbol{b}) = \mathrm{E}_{\boldsymbol{X}}(\mathrm{E}(\boldsymbol{b}|\boldsymbol{X})) = \mathrm{E}_{\boldsymbol{X}}(\beta) = \beta.$$

Viés: omissão de variáveis relevantes

Suponha que o modelo correto seja

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

sendo p_1 e p_2 os números de colunas em $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$, respectivamente.

A regressão de $\boldsymbol{y}$ sobre $\boldsymbol{X}_1$ (sem incluir $\boldsymbol{X}_2$) resulta no estimador

$$\boldsymbol{b}_1 = (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{y} = \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_2\boldsymbol{\beta}_2 + (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{\varepsilon}.$$

Logo,

$$\mathrm{E}(\boldsymbol{b}_1|\boldsymbol{X}) = \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_2\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1 + \boldsymbol{P}_{1.2}\boldsymbol{\beta}_2,$$

sendo $\boldsymbol{P}_{1.2} = (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_2$.

Portanto, $\boldsymbol{b}_1$ é um estimador tendencioso de $\boldsymbol{\beta}_1$, exceto para os casos que $\boldsymbol{X}_1^\top\boldsymbol{X}_2 = \boldsymbol{0}$ ou $\boldsymbol{\beta}_2 = \boldsymbol{0}$.

A estimação do modelo pode ser vista ao impor **incorretamente** a restrição $\boldsymbol{\beta}_2 = \boldsymbol{0}$. Isto introduz um viés na estimação de $\boldsymbol{b}_1$.

Inclusão de variáveis irrelevantes

Suponha que o modelo de regressão correto seja

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon},$$

mas que estimamos como a modelo de regressão (variáveis extras)

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

Aqui, a inclusão de $\boldsymbol{X}_2$ no modelo de regressão pode ser vista como uma falha em impor que $\boldsymbol{\beta}_2 = \mathbf{0}$. Portanto, não a nada a se provar aqui.

Temos que

$$\mathrm{E}(\boldsymbol{b}|\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{bmatrix}.$$

É possível provar que ao aumentar o número de variáveis desnecessárias no modelo faz com que a precisão das estimativas diminua (a menos que $\boldsymbol{X}_1^\top \boldsymbol{X}_2 = \mathbf{0}$).

Exemplo

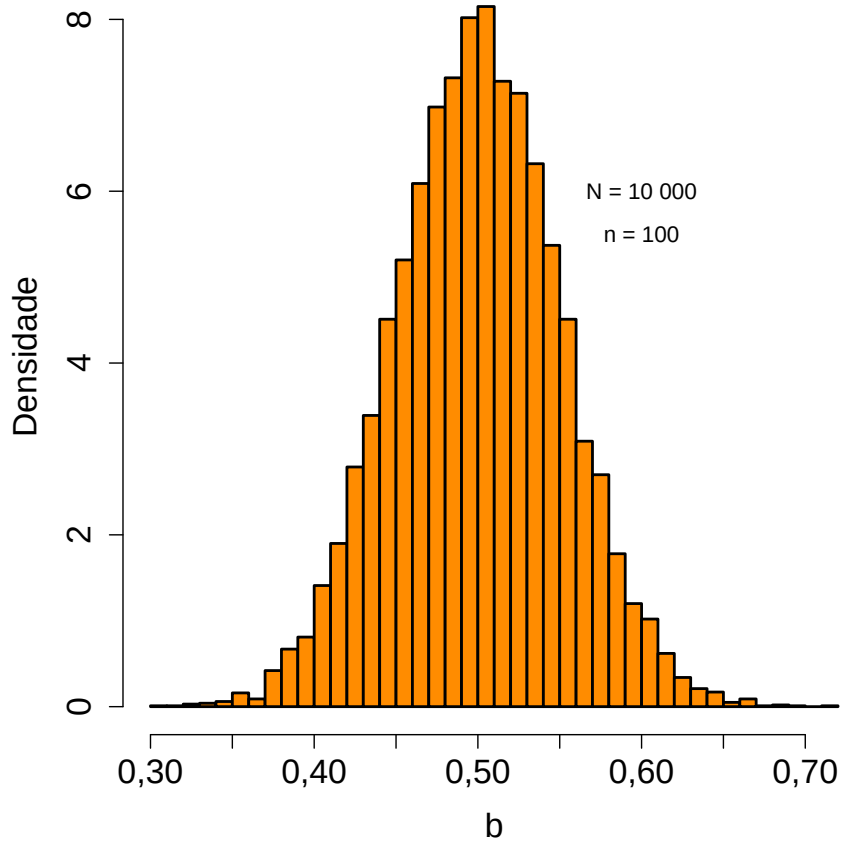


Figura 5: Distribuição amostral do estimador *b* de β na regressão $y = \alpha + \beta x + \varepsilon$.

Ver arquivo exemplo_08.r

Variância amostral do estimador de mínimos quadrados

$$\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top (\boldsymbol{X}\beta + \varepsilon) = \beta + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \varepsilon.$$

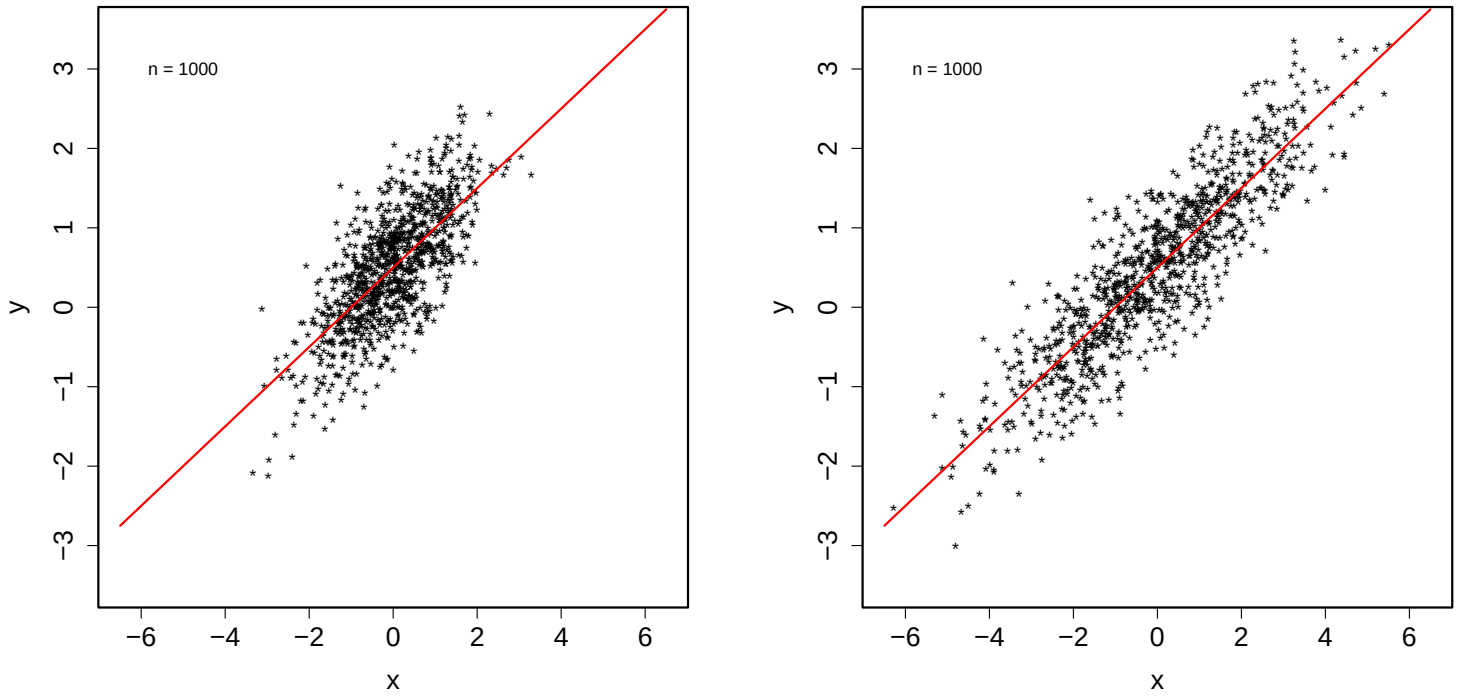
Logo,

$$\begin{aligned}
\text{Var}(\boldsymbol{b}|\boldsymbol{X}) &= \text{E}([\boldsymbol{b} - \beta][\boldsymbol{b} - \beta^\top]|\boldsymbol{X}) \\
&= \text{E}((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \varepsilon\varepsilon^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}|\boldsymbol{X}) \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \text{E}(\varepsilon\varepsilon^\top|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \sigma^2 \boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \quad \text{(pela HP.4)} \\
&= \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}.
\end{aligned}$$

Suponha uma regressão linear simples da forma $y = \beta_1 + \beta_2 x + \varepsilon$. Assim, temos

$$\text{Var}(b_2|\boldsymbol{X}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Exemplo



(a) Menor variância em x.

(b) Maior variância em x.

Figura 6: Variância do estimador b de β na regressão $y = \alpha + \beta x + \varepsilon$.

Ver arquivo exemplo_09.r

Tabela 7: Estimativas para os dados com **menor** dispersão em x.

Parâmetro	Estimativa	Erro padrão	Estat. t	Valor p
β_0	0,52368	0,01579	33,16	$< 10^{-15}$
β_1	0,51920	0,01566	33,15	$< 10^{-15}$

$$\hat{\sigma}_1 = 0,4992 \text{ e } R_1^2 = 0,5240.$$

Tabela 8: Estimativas para os dados com **maior** dispersão em x.

Parâmetro	Estimativa	Erro padrão	Estat. t	Valor p
β_0	0,52294	0,01580	33,09	$< 10^{-15}$
β_1	0,49683	0,00826	60,15	$< 10^{-15}$

$$\hat{\sigma}_2 = 0,4996 \text{ e } R_2^2 = 0,7838.$$

Agora, seja $\boldsymbol{b}_\star = \boldsymbol{C}\boldsymbol{y}$ um outro estimador não tendencioso e linear de $\boldsymbol{\beta}$, em que $\boldsymbol{C}$ é uma matriz $p \times n$. Então,

$$\boldsymbol{\beta} = \mathrm{E}(\boldsymbol{b}_\star|\boldsymbol{X}) = \mathrm{E}(\boldsymbol{C}\boldsymbol{y}|\boldsymbol{X}) = \mathrm{E}(\boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\varepsilon}|\boldsymbol{X}) \quad \Rightarrow \quad \boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}.$$

(existem vários candidatos para $\boldsymbol{C}$ que satisfazem a igualdade) e

$$\mathrm{Var}(\boldsymbol{b}_\star|\boldsymbol{X}) = \sigma^2 \boldsymbol{C}\boldsymbol{C}^\top.$$

Agora, seja $\boldsymbol{D} = \boldsymbol{C} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ tal que $\boldsymbol{D}\boldsymbol{y} = \boldsymbol{b}_\star - \boldsymbol{b}$. Então,
$\boldsymbol{b}_\star = \boldsymbol{D}\boldsymbol{y} + \boldsymbol{b} = (\boldsymbol{D} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{y}$ e

$$\mathrm{Var}(\boldsymbol{b}_\star|\boldsymbol{X}) = \sigma^2 \left[(\boldsymbol{D} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)(\boldsymbol{D} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)^\top\right].$$

Sabemos que $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I} = \boldsymbol{D}\boldsymbol{X} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X} \quad \Rightarrow \quad \boldsymbol{D}\boldsymbol{X} = \boldsymbol{0}$. Portanto,

$$\mathrm{Var}(\boldsymbol{b}_\star|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \sigma^2\boldsymbol{D}\boldsymbol{D}^\top = \mathrm{Var}(\boldsymbol{b}|\boldsymbol{X}) + \sigma^2\boldsymbol{D}\boldsymbol{D}^\top.$$

e $\boldsymbol{D}\boldsymbol{D}^\top$ é uma forma quadrática ($\boldsymbol{a}^\top\boldsymbol{D}\boldsymbol{D}^\top\boldsymbol{a} \geqslant 0$).

Logo, $\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X}) \leqslant \mathrm{Var}(\boldsymbol{b}_\star|\boldsymbol{X})$ para qualquer $\boldsymbol{b}_\star \neq \boldsymbol{b}$ e $\boldsymbol{X}$ fixo.

Teorema de Gauss-Markov

No modelo de regressão linear clássico com matriz de variáveis regressoras $\boldsymbol{X}$, o estimador de mínimos quadrados $\boldsymbol{b}$ é o estimador não tendencioso e linear de variância mínima de β. Para qualquer vetor de constantes $\boldsymbol{v}$, o estimador não tendencioso e linear de variância mínima de $\boldsymbol{v}^\top \beta$ no modelo de regressão clássico é $\boldsymbol{v}^\top \boldsymbol{b}$, sendo $\boldsymbol{b}$ o estimador de mínimos quadrados.

O teorema também implica que o estimador de mínimos quadrados para cada coeficiente também tem variância mínima. Basta tomar um vetor $\boldsymbol{v} = (0, \ldots, 0, 1, 0, \ldots, 0)$ com o valor 1 no coeficiente a ser estimador.

A variância incondicional é dada por

$$\begin{aligned} \text{Var}(\boldsymbol{b}) &= \text{E}_{\boldsymbol{X}}(\text{Var}(\boldsymbol{b}|\boldsymbol{X})) + \text{Var}_{\boldsymbol{X}}(\text{E}(\boldsymbol{b}|\boldsymbol{X})) \\ &= \sigma^2 \text{E}_{\boldsymbol{X}}((\boldsymbol{X}^\top \boldsymbol{X})^{-1}), \end{aligned}$$

pois $\text{E}(\boldsymbol{b}|\boldsymbol{X}) = \beta \quad \Rightarrow \quad \text{Var}_{\boldsymbol{X}}(\text{E}(\boldsymbol{b}|\boldsymbol{X})) = \boldsymbol{0}$.

Estimação da variância

Sabemos que $\text{Var}(\boldsymbol{b}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. Precisamos estimar σ^2.

Temos que $\text{Var}(\varepsilon_i) = \text{E}(\varepsilon_i^2) = \sigma^2$ e e_i é uma estimativa de ε_i. Então,

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

parece um estimador natural de σ^2.

Os resíduos de mínimos quadrados são

$$\boldsymbol{e} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}[\boldsymbol{X}\beta + \varepsilon] = \boldsymbol{M}\varepsilon, \quad \text{pois} \quad \boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}.$$

Este estimador de σ^2 é baseado em $\boldsymbol{e}^\top\boldsymbol{e} = \varepsilon^\top\boldsymbol{M}\varepsilon$.

Note que os resíduos são correlacionados:

$$\text{Var}(\boldsymbol{e}|\boldsymbol{X}) = \text{E}(\boldsymbol{e}\boldsymbol{e}^\top|\boldsymbol{X}) = \text{E}(\boldsymbol{M}\varepsilon\varepsilon^\top\boldsymbol{M}|\boldsymbol{X}) = \boldsymbol{M}\text{E}(\varepsilon\varepsilon^\top|\boldsymbol{X})\boldsymbol{M} = \boldsymbol{M}\sigma^2\boldsymbol{I}\boldsymbol{M} = \sigma^2\boldsymbol{M}.$$

Note que $\boldsymbol{M}$ é uma matriz positiva definida.

Agora, estamos interessados em

$$\mathrm{E}(\boldsymbol{e}^\top \boldsymbol{e}|\boldsymbol{X}) = \mathrm{E}(\varepsilon^\top \boldsymbol{M}\varepsilon|\boldsymbol{X}).$$

A quantidade escalar $\varepsilon^\top \boldsymbol{M}\varepsilon$ é uma matriz 1×1, logo é igual ao seu traço.

Segue-se que

$$\begin{aligned} \mathrm{E}(\mathrm{tr}(\varepsilon^\top \boldsymbol{M}\varepsilon)|\boldsymbol{X}) &= \mathrm{E}(\mathrm{tr}(\boldsymbol{M}\varepsilon\varepsilon^\top)|\boldsymbol{X}) = \mathrm{tr}(\mathrm{E}(\boldsymbol{M}\varepsilon\varepsilon^\top|\boldsymbol{X})) = \mathrm{tr}(\boldsymbol{M}\mathrm{E}(\varepsilon\varepsilon^\top|\boldsymbol{X})) \\ &= \mathrm{tr}(\boldsymbol{M}\sigma^2\boldsymbol{I}) = \sigma^2\mathrm{tr}(\boldsymbol{M}). \end{aligned}$$

O traço de $\boldsymbol{M}$ é

$$\mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) = \mathrm{tr}(\boldsymbol{I}_n) - \mathrm{tr}((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{I}_n) - \mathrm{tr}(\boldsymbol{I}_p) = n - p.$$

Portanto, $\mathrm{E}(\boldsymbol{e}^\top \boldsymbol{e}|\boldsymbol{X}) = (n-p)\sigma^2$.

Logo, o estimador natural de σ^2 é tendencioso. Um estimador não tendencioso de σ^2 é

$$s^2 = \frac{\boldsymbol{e}^\top \boldsymbol{e}}{n-p}.$$

Este estimador é não tendencioso incondicionalmente também,

$$\mathrm{E}(s^2) = \mathrm{E}_{\boldsymbol{X}}(\mathrm{E}(s^2|\boldsymbol{X})) = \mathrm{E}_{\boldsymbol{X}}(\sigma^2) = \sigma^2.$$

Podemos estimar $\mathrm{Var}(\boldsymbol{b}|\boldsymbol{X})$ por $\widehat{\mathrm{Var}}(\boldsymbol{b}|\boldsymbol{X}) = s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$.

A raiz quadrada do k-ésimo elemento da diagonal de $s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ é o erro padrão do estimador b_k.

Normalidade (HP.6)

Se tivermos a hipótese $\varepsilon \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, então

$$\begin{aligned} (\boldsymbol{b}|\boldsymbol{X}) &\sim \mathcal{N}_p(\beta, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}), \quad \text{e} \\ (b_k|\boldsymbol{X}) &\sim \mathcal{N}(\beta_k, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{kk}). \end{aligned}$$

Lembre-se que $\boldsymbol{b} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\beta + \varepsilon) = \beta + \boldsymbol{A}\varepsilon$ com $\boldsymbol{A} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, ou seja, sob **HP.6** $\boldsymbol{b}$ é uma combinação linear de normais. Portanto, $\boldsymbol{b}$ (condicional a $\boldsymbol{X}$) tem distribuição normal.

Propriedades dos estimadores de mínimos quadrados para amostras grandes

Utilizaremos **HP.1** até **HP.5** (sem normalidade).

Consistência de $\boldsymbol{b}$

Suponha que **HP.5a** $(\boldsymbol{x}_i, \varepsilon_i)$, $i = 1, \ldots, n$, seja uma sequência de observações independentes e que

$$\text{plim}\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n} = \boldsymbol{Q}, \quad \text{uma matriz positiva definida.}$$

Temos que

$$\boldsymbol{b} = \beta + \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n}\right)^{-1} \left(\frac{\boldsymbol{X}^\top \varepsilon}{n}\right).$$

Se $\boldsymbol{Q}^{-1}$ existir, então

$$\text{plim}\boldsymbol{b} = \beta + \boldsymbol{Q}^{-1}\text{plim}\left(\frac{\boldsymbol{X}^\top \varepsilon}{n}\right).$$

pois a inversa é uma função contínua da matriz original.

Segue-se

$$\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w}_i = \overline{\boldsymbol{w}}.$$

Portanto,

$$\text{plim}\boldsymbol{b} = \boldsymbol{\beta} + \boldsymbol{Q}^{-1}\text{plim}\overline{\boldsymbol{w}}.$$

Temos que

$$\text{E}(\boldsymbol{w}_i) = \text{E}_{\boldsymbol{x}}(\text{E}(\boldsymbol{w}_i|\boldsymbol{X})) = \text{E}_{\boldsymbol{X}}(\boldsymbol{x}_i\text{E}(\varepsilon_i|\boldsymbol{X})) = \boldsymbol{0} \quad \Rightarrow \quad \text{E}(\overline{\boldsymbol{w}}) = \boldsymbol{0}.$$

Além disso,

$$\begin{aligned}
\text{Var}(\overline{\boldsymbol{w}}|\boldsymbol{X}) &= \text{E}(\overline{\boldsymbol{w}}\,\overline{\boldsymbol{w}}^\top|\boldsymbol{X}) = \frac{1}{n}\boldsymbol{X}^\top\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top|\boldsymbol{X})\boldsymbol{X}\frac{1}{n} = \left(\frac{\sigma^2}{n}\right)\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right) \quad \Rightarrow \\
\text{Var}(\overline{\boldsymbol{w}}) &= \text{E}_{\boldsymbol{X}}(\text{Var}(\overline{\boldsymbol{w}}|\boldsymbol{X})) + \text{Var}_{\boldsymbol{X}}(\text{E}(\overline{\boldsymbol{w}}|\boldsymbol{X})) = \left(\frac{\sigma^2}{n}\right)\text{E}_{\boldsymbol{X}}\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right).
\end{aligned}$$

Se $\lim_{n\to\infty}\text{E}_{\boldsymbol{X}}\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right) = \boldsymbol{Q}$, então $\lim_{n\to\infty}\text{Var}(\overline{\boldsymbol{w}}) = \boldsymbol{0}$ e

$$\text{plim}\frac{\boldsymbol{X}^\top\boldsymbol{\varepsilon}}{n} = \boldsymbol{0} \quad \Rightarrow$$

$$\text{plim}\boldsymbol{b} = \boldsymbol{\beta} + \boldsymbol{Q}^{-1}\boldsymbol{0} = \boldsymbol{\beta}.$$

Assim, $\boldsymbol{b}$ é um estimador consistente de $\boldsymbol{\beta}$ no modelo de regressão linear.

Condições de Grenander

HG.1: Para cada coluna de $\boldsymbol{X}$, $\boldsymbol{x}_k$, se $d_{nk}^2 = \boldsymbol{x}_k^\top \boldsymbol{x}_k$, então $\lim_{n\to\infty} d_{nk}^2 = +\infty$. Assim, $\boldsymbol{x}_k$ não degenera para uma sequência de zeros. A soma de quadrados continuará a crescer com o tamanho da amostra.

HG.2: $\lim_{n\to\infty} x_{ik}^2 / d_{nk}^2 = 0$ para todo $i = 1, \ldots, n$. Esta condição implica que nenhuma observação dominará $\boldsymbol{x}_k^\top \boldsymbol{x}_k$ e quando $n \to \infty$, observações individuais serão menos importantes.

HG.3: Seja $\boldsymbol{R}_n$ a correlação amostral das colunas de $\boldsymbol{X}$, excluindo a variável constante. Então, $\lim_{n\to\infty} \boldsymbol{R}_n = \boldsymbol{C}$, uma matriz positiva definida. Isto implica que a condição de posto completo sempre será satisfeita. Nós já assumimos que $\boldsymbol{X}$ tem posto completo para uma amostra finita, então esta hipótese assegura que a condição nunca será violada.

Estas condições garantem que a matriz de dados são “bem comportadas” em amostras grandes mesmo nos casos envolvendo séries temporais com tendência temporal, séries polinômiais e variáveis de tendência (as hipóteses anteriores eram muito restritivas).

Normalidade assintótica de $\boldsymbol{b}$

Utilizaremos a **HP.3** (exogeneidade) e adicionaremos a hipótese que as observações são independentes.

Sabemos que

$$\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta}) = \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)\boldsymbol{X}^\top \boldsymbol{\varepsilon} \quad \Rightarrow$$

$$\left[\text{plim}\left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n}\right)^{-1}\right]\left(\frac{1}{\sqrt{n}}\right)\boldsymbol{X}^\top \boldsymbol{\varepsilon} = \boldsymbol{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\boldsymbol{X}^\top \boldsymbol{\varepsilon}.$$

Mas

$$\left(\frac{1}{\sqrt{n}}\right)\boldsymbol{X}^\top \boldsymbol{\varepsilon} = \sqrt{n}(\overline{\boldsymbol{w}} - \text{E}(\overline{\boldsymbol{w}})) \quad \text{sendo} \quad \text{E}(\overline{\boldsymbol{w}}) = \boldsymbol{0}.$$

Podemos utilizar a versão multivariada do teorema central do limite de Lindeberg-Feller para obter a distribuição limite de $\sqrt{n}\overline{\boldsymbol{w}}$.

Mas $\overline{\boldsymbol{w}}$ é a média de n vetores aleatórios independentes $\boldsymbol{w}_i = \boldsymbol{x}_i\varepsilon_i$ com média $\mathbf{0}$ e variância $\text{Var}(\boldsymbol{w}_i) = \text{Var}(\boldsymbol{x}_i\varepsilon_i) = \sigma^2\text{E}(\boldsymbol{x}_i\boldsymbol{x}_i^\top) = \sigma^2\boldsymbol{Q}_i$. Segue-se que a variância de $\sqrt{n}\overline{\boldsymbol{w}}$ é

$$\sigma^2\overline{\boldsymbol{Q}}_n = \sigma^2\left(\frac{1}{n}\right)[\boldsymbol{Q}_1 + \boldsymbol{Q}_2 + \cdots + \boldsymbol{Q}_n].$$

Se a soma não for dominada por nenhum termo e se os regressores forem bem comportados,

$$\lim_{n\to\infty}\sigma^2\overline{\boldsymbol{Q}}_n = \sigma^2\boldsymbol{Q}.$$

Portanto, se $[\boldsymbol{x}_i\varepsilon_i]$, $i = 1, \ldots, n$, forem vetores independentes distribuídos com média $\mathbf{0}$ e variância $\sigma^2\boldsymbol{Q}_i$ e se $\text{plim}\dfrac{\boldsymbol{X}^\top\boldsymbol{X}}{n} = \boldsymbol{Q}$, então

$$\frac{1}{\sqrt{n}}\boldsymbol{X}^\top\varepsilon \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{Q}).$$

Segue-se que

$$\boldsymbol{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\boldsymbol{X}^{\top}\varepsilon \xrightarrow{d} \mathcal{N}(\boldsymbol{Q}^{-1}\mathbf{0}, \boldsymbol{Q}^{-1}\sigma^2\boldsymbol{Q}\boldsymbol{Q}^{-1}) \Rightarrow \sqrt{n}(\boldsymbol{b}-\beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{Q}^{-1}).$$

Teorema

Se $\{\varepsilon_i\}$ são independentemente distribuídos com média 0 e variância σ^2 e x_{ik} é tal que as condições de Grenander são satisfeitas, então

$$\boldsymbol{b} \overset{a}{\sim} \mathcal{N}\left(\beta, \frac{\sigma^2}{n}\boldsymbol{Q}^{-1}\right).$$

Na prática, é necessário estimar $(1/n)\boldsymbol{Q}^{-1}$ com $(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}$ e σ^2 com $\boldsymbol{e}^{\top}\boldsymbol{e}/(n-p)$.

Consistência de s^2

Temos a variância assintótica de $\boldsymbol{b}$: $(\sigma^2/n)\boldsymbol{Q}^{-1}$.

Podemos restringir a atenção a s^2 como estimador consistente de σ^2.

$$\begin{aligned} s^2 &= \frac{1}{n-p}\varepsilon^\top \boldsymbol{M}\varepsilon = \frac{1}{n-p}\left[\varepsilon^\top\varepsilon - \varepsilon^\top \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\varepsilon\right] \\ &= \frac{n}{n-p}\left[\frac{\varepsilon^\top\varepsilon}{n} - \left(\frac{\varepsilon^\top\boldsymbol{X}}{n}\right)\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{X}^\top\varepsilon}{n}\right)\right]. \end{aligned}$$

Mas

$$\operatorname{plim}\left(\frac{\varepsilon^\top\boldsymbol{X}}{n}\right)\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{X}^\top\varepsilon}{n}\right) = 0, \quad \text{pois}$$

$$\operatorname{plim}\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right)^{-1} = \boldsymbol{Q}^{-1} \quad \text{e} \quad \operatorname{plim}\left(\frac{\boldsymbol{X}^\top\varepsilon}{n}\right) = 0.$$

Agora, vamos analizar

$$\overline{\varepsilon^2} = \frac{\varepsilon^\top\varepsilon}{n} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2.$$

Neste caso, $\mathrm{E}(\varepsilon_i^2) = \sigma^2$. Pelo teorema de Markov, precisamos de $\mathrm{E}|\varepsilon_i^2|^{1+\delta} < \infty$, $\delta > 0$, tal que a hipótese mínima até agora é que ε_i tenha momentos finitos um pouco maior que 2.

Além disso, se assumirmos que cada ε_i tem a mesma distribuição, então pelo teorema de Khinchine, momentos finitos até 2 é suficiente.

Convergência em média quadrática exigiria que $\mathrm{E}(\varepsilon_i^4) = \phi_\varepsilon < \infty$.

Então, os termos da soma são independentes, com média σ^2 e variância $\phi_\varepsilon - \sigma^4$. Logo, sob condições fracas,

$$\operatorname{plim} \frac{\boldsymbol{e}^\top \boldsymbol{e}}{n-p} = \operatorname{plim} \frac{\varepsilon^\top \boldsymbol{M} \varepsilon}{n-p} = \operatorname{plim} s_n^2 = \sigma^2 \;\Rightarrow\; \operatorname{plim} s_n^2 \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n} \right)^{-1} = \sigma^2 \boldsymbol{Q}^{-1}.$$

O estimador apropriado da matriz de covariância assintótica de $\boldsymbol{b}$ é

$$\boldsymbol{S}_{\boldsymbol{b}} = s^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

Distribuição assintótica de uma função de $\boldsymbol{b}$

Seja $\boldsymbol{f}(\boldsymbol{b})$ um conjunto de J funções contínuas, e continuamente diferenciáveis do estimador de mínimos quadrados, e seja

$$\boldsymbol{C}(\boldsymbol{b}) = \frac{\partial \boldsymbol{f}(\boldsymbol{b})}{\partial \boldsymbol{b}^\top},$$

sendo $\boldsymbol{C}$ uma matriz $J \times p$ cuja j-ésima linha é o vetor de derivadas da j-ésima função com respeito a $\boldsymbol{b}^\top$.

Pelo teorema de Slutsky,

$$\text{plim}\boldsymbol{f}(\boldsymbol{b}) = \boldsymbol{f}(\beta) \quad \text{e} \quad \text{plim}\boldsymbol{C}(\boldsymbol{b}) = \frac{\partial \boldsymbol{f}(\beta)}{\partial \beta^\top} = \boldsymbol{\Gamma}.$$

Teorema

Se $\boldsymbol{f}(\boldsymbol{b})$ é um conjunto de funções contínuas e continuamente diferenciável de $\boldsymbol{b}$ tal que $\boldsymbol{\Gamma} = \frac{\partial \boldsymbol{f}(\beta)}{\partial \beta^\top}$, e se $\boldsymbol{b} \overset{a}{\sim} \mathcal{N}\left(\beta, \frac{\sigma^2}{n}\boldsymbol{Q}^{-1}\right)$, então

$$\boldsymbol{f}(\boldsymbol{b}) \overset{a}{\sim} \mathcal{N}\left(\boldsymbol{f}(\beta), \boldsymbol{\Gamma}\left(\frac{\sigma^2}{n}\boldsymbol{Q}^{-1}\right)\boldsymbol{\Gamma}^\top\right).$$

A prova do teorema utiliza a aproximação de Taylor:

$$\boldsymbol{f}(\boldsymbol{b}) = \boldsymbol{f}(\boldsymbol{\beta}) + \boldsymbol{\Gamma}(\boldsymbol{b} - \boldsymbol{\beta}) + \text{outros termos.}$$

Na prática, estimaremos a variância de $\boldsymbol{f}(\boldsymbol{b})$ por $\boldsymbol{S}_{\boldsymbol{f}(\boldsymbol{b})} = s^2 \boldsymbol{C}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{C}^\top$.

Exemplo - demanda por gasolina

$$\begin{aligned} \ln(\text{G/Pop})_t &= \beta_1 + \gamma \ln(\text{G/Pop})_{t-1} + \beta_2 \ln \text{Preço}_t + \beta_3 \ln(\text{Renda/Pop})_t \\ &+ \beta_4 \ln(\text{PCN})_t + \beta_5 \ln(\text{PCU})_t + \varepsilon_t, \end{aligned}$$

sendo PCN o preço dos carros novos, PCU o preço dos carros usados e G o consumo de gasolina (gasto total (`GasExp`) divido pelo preço (`GasP`)).

- A elasticidade a longo prazo do preço é $\epsilon_2 = \dfrac{\beta_2}{1-\gamma}$.
- A elasticidade a longo prazo da renda é $\epsilon_3 = \dfrac{\beta_3}{1-\gamma}$.

Utilizaremos mínimos quadrados para estimar os parâmetros do modelo.

Ver arquivo exemplo_10.r

Sejam $\boldsymbol{x}_t = (1, \ln \text{Preço}_t, \ln(\text{Renda/Pop})_t, \ln(\text{PCN})_t, \ln(\text{PCU})_t)^\top$, $y_t = \ln(\text{G/Pop})_t$ e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_5)^\top$.

Então, podemos reescrever o modelo como:

$$\begin{aligned} y_t &= \boldsymbol{x}_t^\top \boldsymbol{\beta} + \gamma y_{t-1} + \varepsilon_t \\ y_t - \frac{1}{1-\gamma}\boldsymbol{x}_t^\top \boldsymbol{\beta} &= \gamma\left(y_{t-1} - \frac{1}{1-\gamma}\boldsymbol{x}_t^\top \boldsymbol{\beta}\right) + \varepsilon_t \\ y_t - \mu_t &= \gamma(y_{t-1} - \mu_t) + \varepsilon_t \\ z_t &= \gamma z_{t-1} + \varepsilon_t, \end{aligned}$$

sendo $z_t = y_t - \mu_t$, $z_{t-1} = y_{t-1} - \mu_t$ e $\mu_t = \dfrac{1}{1-\gamma}\boldsymbol{x}_t^\top \boldsymbol{\beta}$.

Temos, portanto, um modelo autoregressivo de ordem 1 (AR(1)) estacionário (por hipótese pois estaremos removendo todas as tendências com os regressores em $\boldsymbol{x}_t$).

Da estacionariedade do modelo AR(1), temos que

$\mathrm{E}(z_t) = \gamma \mathrm{E}(z_{t-1}) + \mathrm{E}(\varepsilon) \Rightarrow \mathrm{E}(z_t) = 0 \Rightarrow \mathrm{E}(y_t) = \mu_t$ (média de longo prazo).

Agora, podemos calcular as elasticidade de longo prazo do consumo em relação ao preço ou a renda.

A elasticidade a longo prazo do consumo em relação ao preço:

$$\epsilon_2 = \frac{\partial \ln(\text{G/Pop})_t}{\partial \ln \text{Preço}_t} = \frac{\partial \mu_t}{\partial \ln \text{Preço}_t} = \frac{\beta_2}{1-\gamma}.$$

A elasticidade a longo prazo do consumo em relação a renda:

$$\epsilon_3 = \frac{\partial \ln(\text{G/Pop})_t}{\partial \ln(\text{Renda/Pop})_t} = \frac{\partial \mu_t}{\partial \ln(\text{Renda/Pop})_t} = \frac{\beta_3}{1-\gamma}.$$

Seja $\boldsymbol{\theta} = (\beta_1, \gamma, \beta_2, \ldots, \beta_5)^\top$. Então,

$$\frac{\partial \epsilon_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(0, \frac{\beta_2}{(1-\gamma)^2}, \frac{1}{1-\gamma}, 0, 0, 0\right)^\top$$

$$\frac{\partial \epsilon_3(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(0, \frac{\beta_3}{(1-\gamma)^2}, 0, \frac{1}{1-\gamma}, 0, 0\right)^\top.$$

Estimação por máxima verossimilhança

Suponha **HP.6** seja verdade. Então, o estimador dos coeficientes da regressão por mínimos quadrados coincide com o estimador dos coeficientes por máxima verossimilhança.

O logaritmo da função de verossimilhança baseada em uma amostra aleatória de tamanho n com erros normais é dada por

$$\ell(\boldsymbol{\beta}, \sigma^2) = \ln L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

O estimador de máxima verossimilhança de $(\boldsymbol{\beta}, \sigma^2)$ é dado por

$$\begin{aligned} \boldsymbol{b}_{EMV} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}, \quad \text{e} \\ \widehat{\sigma}^2_{EMV} &= \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = \frac{1}{n}\boldsymbol{e}^\top\boldsymbol{e}. \end{aligned}$$

Por ser um estimador de máxima verossimilhança, $\boldsymbol{b}_{EMV}$ é assintoticamente eficiente entre os estimadores consistentes, e assintoticamente normal.

Veremos estimadores de máxima verossimilhança mais adiante.

Estimação por intervalo

Começaremos utilizando todas as hipóteses do modelo de regressão linear, inclusive **HP.6**.

Dado $\varepsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$, temos que

$$\begin{aligned} (\boldsymbol{b}|\boldsymbol{X}) &\sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}), \quad \text{e} \\ (b_k|\boldsymbol{X}) &\sim \mathcal{N}(\beta_k, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{kk}). \end{aligned}$$

Definamos s_{kk} o k-ésimo elemento da diagonal de $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

Daí, temos

$$Z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 s_{kk}}} \sim \mathcal{N}(0, 1).$$

Mas σ^2 é desconhecido.

Resultado 1

Os autovalores de uma matriz idempotente são compostos de 0 (zeros) ou 1 (uns).

Resultado 2

O posto de uma matriz idempotente é igual ao seu traço (um número inteiro).

Resultado 3

A única matriz idempotente não singular é a matriz identidade.

Resultado 4

Se $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ e $\boldsymbol{C}$ é uma matriz tal que $\boldsymbol{C}^\top \boldsymbol{C} = \boldsymbol{I}$, então $\boldsymbol{C}^\top \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Resultado 5

Se $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ e $\boldsymbol{A}$ é uma matriz idempotente, então $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \sim \chi^2_m$ sendo m o número de raizes unitárias de $\boldsymbol{A}$ (que é igual ao posto de $\boldsymbol{A}$).

Resultado 6

Se $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{L}\boldsymbol{x}$ uma função linear e $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ uma forma quadrática idempotente, então $\boldsymbol{L}\boldsymbol{x}$ e $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ são independentes se $\boldsymbol{L}\boldsymbol{A} = \boldsymbol{0}$.

A quantidade

$$\frac{(n-p)S^2}{\sigma^2} = \frac{\boldsymbol{e}^\top \boldsymbol{e}}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)^\top \boldsymbol{M} \left(\frac{\varepsilon}{\sigma}\right)$$

é uma forma quadrática idempotente do vetor normal (ε/σ) com $\text{tr}(\boldsymbol{M}) = n - p$ e a forma linear dada por

$$\left(\frac{\boldsymbol{b} - \beta}{\sigma}\right) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \left(\frac{\varepsilon}{\sigma}\right).$$

Temos que $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{M} = \boldsymbol{0}$ pois $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$. Daí, temos que a independência.

Teorema (independência de $\boldsymbol{b}$ e s^2)

Se ε é um vetor com distribuição normal, então o estimador $\boldsymbol{b}$ de mínimos quadrados dos coeficientes é estatisticamente independente do vetor de resíduos $\boldsymbol{e}$ e, portanto, todas as funções de $\boldsymbol{e}$, incluindo s^2.

Destes resultados, temos que

$$\frac{b_k - \beta_k}{\sqrt{s^2 s_{kk}}} \sim t_{n-p}.$$

Consequentemente,

$$\text{IC}_{100(1-\alpha)\%}(\beta_k) = (b_k - t_{n-p,1-\alpha/2}\sqrt{s^2 s_{kk}},\, b_k + t_{n-p,1-\alpha/2}\sqrt{s^2 s_{kk}}).$$

Exemplo (continuação) - demanda por Gasolina

$$\ln(\text{G/Pop})_t = \beta_1 + \beta_2 \ln \text{Preço}_t + \beta_3 \ln(\text{Renda/Pop})_t + \beta_4 \ln(\text{PCN})_t + \beta_5 \ln(\text{PCU})_t + \varepsilon_t.$$

Ver arquivo exemplo_10.r

Tabela 9: Intervalos de confiança de 95%.

Variável	2,5%	97,5%
Constante	-22,7264	-19,6958
ln Preço	-0,1093	0,0668
ln Renda/Pop	0,9395	1,2522
ln PCN	-0,6896	-0,0576
ln PCU	-0,1878	0,2279

Os intervalos de confiança para ϵ_2 e ϵ_3 baseados em teoria assintótica:

$$\text{IC}_{95\%}(\epsilon_2) = (-0,7099;\ -0,1129) \quad \text{e} \quad \text{IC}_{95\%}(\epsilon_3) = (0,6523;\ 1,2888).$$

Combinação linear dos coeficientes

- Seja $\boldsymbol{w}$ um vetor $p \times 1$ de constantes.
- Estamos interessados na combinação linear $\alpha = \boldsymbol{w}^\top \boldsymbol{\beta}$.
- Estimamos α com $a = \boldsymbol{w}^\top \boldsymbol{b}$ sendo $\boldsymbol{b}$ o estimador de mínimos quadrados.
- Como $\boldsymbol{b}$ tem distribuição normal, temos

$$a \sim \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{\beta}, \sigma^2 \boldsymbol{w}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{w}).$$

- Ainda estimamos σ^2 com s^2.
- Estimamos a variância de a com $s_a^2 = s^2 \boldsymbol{w}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{w}$.
- Segue-se que $(a - \alpha)/s_a \sim t_{n-p}$ (pois a e s_a^2 são independentes).
- Daí, podemos construir intervalos de confiança para α.
- Podemos testar somas ou diferenças entre os coeficientes.

Predição

- Suponha que desejamos predizer o valor de $y_\star$ associado ao vetor de regressores $\boldsymbol{x}_\star$.
- Este valor seria $y_\star = \boldsymbol{x}_\star^\top \boldsymbol{\beta} + \varepsilon_\star$.
- Segue do teorema de Gauss-Markov que $\widehat{y}_\star = \boldsymbol{x}_\star^\top \boldsymbol{b}$ é o estimador não tendencioso linear de variância mínima de $\mathrm{E}(y_\star|\boldsymbol{x}_\star)$.
- O erro de previsão é $e_\star = y_\star - \widehat{y}_\star = (\boldsymbol{\beta} - \boldsymbol{b})^\top \boldsymbol{x}_\star + \varepsilon_\star$.
- A variância de predição para este estimador é

$$\mathrm{Var}(e_\star|\boldsymbol{X}, \boldsymbol{x}_\star) = \sigma^2 + \mathrm{Var}((\boldsymbol{\beta} - \boldsymbol{b})^\top \boldsymbol{x}_\star|\boldsymbol{X}, \boldsymbol{x}_\star) = \sigma^2(1 + \boldsymbol{x}_\star^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_\star).$$

- Podemos estimar esta variância ao substituir σ^2 pelo estimador s^2.
- Assim, um intervalo de predição de $y_\star$ é dada por

$$\mathrm{IC}_{100(1-\alpha)\%}(y_\star) = (\widehat{y}_\star - t_{\alpha/2;n-p} \times \mathrm{ep}(e_\star), \widehat{y}_\star + t_{\alpha/2;n-p} \times \mathrm{ep}(e_\star)).$$

Exemplo

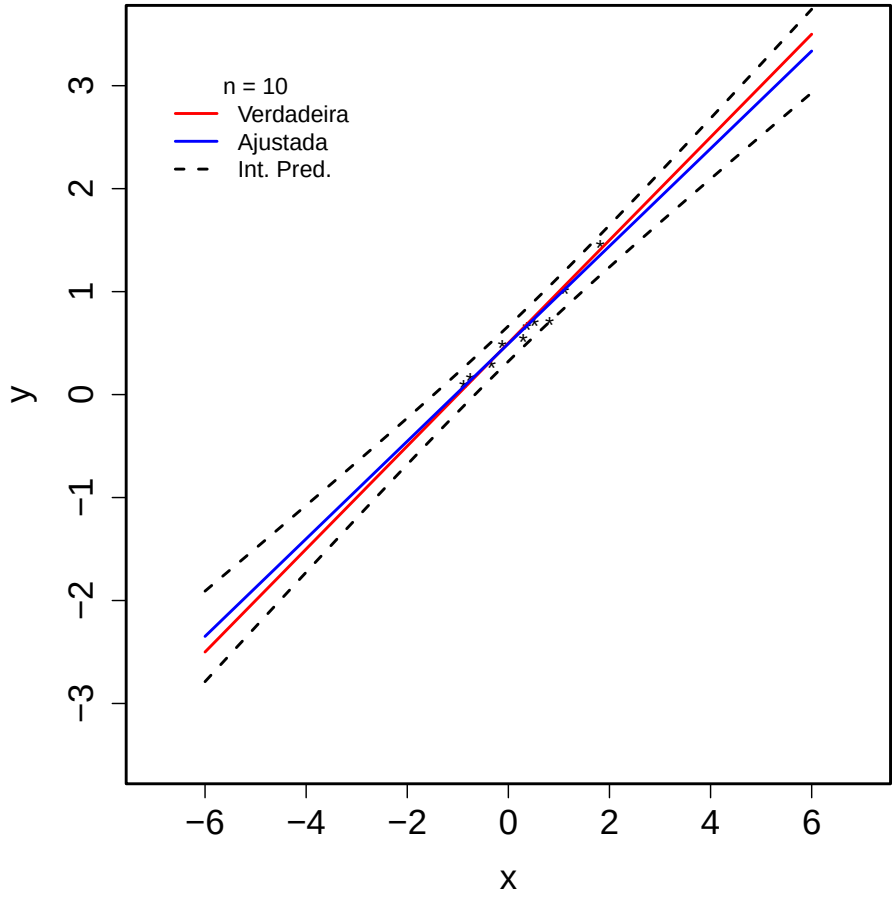


Figura 7: Intervalo de predição.

Ver arquivo exemplo_11.r

Multicolinearidade

Quando os regressores são altamente correlacionados, temos que

- pequenas mudanças nos dados podem produzir grandes variações na estimativas dos parâmetros;
- coeficientes podem ter erros padrões grandes e baixos níveis de significância mesmo que eles sejam conjuntamente significativos e o R^2 para a regressão seja elevado; e
- coeficientes podem ter um sinal equivocado ou magnitudes implausíveis.

Seja $\boldsymbol{X}$ a matriz de variáveis regressoras com uma constante e $p-1$ variáveis medidas como desvios de suas médias, $\boldsymbol{x}_k = \boldsymbol{z}_k - \overline{z}_k \mathbf{1}$.

Seja $\boldsymbol{X}_{(k)}$ a matriz com todas as variáveis exceto $\boldsymbol{x}_k$. Então, o k-ésimo elemento de $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ é dado por

$$\begin{aligned}
(\boldsymbol{x}_k^\top \boldsymbol{M}_{(k)} \boldsymbol{x}_k)^{-1} &= \left[\boldsymbol{x}_k^\top \boldsymbol{x}_k - \boldsymbol{x}_k^\top \boldsymbol{X}_{(k)} (\boldsymbol{X}_{(k)}^\top \boldsymbol{X}_{(k)})^{-1} \boldsymbol{X}_{(k)}^\top \boldsymbol{x}_k\right]^{-1} \\
&= \left[\boldsymbol{x}_k^\top \boldsymbol{x}_k \left(1 - \frac{\boldsymbol{x}_k^\top \boldsymbol{X}_{(k)} (\boldsymbol{X}_{(k)}^\top \boldsymbol{X}_{(k)})^{-1} \boldsymbol{X}_{(k)}^\top \boldsymbol{x}_k}{\boldsymbol{x}_k^\top \boldsymbol{x}_k}\right)\right]^{-1} \\
&= \frac{1}{(1 - R_k^2) s_{kk}},
\end{aligned}$$

sendo R_k^2 o coeficiente de determinação da regressão de x_k sobre todas as outras variáveis do modelo e $s_{kk} = \sum_{i=1}^n (x_{ik} - \overline{x}_k)^2$.

A variância do estimador de mínimos quadrados do k-ésimo coeficiente é σ^2 vezes a razão acima,

$$\text{Var}(b_k | \boldsymbol{X}) = \frac{\sigma^2}{(1 - R_k^2) \sum_{i=1}^n (x_{ik} - \overline{x}_k)^2}.$$

Com as outras coisas iguais,

- quanto maior a correlação de x_k com as outras variáveis, maior será a variância de b_k, devido a multicolinearidade;
- quanto maior a variância de x_k, menor será a variância de b_k; e
- quanto melhor o ajuste geral da regressão, menor será a variância devido a σ^2.

Dados reais nunca serão ortogonais ($R_k^2 = 0$). Então, a multicolinearidade estará sempre presente.

Quando a multicolinearidade é um problema?

Podemos utilizar o **número de condição** de $\boldsymbol{X}^\top \boldsymbol{X}$: a raiz quadrada da razão da maior raiz característica de $\boldsymbol{X}^\top \boldsymbol{X}$ (sendo as colunas padronizadas para tamanho unitário) sobre a menor. Valores maiores que 20 sugerem uma indicação de problema.

Possíveis caminhos são

- aumentar o número de observações;
- retirar as variáveis mais problemáticas do modelo; e
- utilizar componentes principais (pode haver dificuldade de interpretação).

Exemplo - dados de Longley

Tabela 10: Correlação entre as variáveis.

	Emprego	Preço	PIB	Militar	Ano
Emprego	1,0000	0,9709	0,9836	0,4573	0,9713
Preço	0,9709	1,0000	0,9916	0,4647	0,9911
PIB	0,9836	0,9916	1,0000	0,4464	0,9953
Militar	0,4573	0,4647	0,4464	1,0000	0,4172
Ano	0,9713	0,9911	0,9953	0,4172	1,0000

Ver arquivo exemplo_12.r

Tabela 11: Ajuste do modelo com **15** observações.

	Estimativa	Erro padrão	Estat. *t*	Valor *p*
Constante	1459415,1	714182,9	2,04348	0,06825
Ano	-721,756	369,985	-1,95077	0,07965
Preço	-181,123	135,525	-1,33646	0,21101
PIB	0,09107	0,02026	4,49478	0,00115
Militar	-0,07494	0,26113	-0,28698	0,77999

Tabela 12: Ajuste do modelo com **16** observações.

	Estimativa	Erro padrão	Estat. *t*	Valor *p*
Constante	1169087,5	835902,4	1,39859	0,18949
Ano	-576,464	433,487	-1,32983	0,21049
Preço	-19,7681	138,893	-0,14233	0,88940
PIB	0,06439	0,01995	3,22746	0,00805
Militar	-0,01015	0,30857	-0,03288	0,97436

Testes de hipóteses

Estamos interessados em implicações testáveis. Por exemplo, o modelo apresenta algumas restrições testáveis.

Suponha o seguinte modelo de investimento (I_t):

$$\mathcal{M}_1 : \quad \ln I_t = \beta_1 + \beta_2 \iota_t + \beta_3 \Delta p_t + \beta_4 \ln y_t + \beta_5 t + \varepsilon_t, \tag{1}$$

sendo

- ι_t: taxa de juros nominal;
- Δp_t: taxa de inflação; e
- $\ln y_t$: logaritmo da produção real.

Entretanto, uma teoria alternativa afirma que "investidores se preocupam com a taxa de juros real".

$$\ln I_t = \beta_1 + \beta_2(\iota_t - \Delta p_t) + \beta_3 \Delta p_t + \beta_4 \ln y_t + \beta_5 t + \varepsilon_t.$$

Não temos como testar a hipótese, pois ι_t e Δp_t estão em ambos os modelos.

Nota: $\iota_t - \Delta p_t$ é conhecida como a equação de Fisher.

Considere agora a alternativa em que "investidores se preocupam *somente* com a taxa de juros real

$$\mathcal{M}_2: \quad \ln I_t = \beta_1 + \beta_2(\iota_t - \Delta p_t) + \beta_4 \ln y_t + \beta_5 t + \varepsilon_t. \tag{2}$$

Agora, em relação ao modelo $\mathcal{M}_1$, temos a restrição $\beta_2 + \beta_3 = 0$.

Os modelos em (1) e (2) são encaixados.

Se os investidores não se preocuparem com a inflação (modelo $\mathcal{M}_3$), então $\beta_3 = 0$. Neste caso, o modelo $\mathcal{M}_3$ também seria encaixado com $\mathcal{M}_1$.

Então, suponha os seguintes modelos:

Modelo $\mathcal{M}_3$: Investidores se preocupam somente com a taxa de juros nominal ($\beta_1, \beta_2, 0, \beta_4, \beta_5$).

Modelo $\mathcal{M}_4$: Investidores se preocupam somente com a inflação ($\beta_1, 0, \beta_3, \beta_4, \beta_5$).

Estes dois modelos ($\mathcal{M}_3$ e $\mathcal{M}_4$) não são encaixados. Eles têm o mesmo número de parâmetros, mas diferentes covariáveis.

A seguir trataremos de modelos encaixados.

O modelo de regressão linear

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Restrições lineares:

$$\begin{aligned} r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1p}\beta_p &= q_1 \\ r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2p}\beta_p &= q_2 \\ \vdots \qquad\qquad\qquad\quad & \vdots \quad \vdots \\ r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{Jp}\beta_p &= q_J. \end{aligned}$$

Podemos organizar em notação matricial $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{q}$.

A matriz $\boldsymbol{R}$ tem J linhas e J restrições, então $J < p$. Estamos eliminando a possibilidade de $J = p$ para evitar solução degenerada. As linhas de $\boldsymbol{R}$ devem ser linearmente independentes.

A restrição $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{q}$ impõe J restrições em p parâmetros livres. Então, existem $p - J$ parâmetros livres.

Abordagem para testes de hipóteses

Utilizaremos a hipótese de normalidade **HP.6**.

Queremos testar:

$$H_0: \quad \boldsymbol{R}\beta - \boldsymbol{q} = \boldsymbol{0}$$
$$H_1: \quad \boldsymbol{R}\beta - \boldsymbol{q} \neq \boldsymbol{0}.$$

Exemplos:

1. $\beta_j = 0$, $\boldsymbol{R} = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$, $q = 0$;
2. $\beta_k = \beta_j$, $\boldsymbol{R} = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & -1 & 0 & \cdots & 0 \end{bmatrix}$, $q = 0$;
3. $\beta_2 + \beta_3 + \beta_4 = 1$, $\boldsymbol{R} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & \cdots & 0 \end{bmatrix}$, $q = 1$;
4. $\beta_2 + \beta_3 = 1$, $\beta_4 + \beta_6 = 0$, $\beta_5 + \beta_6 = 0$,

$$\boldsymbol{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad \text{e} \quad \boldsymbol{q} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix};$$

5. $\beta_2 = \beta_3 = \cdots = \beta_p = 0$, $\boldsymbol{R} = [\boldsymbol{0} : \boldsymbol{I}_{p-1}]$, $\boldsymbol{q} = \boldsymbol{0}$.

Dado $\boldsymbol{b}$, estamos interessados no vetor de discrepâncias $\boldsymbol{m} = \boldsymbol{Rb} - \boldsymbol{q}$.

Em termos estatísticos, $\boldsymbol{m}$ é significativamente diferente de $\mathbf{0}$?

Temos que
$\boldsymbol{b} \sim \mathcal{N}(\beta, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}) \quad \Rightarrow \quad \boldsymbol{m} \sim \mathcal{N}(\boldsymbol{R}\beta - \boldsymbol{q}, \sigma^2 \boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top)$.

Sob H_0, $\boldsymbol{m} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top)$ e $\boldsymbol{R}\beta = \boldsymbol{q}$ tal que
$(1/\sigma)[\boldsymbol{Rb} - \boldsymbol{q}] = (1/\sigma)\boldsymbol{R}(\boldsymbol{b} - \beta) = \boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top(\varepsilon/\sigma) = \boldsymbol{D}(\varepsilon/\sigma)$.

Podemos testar a H_0 com o critério de Wald. Condicional a $\boldsymbol{X}$, temos

$$\begin{aligned} W &= \boldsymbol{m}^\top \left[\text{Var}(\boldsymbol{m}|\boldsymbol{X})\right]^{-1} \boldsymbol{m} \\ &= \frac{(\boldsymbol{Rb} - \boldsymbol{q})^\top \left[\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1} (\boldsymbol{Rb} - \boldsymbol{q})}{\sigma^2} \\ &= \frac{1}{J}\left(\frac{\varepsilon}{\sigma}\right)^\top \boldsymbol{D}^\top \left[\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1} \boldsymbol{D}\left(\frac{\varepsilon}{\sigma}\right), \end{aligned}$$

uma forma quadrática idempotente em (ε/σ) com

$$\text{tr}\left(\boldsymbol{D}^\top \left[\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1} \boldsymbol{D}\right) = \text{tr}\left(\left[\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1} \boldsymbol{D}\boldsymbol{D}^\top\right) = \text{tr}(\boldsymbol{I}_J) = J.$$

Logo, $W \sim \chi^2_J$.

Mas a estatística W não pode ser utilizada porque desconhecemos σ^2.

Podemos utilizar s^2 (regressão) para estimar σ^2.

Resultado 7
Dado $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, se $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ e $\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}$ são duas formas quadráticas idempotentes em $\boldsymbol{x}$, então $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ e $\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}$ são independentes se $\boldsymbol{AB} = \boldsymbol{0}$.

Note que dado $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, temos que $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ e $\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}$ têm distribuição qui-quadrado.

Sabemos que $\boldsymbol{b}$ e s^2 são independentes. Isto implica que W e s^2 são independentes. Logo,

$$\begin{aligned} F &= \frac{W/J}{(n-p)s^2/[\sigma^2(n-p)]} \\ &= \frac{(\boldsymbol{R}\boldsymbol{b}-\boldsymbol{q})^\top \left[\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1}(\boldsymbol{R}\boldsymbol{b}-\boldsymbol{q})}{Js^2} \sim \mathcal{F}_{J,n-p}. \end{aligned}$$

Teste de hipóteses com uma restrição

Agora, considere uma abordagem alternativa para testar a hipótese

$$H_0: \ r_1\beta_1 + r_2\beta_2 + \cdots + r_p\beta_p = \boldsymbol{r}^\top \boldsymbol{\beta} = q.$$

Sabemos que $\widehat{q} = \boldsymbol{r}^\top \boldsymbol{b}$ é o estimador de $q = \boldsymbol{r}^\top \boldsymbol{\beta}$ e que

$$\widehat{q} = \boldsymbol{r}^\top \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{r}^\top \boldsymbol{\beta}, \sigma^2 \boldsymbol{r}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{r}).$$

Logo,

$$t = \frac{\widehat{q} - q}{\text{ep}(\widehat{q})} \sim t_{n-p}, \quad \text{sendo} \quad \text{ep}(\widehat{q}) = s\sqrt{\boldsymbol{r}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{r}}.$$

Note que $t^2 \sim \mathcal{F}_{1,n-p}$. Portanto, os testes são equivalentes.

Exemplo 12 - investimento restrito

Analizaremos dados trimestrais dos EUA entre jan/1950 e dez/2000. O modelo é dado por

$$\ln I_t = \beta_1 + \beta_2 \iota_t + \beta_3 \Delta p_t + \beta_4 \ln y_t + \beta_5 t + \varepsilon_t.$$

Primeiro queremos testar $H_0 : \beta_2 + \beta_3 = 0$ (somente taxa de juros real).

- I_t é o investimento real (`Realinvs`);
- ι_t é a taxa de juros (`Tbilrate`);
- Δp_t é a inflação medida por variação no logaritmo do índice de preço ao consumidor (`CPI_U`); e
- $\ln y_t$ é o logaritmo do PIB (`Realgdp`).

Uma observação é “perdida" ao calcular Δp_t.

Ver arquivo exemplo_13.r

Tabela 13: Resumo do ajuste do modelo irrestrito.

	Estimativa	Erro padrão	Estat. t	Valor p
Constante	-9,134092	1,366459	-6,684	$< 10^{-10}$
Taxa de juros	-0,008598	0,003196	-2,691	0,00774
Inflação	0,003306	0,002337	1,415	0,15872
ln PIB	1,930156	0,183272	10,532	$< 10^{-15}$
Tempo	-0,005659	0,001488	-3,803	0,00019

Além disso, $\hat{\sigma} = 0,08618$ e $R^2 = 0,9798$.

Ainda, $F = 2.395$ com 4 e 198 graus de liberdade, e valor $p < 10^{-15}$.

Com os resultados do modelo irrestrito, temos

- $\widehat{q} = b_2 + b_3 = -0,008598 + 0,003306 = -0,00529$,
- $ep(\widehat{q}) = \sqrt{ep(b_2)^2 + ep(b_3)^2 + 2cov(b_2, b_3)} = \sqrt{0,00320^2 + 0,00234^2 + 2 * (-3,717e-06)} = 0,002878$, e
- $t = \widehat{q}/ep(\widehat{q}) = -1,838$.
- Como $|t| < t_{0,975;198} = 1,972$, a hipótese nula não é rejeitada.
- Podemos estimar o modelo supondo $\beta_2 = -\beta_3$. **(Próxima página.)**

Ver arquivo exemplo_13.r

Tabela 14: Resumo do ajuste do modelo restrito.

	Estimativa	Erro padrão	Estat. t	Valor p
Constante	-7,90716	1,20063	-6,59	$< 10^{-10}$
Taxa de juros real	-0,00443	0,00227	-1,95	0,0526
ln PIB	1,76406	0,16056	10,99	$< 10^{-15}$
Tempo	-0,00440	0,00133	-3,31	0,0011

Além disso, $\hat{\sigma} = 0,0867$ e $R^2 = 0,979$.

Ainda, $F = 3.150$ com 3 e 198 graus de liberdade, e valor $p < 10^{-15}$.

Agora, queremos testar (conjuntamente) a seguinte hipótese nula:

- $\beta_2 + \beta_3 = 0$ (taxa de juros real),
- $\beta_4 = 1$ (a propensidade marginal a investir é 1), e
- $\beta_5 = 0$ (não existe tendência temporal).

Ainda utilizando o ajuste no modelo irrestrito, temos

$$\boldsymbol{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{q} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{m} = \boldsymbol{Rb} - \boldsymbol{q} = \begin{bmatrix} -0,005292 \\ 0,930156 \\ -0,005659 \end{bmatrix} \quad \text{e}$$

$$\widehat{\text{Var}}(\boldsymbol{m}|\boldsymbol{X}) = \begin{bmatrix} 8,238e-06 & -0,0002586 & 1,956e-06 \\ -2,586e-04 & 0,0335887 & -2,718e-04 \\ 1,956e-06 & -0,0002718 & 2,214e-06 \end{bmatrix}.$$

Assim, $F = \dfrac{\boldsymbol{m}^\top \left[\widehat{\text{Var}}(\boldsymbol{m}|\boldsymbol{X})\right]^{-1} \boldsymbol{m}}{J} = 109,8$, sendo $J = 3$.

Como $F > F_{3;\,198;\,0,95} = 2,65$, a hipótese nula é rejeitada (valor $p \approx 0$).

Predição para investimento

Queremos predizer o primeiro trimestre de 2001 ao nível $\alpha = 0,05$.

- Taxa de juros: 4,48 (`90-day T-Bill`);
- Inflação: $528,0$ (`CPI_U`);
- PIB: $9316,8$ (`realGDP`); e
- Tempo: 204.

Ver arquivo exemplo_13.r

O intervalor de predição é dado por:

$$IC_{95\%}(\hat{y}_{\star,204}) = (7,159;\ 7,503).$$

A taxa anual de investimento real no primeiro trimestre de 2001 foi 1.721.

O logaritmo é 7,4507. Assim, a predição contém este valor.

Estimadores de mínimos quadrados restritos

- Sabemos que $\boldsymbol{b}$ é escolhido para minimizar $\boldsymbol{e}^\top \boldsymbol{e}$.
- Sendo $R^2 = 1 - \dfrac{\boldsymbol{e}^\top \boldsymbol{e}}{\boldsymbol{y}^\top \widetilde{\boldsymbol{M}} \boldsymbol{y}}$ e $\boldsymbol{y}^\top \widetilde{\boldsymbol{M}} \boldsymbol{y}$ é uma constante que não envolve $\boldsymbol{b}$, então $\boldsymbol{b}$ é escolhido para maximizar R^2.
- O estimador de mínimos quadrados restrito é obtido pela solução de

$$\min_{S(\boldsymbol{b})} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \quad \text{sujeito a} \quad \boldsymbol{R}\boldsymbol{b} = \boldsymbol{q}.$$

- Uma função Lagrangeana para este problema pode ser escrita como

$$L(\boldsymbol{b}_\star, \boldsymbol{\lambda}_\star) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_\star)^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_\star) + 2\boldsymbol{\lambda}_\star^\top (\boldsymbol{R}\boldsymbol{b}_\star - \boldsymbol{q}).$$

- Como $\boldsymbol{\lambda}$ é irrestrito, utilizamos $2\boldsymbol{\lambda}$ como restrição.
- A solução de $\boldsymbol{b}_\star$ e $\boldsymbol{\lambda}_\star$ satisfazem as condições necessárias:

$$\begin{aligned} \frac{\partial L(\boldsymbol{b}_\star, \boldsymbol{\lambda}_\star)}{\partial \boldsymbol{b}_\star} &= -2\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_\star) + 2\boldsymbol{R}^\top \boldsymbol{\lambda} = \boldsymbol{0} \\ \frac{\partial L(\boldsymbol{b}_\star, \boldsymbol{\lambda}_\star)}{\partial \boldsymbol{\lambda}_\star} &= 2(\boldsymbol{R}\boldsymbol{b}_\star - \boldsymbol{q}) = \boldsymbol{0}. \end{aligned}$$

$$\begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{R}^\top \\ \boldsymbol{R} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_\star \\ \lambda_\star \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{y} \\ \boldsymbol{q} \end{bmatrix} \quad \text{ou} \quad \boldsymbol{A}\boldsymbol{d}_\star = \boldsymbol{v}.$$

Se a matriz $\boldsymbol{A}$ for não singular, então $\boldsymbol{d}_\star = \boldsymbol{A}^{-1}\boldsymbol{v}$.

Se $\boldsymbol{X}^\top \boldsymbol{X}$ for não singular, então uma solução explícita para $\boldsymbol{b}_\star$ e $\lambda_\star$ pode ser obtida usando a fórmula da inversa particionada:

$$\begin{aligned} \boldsymbol{b}_\star &= \boldsymbol{b} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top \left[\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1} (\boldsymbol{R}\boldsymbol{b} - \boldsymbol{q}) = \boldsymbol{b} - \boldsymbol{C}\boldsymbol{m} \quad \text{e} \\ \lambda_\star &= \left[\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1} (\boldsymbol{R}\boldsymbol{b} - \boldsymbol{q}). \end{aligned}$$

Além disso,

$$\text{Var}(\boldsymbol{b}_\star | \boldsymbol{X}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} - \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top \left[\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1} \boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

(A prova está na próxima página.)

Logo, $\text{Var}(\boldsymbol{b}_\star | \boldsymbol{X})$ é a variância de $\boldsymbol{b}$ reduzida pelas restrições.

Prova:

$$\begin{aligned}
\text{Var}(\boldsymbol{b}_\star|\boldsymbol{X}) &= \text{Var}(\boldsymbol{b}|\boldsymbol{X}) + \text{Var}(\boldsymbol{Cm}|\boldsymbol{X}) - 2\text{Cov}(\boldsymbol{b}, \boldsymbol{Cm}|\boldsymbol{X}) \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \boldsymbol{C}\text{Var}(\boldsymbol{Rb}-\boldsymbol{q}|\boldsymbol{X})\boldsymbol{C}^\top - 2\text{Cov}(\boldsymbol{b}, \boldsymbol{Rb}-\boldsymbol{q}|\boldsymbol{X})\boldsymbol{C}^\top \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \boldsymbol{CR}\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{C}^\top - 2\text{Cov}(\boldsymbol{b}, \boldsymbol{b}|\boldsymbol{X})\boldsymbol{R}^\top\boldsymbol{C}^\top \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \sigma^2\boldsymbol{CR}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{C}^\top - 2\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{C}^\top.
\end{aligned}$$

Note que

$$\begin{aligned}
\boldsymbol{CR}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\left[\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top.
\end{aligned}$$

Daí, temos que

$$\begin{aligned}
\text{Var}(\boldsymbol{b}_\star|\boldsymbol{X}) &= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{C}^\top - 2\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{C}^\top \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} - \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{C}^\top \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} - \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\left[\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.
\end{aligned}$$

A solução restrita de mínimos quadrados não pode ser melhor que a irrestrita:

$$\begin{aligned} \boldsymbol{e}_\star &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_\star = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{b}_\star = \boldsymbol{e} - \boldsymbol{X}(\boldsymbol{b}_\star - \boldsymbol{b}) \quad \Rightarrow \\ \boldsymbol{e}_\star^\top \boldsymbol{e}_\star &= \boldsymbol{e}^\top \boldsymbol{e} + (\boldsymbol{b}_\star - \boldsymbol{b})^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{b}_\star - \boldsymbol{b}) \geqslant \boldsymbol{e}^\top \boldsymbol{e} \quad \Rightarrow \\ R_\star^2 &\leqslant R^2. \end{aligned}$$

A perda de ajuste é dada por

$$\boldsymbol{e}_\star^\top \boldsymbol{e}_\star - \boldsymbol{e}^\top \boldsymbol{e} = (\boldsymbol{R}\boldsymbol{b} - \boldsymbol{q})^\top \left[\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1} (\boldsymbol{R}\boldsymbol{b} - \boldsymbol{q}).$$

Esta é a expressão no numerador do teste F:

$$\begin{aligned} F &= \frac{(\boldsymbol{R}\boldsymbol{b} - \boldsymbol{q})^\top \left[\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top\right]^{-1} (\boldsymbol{R}\boldsymbol{b} - \boldsymbol{q})}{Js^2} = \frac{(\boldsymbol{e}_\star^\top \boldsymbol{e}_\star - \boldsymbol{e}^\top \boldsymbol{e})/J}{\boldsymbol{e}^\top \boldsymbol{e}/(n-p)} \\ &= \frac{(R^2 - R_\star^2)/J}{(1-R^2)/(n-p)} \sim \mathcal{F}_{J,n-p}. \end{aligned}$$

Se a restrição for $\beta_2 = \cdots = \beta_p = 0$ (exceto intercepto), então temos $p-1$ restrições e o **teste** $\mathcal{F}$ usual de significância da regressão.

Exemplo - função de produção

Função de produção de Cobb-Douglas para a indústria de metal.

Regressão de mínimos quadrados no logaritmo da produção (valor adicionado) sobre uma constante, logaritmo do trabalho e produção de capital.

Uma generalização do modelo de Cobb-Douglas é o modelo **translog**:

$$\begin{aligned} \ln y &= \beta_1 + \beta_2 \ln \mathrm{TR} + \beta_3 \ln \mathrm{CP} \\ &+ \beta_4 \left[\frac{1}{2}(\ln \mathrm{TR})^2\right] + \beta_5 \left[\frac{1}{2}(\ln \mathrm{CP})^2\right] + \beta_6 \ln \mathrm{TR} \times \ln \mathrm{CP} + \varepsilon, \end{aligned}$$

sendo TR a variável trabalho e CP a variável capital.

Veremos depois que este modelo relaxa a hipótese de elasticidade unitária de substituição. O modelo de Cobb-Douglas é obtido com $\beta_4 = \beta_5 = \beta_6 = 0$.

Sem **HP.6 (normalidade)** os testes podem ser aproximados para amostras grandes.

Ver arquivo exemplo_14.r

Tabela 15: Resumo do ajuste do modelo translog.

	Estimativa	Erro padrão	Estat. t	Valor p
Constante	0,94420	2,91075	0,324	0,7489
ln TR	3,61364	1,54807	2,334	0,0296
ln CP	-1,89311	1,01626	-1,863	0,0765
$(\ln \text{TR})^2/2$	-0,96405	0,70738	-1,363	0,1874
$(\ln \text{CP})^2/2$	0,08529	0,29261	0,291	0,7735
$(\ln \text{TR})(\ln \text{CP})$	0,31239	0,43893	0,712	0,4845

Além disso, $\hat{\sigma} = 0,1799$ e $R^2 = 0,9549$.

Ainda, $F = 88,85$ com 5 e 21 graus de liberdade, e valor $p < 10^{-12}$.

Note o valor da estatística do **teste** $\mathcal{F}$ e seu valor p. Os demais resultados e outros testes estão descritos no arquivo R.

Restrições não lineares

Trabalharemos com uma restrição e o teste do tipo $H_0: \ c(\beta) = q$.

A estatística de teste é dada por

$$Z = \frac{c(\boldsymbol{b}) - q}{\text{ep}(c(\boldsymbol{b}))} \approx \mathcal{N}(0,1), \quad \text{sendo} \quad c(\boldsymbol{b}) \approx c(\beta) + \left(\frac{\partial c(\beta)}{\partial \beta}\right)^{\top} (\boldsymbol{b} - \beta).$$

O resultado é baseado em consistência e não em não tendeciosidade visto que, em geral, $\text{E}(g(X)) \neq g\left(\text{E}(X)\right)$. Precisamos estimar o erro padrão do estimador.

$$\text{Var}(c(\boldsymbol{b})) \approx \left(\frac{\partial c(\beta)}{\partial \beta}\right)^{\top} \text{Var}(\boldsymbol{b}|\boldsymbol{X}) \left(\frac{\partial c(\beta)}{\partial \beta}\right),$$

usando $\boldsymbol{h}(\boldsymbol{b})$ para estimar $\boldsymbol{h}(\beta) = \dfrac{\partial c(\beta)}{\partial \beta}$.

Assim,

$$\text{ep}(c(\boldsymbol{b})) = s\sqrt{\boldsymbol{h}(\boldsymbol{b})^{\top}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{h}(\boldsymbol{b})}.$$

Exemplo - propensidade marginal de longo prazo a consumir

Dados trimestrais de consumo agregados dos EUA (c_t) e renda pessoal (y_t) de 1950 a 2000. O modelo é dado por:

$$\ln c_t = \alpha + \beta \ln y_t + \gamma \ln c_{t-1} + \varepsilon_t.$$

A propensidade marginal a consumir (PMC) a curto prazo (elasticidade, pois as variáveis estão no logaritmo) é β.

A PMC de longo prazo é $\delta = \dfrac{\beta}{1-\gamma}$.

Considere o teste $H_0 : \ \delta = 1$. Temos,

$$h_1(\alpha, \beta, \gamma) = \frac{\partial \delta}{\partial \beta} = \frac{1}{1-\gamma} \quad \text{e} \quad h_2(\alpha, \beta, \gamma) = \frac{\partial \delta}{\partial \gamma} = \frac{\beta}{(1-\gamma)^2}.$$

Ver arquivo exemplo_15.r

Tabela 16: Resumo do ajuste do modelo de consumo.

	Estimativa	Erro padrão	Estat. t	Valor p
Constante	0,003142	0,010553	0,298	0,76624
$\ln(y_t)$	0,074958	0,028727	2,609	0,00976
$\ln(c_{t-1})$	0,924625	0,028594	32,337	$< 10^{-15}$

Além disso, $\hat{\sigma} = 0,008742$ e $R^2 = 0,9997$.

Ainda, $F = 3,476 \times 10^5$ com 2 e 200 graus de liberdade, e valor $p < 10^{-15}$.

Para o teste $H_0 : \delta = 1$, temos que

$$\hat{\delta} = 0,9945, \quad \text{ep}(\hat{\delta}) = 0,01636, \quad z_{obs} = 0,3386 \quad \text{e} \quad z_{0,975} = 1,96.$$

Como $|z_{obs}| < 1,96$, então não rejeitamos a hipótese nula que $\delta = 1$.

Critérios de seleção de modelos

Definição

A função de verossimilhança de θ é a função que associa o valor $f_{\boldsymbol{y}}(\boldsymbol{y}|\theta)$ para cada θ é definida da seguinte maneira

$$\begin{aligned} L(.;\boldsymbol{y}) : \boldsymbol{\Theta} &\rightarrow \mathbb{R}^{+} \\ \boldsymbol{\theta} &\rightarrow L(\boldsymbol{\theta};\boldsymbol{y}) = f_{\boldsymbol{y}}(\boldsymbol{y}|\boldsymbol{\theta}). \end{aligned}$$

Definição

O critério de informação de Akaike (AIC) e o critério de informação bayesiana (BIC) para o modelo $\mathcal{M}_j$ são dados por

$$\begin{aligned} AIC(\mathcal{M}_j) &= -2\ln L(\widehat{\boldsymbol{\theta}}_j;\boldsymbol{y}) + 2p_j \quad \text{e} \\ BIC(\mathcal{M}_j) &= -2\ln L(\widehat{\boldsymbol{\theta}}_j;\boldsymbol{y}) + p_j \ln(n), \end{aligned}$$

sendo p_j o número de parâmetros e $\widehat{\boldsymbol{\theta}}_j$ a estimativa de máxima verossimilhança do modelo $\mathcal{M}_j$.

Quanto menor o AIC (ou BIC), melhor o modelo ajustado.

Seleção bayesiana de modelos

Temos M modelos, $m = 1, 2, \ldots, M$, $f_m(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\theta}_m)$, sendo $f_m(.)$ a densidade conjunta, $\boldsymbol{y}$ e $\boldsymbol{Z}$ os dados, e $\boldsymbol{\theta}_m$ o vetor de parâmetros do $m-$ésimo modelo.

Suponha que $m^\star$ seja o modelo correto, desconhecido pelo analista, mas que seja um dos M modelos.

Seja $g(m) = \pi_m$ a probabilidade a priori do modelo m. Assim, a probabilidade a posteriori do m-ésimo modelo é dada por

$$\omega_m = g(m|\boldsymbol{y}, \boldsymbol{Z}) = \frac{f_m(\boldsymbol{y}|\boldsymbol{Z}, m)\pi_m}{\sum_{r=1}^{M} f_r(\boldsymbol{y}|\boldsymbol{Z}, r)\pi_r},$$

sendo $f_m(\boldsymbol{y}|\boldsymbol{Z}, m)$ a função de verossimilhança marginal do modelo m,

$$f_m(\boldsymbol{y}|\boldsymbol{Z}, m) = \int f_m(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\theta}_m, m) f_m(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m.$$

Seja $\widehat{\boldsymbol{\theta}}_m$ uma estimativa bayesiana, em geral a média a posteriori, dos parâmetros do modelo m. Então, uma possível aproximação é dada por

$$f_m(\boldsymbol{y}|\boldsymbol{Z}, m) \simeq \hat{f}_m(\boldsymbol{y}|\boldsymbol{Z}, m) = f_m(\boldsymbol{y}|\boldsymbol{Z}, \widehat{\boldsymbol{\theta}}_m, m) f_m(\widehat{\boldsymbol{\theta}}_m).$$

Daí, a densidade de previsão média do modelo bayesiano pode ser aproximada por

$$\bar{f} = \sum_{r=1}^{M} f_r(\boldsymbol{y}|\boldsymbol{Z}, \widehat{\boldsymbol{\theta}}_r, r) f_r(\widehat{\boldsymbol{\theta}}_r).$$

Assim,

$$\hat{\omega}_m = \frac{\hat{f}_m(\boldsymbol{y}|\boldsymbol{Z}, m)}{\bar{f}}, \quad \text{para} \quad m = 1, 2, \ldots, M.$$

Voltaremos a este ponto quando tratarmos dos modelos bayesianos.

Forma funcional e mudança de estrutura

Analisaremos maneiras de mudar a estrutura da variável dependente - no contexto de regressão - utilizando covariáveis.

Dados dicotômicos

- Os dados dicotômicos (binário, *dummy*) são muito utilizados em análise de regressão.
- Em geral, utiliza-se estas variáveis em conjunto com outras contínuas.

Exemplo - rendimentos

- $T = 428$ participantes do mercado de trabalho formal;
- Uma equação de (semilog) ganhos:

  $$\ln(\text{rendimentos}) = \beta_1 + \beta_2 \text{idade} + \beta_3 \text{idade}^2 + \beta_4 \text{educação} + \beta_5 \text{filhos} + \varepsilon;$$

- rendimentos: salário por hora vezes horas trabalhadas;
- educação: anos de escola; e
- filhos: variável dicotômica indicando (1) a presença de filhos menores de 18 anos.

Ver arquivo exemplo_16.r

Tabela 17: Resumo do ajuste do modelo de rendimentos.

Parâmetro	Estimativa	Erro padrão	Estat. t	Valor p
β_1	3,24010	1,76743	1,833	0,06747
β_2	0,20056	0,08386	2,392	0,01721
β_3	-0,00231	0,00099	-2,345	0,01947
β_3	0,06748	0,02525	2,672	0,00782
β_5	-0,35122	0,14753	-2,380	0,01773

Além disso, $\hat{\sigma} = 1,19$ e $R^2 = 0,041$.

O valor $-0,3511952$ é considerado como um efeito extremamente grande (semilog).

Note que $\epsilon = \dfrac{\partial y/y}{\partial x/x} = \dfrac{\partial \ln y}{\partial \text{filhos}} \times \text{filhos} = \beta_k \text{filhos}$ (**é um modelo semilog**).

$1 - \exp(-0,3511952) = 0,2961537 \simeq 1/3$ (mulheres com filhos recebem quase 1/3 menos).

Possíveis aplicações de variáveis dicotômicas

- Estudo do efeito de tratamentos em algum tipo de resposta.
- Efeito de diploma universitário na renda ao longo da vida.
- Diferenças de sexo em relação ao salário na iniciativa privada.
- Modelo envolvendo uma variável dicotômica:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \delta d_i + \varepsilon_i.$$

- É prática comum incluir variáveis dicotômicas na regressão para modelar algo somente em uma observação.

  Uma variável dicotômica que toma valor 1 (um) somente para uma observação tem o efeito de excluir esta observação do cálculo dos coeficientes de mínimos quadrados e a variância do estimador (mas não o R^2).

- Diversas categorias (por exemplo, efeitos sazonais):

$$C_t = \beta_1 + \beta_2 x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \varepsilon_t,$$

  sendo x_t os ganhos.

- Uma formulação alternativa é dada por

$$C_t = \beta x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \delta_4 D_{t4} + \varepsilon_t.$$

- Diversos grupos de variáveis dicotômicas.

$$y_{it} = \alpha + \beta x_{it} + \delta_i + \theta_t + \varepsilon_{it}.$$

- Precisamos retirar uma dicotômica de cada grupo.

Exemplo - companhias aéreas

O estudo trata da eficiência na produção de serviços de companhias aéreas.

Temos observações de 6 firmas durante 15 anos.

O modelo é dado por

$$\begin{aligned} \ln C_{i,t} &= \beta_1 + \beta_2 \ln Q_{i,t} + \beta_3 (\ln Q_{i,t})^2 + \beta_4 \ln P_{i,t} + \beta_5 \text{Cargas}_{i,t} \\ &+ \sum_{t=1}^{14} \theta_t D_{i,t} + \sum_{i=1}^{5} \delta_i F_{i,t} + \varepsilon_{i,t} \end{aligned}$$

Ver arquivo exemplo_17.r

Em várias aplicações, variáveis dicotômicas são utilizadas para fatores qualitativos somente.

Suponha o interesse no seguinte modelo:

$$\text{renda} = \beta_1 + \beta_2 \text{idade} + \text{efeito de educação} + \varepsilon.$$

Ensino médio (EM), graduação (GR), mestrado (MS) e doutorado (DS).

Uma maneira **insatisfatória** de modelar é considerar a variável E igual a 0 para o primeiro grupo, 1 para o segundo, 2 para o terceiro e 3 para o quarto.

$$\text{renda} = \beta_1 + \beta_2 \text{idade} + \beta_3 E + \varepsilon.$$

Um modelo mais flexível seria um com variáveis dicotômicas:

$$\text{renda} = \beta_1 + \beta_2 \text{idade} + \delta_1 \text{GR} + \delta_2 \text{MS} + \delta_3 \text{DS} + \varepsilon.$$

$$\begin{aligned}
\text{Ensino Médio:} \quad & \text{E(renda}|\text{idade,EM)} && = \beta_1 + \beta_2 \text{idade}, \\
\text{Graduação:} \quad & \text{E(renda}|\text{idade,GR)} && = \beta_1 + \beta_2 \text{idade} + \delta_1, \\
\text{Mestrado:} \quad & \text{E(renda}|\text{idade,MS)} && = \beta_1 + \beta_2 \text{idade} + \delta_2, \\
\text{Doutorado:} \quad & \text{E(renda}|\text{idade,DS)} && = \beta_1 + \beta_2 \text{idade} + \delta_3.
\end{aligned}$$

Regressão *spline* (função definida segmentarmente por polinômios)

A função que desejamos estimar é

$$\begin{aligned} \text{E(renda|idade)} &= \alpha^0 + \beta^0 \text{idade}, \quad \text{se} \quad \text{idade} < 18, \\ &\quad \alpha^1 + \beta^1 \text{idade}, \quad \text{se} \quad 18 \leqslant \text{idade} < 22, \\ &\quad \alpha^2 + \beta^2 \text{idade}, \quad \text{se} \quad \text{idade} \geqslant 22. \end{aligned}$$

Os valores 18 e 22 são chamados de nós. Seja $d_1 = 1$, se idade$\geqslant t_1^\star$ e $d_2 = 1$, se idade$\geqslant t_2^\star$, sendo $t_1^\star = 18$ e $t_2^\star = 22$. Temos a combinação das equações dada por

$$\text{renda} = \beta_1 + \beta_2 \text{idade} + \gamma_1 d_1 + \delta_1 d_1 \text{idade} + \gamma_2 d_2 + \delta_2 d_2 \text{idade} + \varepsilon.$$

Os coeficientes angulares dos três segmentos são β_2, $\beta_2 + \delta_1$ e $\beta_2 + \delta_1 + \delta_2$.

Para fazer com que a função seja contínua, precisamos impor que

$$\begin{aligned} \beta_1 + \beta_2 t_1^\star &= (\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_1^\star \qquad \text{e} \\ (\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_2^\star &= (\beta_1 + \gamma_1 + \gamma_2) + (\beta_2 + \delta_1 + \delta_2) t_2^\star. \end{aligned}$$

Isto implica restrições lineares nos coeficientes:

$$\gamma_1 + \delta_1 t_1^\star = 0 \quad \text{and} \quad \gamma_2 + \delta_2 t_2^\star = 0.$$

Assim,

$$\text{renda} = \beta_1 + \beta_2 \text{idade} + \delta_1 d_1(\text{idade} - t_1^\star) + \delta_2 d_2(\text{idade} - t_2^\star) + \varepsilon.$$

Os estimadores de mínimos quadrados restritos são obtidos pela regressão múltipla:

$$\begin{array}{lcl} x_1 & = & \text{idade}, \\ x_2 & = & \text{idade} - 18, \quad \text{se} \quad \text{idade} \geqslant 18 \quad \text{e} \quad 0 \quad \text{caso contrário}, \\ x_3 & = & \text{idade} - 22, \quad \text{se} \quad \text{idade} \geqslant 22 \quad \text{e} \quad 0 \quad \text{caso contrário}. \end{array}$$

Podemos testar a hipótese nula $H_0 : \ \delta_1 = \delta_2 = 0$ que implica mesmo coeficiente angular.

Não linearidade nas variáveis

Seja $\boldsymbol{z} = \{z_1, z_2, \ldots, z_L\}$ um conjunto de L variáveis independentes.

Seja $f_1, f_2, \ldots, f_p$ um conjunto de p funções linearmente independentes de $\boldsymbol{z}$.

Seja $g(y)$ uma função observável de y.

$$\begin{aligned} g(y) &= \beta_1 f_1(\boldsymbol{z}) + \beta_2 f_2(\boldsymbol{z}) + \cdots + \beta_p f_p(\boldsymbol{z}) + \varepsilon \\ &= \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \\ &= \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon. \end{aligned}$$

Um modelo comum é o log-linear

$$\ln y = \ln \alpha + \sum\nolimits_k \beta_k \ln z_k + \varepsilon = \beta_1 + \sum\nolimits_k \beta_k x_k + \varepsilon.$$

Neste modelo, os coeficientes angulares são elasticidades:

$$\left(\frac{\partial y}{\partial x_k}\right)\left(\frac{x_k}{y}\right) = \frac{\partial \ln y}{\partial ln x_k} = \beta_k.$$

No modelo log-linear,

- Mudanças são medidas em termos proporcionais ou porcentagem.
- β_k mede a mudança percentual em y associada com a mudança de 1 porcento em x_k.

Termos de interação

Um modelo relacionando distância de frenagem D com velocidade S e humidade da pista W,

$$D = \beta_1 + \beta_2 S + \beta_3 W + \beta_4 SW + \varepsilon.$$

Neste modelo,

$$\frac{\partial \mathrm{E}(D|S,W)}{\partial S} = \beta_2 + \beta_4 W.$$

Isto implica que o efeito marginal da velocidade na distância de frenagem cresce quando a pista fica mais molhada.

$$\mathrm{Var}\left(\frac{\partial \widehat{\mathrm{E}}(D|S,W)}{\partial S}\right) = \mathrm{Var}(\widehat{\beta}_2) + W^2\mathrm{Var}(\widehat{\beta}_4) + 2W\mathrm{Cov}(\widehat{\beta}_2, \widehat{\beta}_4).$$

Podemos tentar identificar não linearidades:

- Análise de resíduos;
- Regressão linear por partes; e
- Regressão polinomial.

Definição de linearidade intrínseca

No modelo de regressão linear clássico, se os p parâmetros $\beta_1, \beta_2, \ldots, \beta_p$ podem ser escritos como p funções biunívocas, possivelmente não lineares, de um conjunto de p parâmetros $\theta_1, \theta_2, \ldots, \theta_p$, então o modelo é intrinsecamente linear em θ.

Exemplo - regressão linear intrínseca

Considere a estimação dos parâmetros do modelo

$$f(y|x, \rho, \beta) = \frac{(\beta + x)^{-\rho}}{\Gamma(\rho)} y^{\rho-1} \exp\{-y/(\beta + x)\}$$

por máxima verossimilhança.

Ver arquivo exemplo_18.r

- Neste modelo, $\mathrm{E}(y|x) = \beta\rho + \rho x$.
- Um modelo de regressão linear intrínseco seria $\mathrm{E}(y|x) = \beta_1 + \beta_2 x$.
- Podemos estimar β_1 e β_2 por mínimos quadrados.
- Podemos estimar também β por b_1/b_2.
- A estimação do erro padrão de β se dá através do método delta,

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = \left[\frac{\partial\widehat{\beta}}{\partial b_1}\right]^2 \widehat{\mathrm{Var}}(b_1) + \left[\frac{\partial\widehat{\beta}}{\partial b_2}\right]^2 \widehat{\mathrm{Var}}(b_2) + 2\left[\frac{\partial\widehat{\beta}}{\partial b_1}\right]\left[\frac{\partial\widehat{\beta}}{\partial b_2}\right] \widehat{\mathrm{Cov}}(b_1, b_2).$$

Exemplo - função de produção de elasticidade constante de substituição

O modelo é dado por

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln(\delta K^{-\rho} + (1-\delta)L^{-\rho}) + \varepsilon$$

A aproximação de Taylor em torno do ponto $\rho = 0$ é

$$\begin{aligned} \ln y &= \ln \gamma + \nu\delta \ln K + \nu(1-\delta) \ln L + \rho\nu\delta(1-\delta) \left\{ -\frac{1}{2} [\ln K - \ln L]^2 \right\} + \varepsilon^\star \\ &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon^\star, \end{aligned}$$

sendo $x_1 = 1$, $x_2 = \ln K$, $x_3 = \ln L$ e $x_4 = -(1/2)\ln^2(K/L)$.

As transformações são dadas por

$$\begin{aligned} \beta_1 &= \ln \gamma & \gamma &= \exp(\beta_1) \\ \beta_2 &= \nu\delta & \delta &= \beta_2/(\beta_2+\beta_3) \\ \beta_3 &= \nu(1-\delta) & \nu &= \beta_2 + \beta_3 \\ \beta_2 &= \rho\nu\delta(1-\delta) & \rho &= \beta_4(\beta_2+\beta_3)/(\beta_2\beta_3). \end{aligned}$$

Estimativas de β_1, β_2, β_3 e β_4 podem ser obtidas por mínimos quadrados.

Para estimar a matriz de covariância de $\boldsymbol{\theta} = (\gamma, \delta, \nu, \rho)$ utilizamos o método delta.

$$\boldsymbol{C} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}^\top} = \begin{bmatrix} \exp\{\beta_1\} & 0 & 0 & 0 \\ 0 & \dfrac{\beta_3}{(\beta_2+\beta_3)^2} & -\dfrac{\beta_2}{(\beta_2+\beta_3)^2} & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -\dfrac{\beta_3\beta_4}{\beta_2^2\beta_3} & -\dfrac{\beta_2\beta_4}{\beta_2\beta_3^2} & \dfrac{\beta_2+\beta_3}{\beta_2\beta_3} \end{bmatrix}.$$

A matriz de covariâncias de $\boldsymbol{\theta}$ é, então, estimada por

$$\widehat{\boldsymbol{C}} \left[s^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \right] \widehat{\boldsymbol{C}}^\top.$$

Vale ressaltar que nem todos os modelos são intrinsecamente lineares.

Mudança estrutural

- Suponha o modelo de regressão múltipla $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$.
- Queremos testar a hipótese de que $\beta_1 = \beta_2$ para dois conjuntos diferentes da mesma variável.
- Podemos pensar no modelo

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

- A solução é

$$\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \begin{bmatrix} \boldsymbol{X}_1^\top \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2^\top \boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_1 \boldsymbol{y}_1 \\ \boldsymbol{X}_2 \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}$$

- A soma total de quadrados é $\boldsymbol{e}^\top \boldsymbol{e} = \boldsymbol{e}_1^\top \boldsymbol{e}_1 + \boldsymbol{e}_2^\top \boldsymbol{e}_2$.
- Temos, $H_0 : \ \boldsymbol{R}\beta = \boldsymbol{q}$, sendo $\boldsymbol{R} = [\boldsymbol{I} \ \ -\boldsymbol{I}]$ e $\boldsymbol{q} = \boldsymbol{0}$.
- Prossegue-se com o teste F.
- O número de graus de liberdade é J e $n_1 + n_2 - 2p$. J é o número de colunas em $\boldsymbol{X}_2$.

Omissão de variáveis relevantes

Suponha que o modelo correto seja $\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \varepsilon$.

Se regredimos $\boldsymbol{y}$ sobre $\boldsymbol{X}_1$, excluindo $\boldsymbol{X}_2$, temos

$$\boldsymbol{b}_1 = (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top \boldsymbol{y} = \beta_1 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top \boldsymbol{X}_2\beta_2 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top \varepsilon$$

$$\mathrm{E}(\boldsymbol{b}_1|\boldsymbol{X}) = \beta_1 + \boldsymbol{P}_{1,2}\beta_2 \quad \text{e} \quad \mathrm{Var}(\boldsymbol{b}_1|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1},$$

sendo $P_{1,2} = (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top \boldsymbol{X}_2$. Cada coluna de $\boldsymbol{P}_{1,2}$ é o resultado da regressão das colunas de $\boldsymbol{X}_2$ sobre $\boldsymbol{X}_1$.

Com $\boldsymbol{X}_2$ incluído na regressão, teríamos

$$\mathrm{Var}(\boldsymbol{b}_{1,2}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}_1^\top \boldsymbol{M}_2\boldsymbol{X}_1)^{-1} = \sigma^2[\boldsymbol{X}_1^\top \boldsymbol{X}_1 - \boldsymbol{X}_1^\top \boldsymbol{X}_2(\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top \boldsymbol{X}_1]^{-1},$$

pois $\boldsymbol{M}_2 = \boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top$.

Podemos comparar as inversas das matrizes (precisão)

$$[\mathrm{Var}(\boldsymbol{b}_1|\boldsymbol{X})]^{-1} - [\mathrm{Var}(\boldsymbol{b}_{1,2}|\boldsymbol{X})]^{-1} = (1/\sigma^2)\boldsymbol{X}_1^\top \boldsymbol{X}_2(\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top \boldsymbol{X}_1,$$

que é não negativa definida.

Concluímos que apesar de $\boldsymbol{b}_1$ ser tendencioso, sua variância nunca é maior do que a de $\boldsymbol{b}_{1,2}$.

Modelo de regressão generalizado e heterocedasticidade

Trataremos de modelos que não satisfazem a hipótese **HP.4** de homocedasticidade e correlação nula.

O modelo de regressão linear generalizado é dado por

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) &= \mathbf{0} \\ \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top|\boldsymbol{X}) &= \sigma^2\boldsymbol{\Omega} = \boldsymbol{\Sigma}, \end{aligned}$$

em que $\boldsymbol{\Sigma}$ é uma matriz positiva definida.

No modelo (somente) heterocedástico:

$$\sigma^2\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}.$$

No modelo (somente) correlacionado:

$$\sigma^2\mathbf{\Omega} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}.$$

Em geral, aparece em modelos de séries temporais (modelo AR(1)).

Conjuntos de dados de painéis, consistindo de observações por seções cruzadas em vários pontos no tempo, podem exibir heterocedasticidade e correlação não nula.

Estimação por mínimos quadrados

Para o modelo com distúrbios esféricos,

$$\mathrm{E}(\varepsilon|\boldsymbol{X}) = \mathbf{0} \quad \text{e} \quad \mathrm{E}(\varepsilon\varepsilon^\top|\boldsymbol{X}) = \sigma^2\boldsymbol{I},$$

o estimador de mínimos quadrados ordinários (OLS) é

$$\boldsymbol{b} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \beta + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\varepsilon.$$

- o melhor estimador linear não tendencioso (BLUE);
- consistente;
- assintoticamente normal; e
- se os erros forem normais, então também é eficiente assintoticamente.

Teorema

Se os regressores e os distúrbios forem não correlacionados, então o estimador de mínimos quadrados para o modelo de regressão linear generalizado é não tendencioso. Condicional a $\boldsymbol{X}$, a variância amostral do estimador de mínimos quadrados é

$$\begin{aligned}
\text{Var}(\boldsymbol{b}|\boldsymbol{X}) &= \text{E}((\boldsymbol{b}-\beta)(\boldsymbol{b}-\beta)^\top|\boldsymbol{X}) \\
&= \text{E}((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\varepsilon\varepsilon^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}|\boldsymbol{X}) \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\sigma^2\Omega\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= \frac{\sigma^2}{n}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}^\top\Omega\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}
\end{aligned}$$

Se ε for normalmente distribuído, então

$$\boldsymbol{b} \sim \mathcal{N}(\beta, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\boldsymbol{X}^\top\Omega\boldsymbol{X}\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}).$$

- Não podemos mais estimar $\text{Var}(\boldsymbol{b}|\boldsymbol{X})$ com $s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$; e
- Os testes t e F já não são mais apropriados.

Propriedades assintóticas de mínimos quadrados

Teorema
Se $\boldsymbol{Q} = \text{plim}(\boldsymbol{X}^\top \boldsymbol{X}/n)$ e $\text{plim}(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}/n)$ são matrizes positivas definidas e finitas, então $\boldsymbol{b}$ é consistente para β. Sob estas hipóteses, $\text{plim}\boldsymbol{b} = \beta$.

A condição no teorema acima depende de $\boldsymbol{X}$ e de $\boldsymbol{\Omega}$.

Teorema
Se os regressores forem suficientemente bem comportados e os termos fora da diagonal da matriz $\boldsymbol{\Omega}$ diminuirem suficientemente rápido, então o estimador de mínimos quadrados é assintoticamente normal com média β e matriz de covariâncias

$$\text{Var}(\boldsymbol{b}) = \frac{\sigma^2}{n} \boldsymbol{Q}^{-1} \text{plim} \left(\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} \right) \boldsymbol{Q}^{-1}.$$

Estimação robusta das matrizes de covariâncias assintóticas

Se $\sigma^2\boldsymbol{\Omega}$ fosse conhecido, então o estimador da matriz de covariância assintótica de $\boldsymbol{b}$ seria

$$\boldsymbol{V}_{OLS} = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}^\top\left[\sigma^2\boldsymbol{\Omega}\right]\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}.$$

Nosso problema é $\sigma^2\boldsymbol{\Omega}$. Assumiremos que $\text{tr}(\boldsymbol{\Omega}) = n$ e $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Omega}$.

Em príncipio, parece que temos $n(n+1)/2$ parâmetros para estimar em $\boldsymbol{\Sigma}$.

Na verdade, precisamos estimar $p(p+1)/2$ parâmetros da matriz

$$\text{plim}\boldsymbol{Q}_\star = \text{plim}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}\boldsymbol{x}_i\boldsymbol{x}_j^\top.$$

Voltaremos a este ponto mais à frente.

Estimação eficiente por mínimos quadrados generalizados

Temos que $\boldsymbol{\Omega}$ é uma matriz positiva definida, logo

$$\boldsymbol{\Omega} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}^{\top},$$

sendo as colunas de $\boldsymbol{C}$ os vetores característicos de $\boldsymbol{\Omega}$ e $\boldsymbol{\Lambda}$ é uma matriz diagonal com as raizes características de $\boldsymbol{\Omega}$.

Seja $\boldsymbol{T} = \boldsymbol{C}\boldsymbol{\Lambda}^{1/2}$, então $\boldsymbol{\Omega} = \boldsymbol{T}\boldsymbol{T}^{\top}$.

Seja $\boldsymbol{P}^{\top} = \boldsymbol{C}\boldsymbol{\Lambda}^{-1/2}$, então $\boldsymbol{\Omega}^{-1} = \boldsymbol{P}^{\top}\boldsymbol{P}$.

Segue-se que

$$\boldsymbol{P}\boldsymbol{y} = \boldsymbol{P}\boldsymbol{X}\beta + \boldsymbol{P}\varepsilon \quad \text{ou} \quad \boldsymbol{y}_{\star} = \boldsymbol{X}_{\star}\beta + \varepsilon_{\star} \Rightarrow \mathrm{E}(\varepsilon_{\star}\varepsilon_{\star}^{\top}|\boldsymbol{X}_{\star}) = \boldsymbol{P}\sigma^2\boldsymbol{\Omega}\boldsymbol{P}^{\top} = \sigma^2\boldsymbol{I}.$$

Então, o modelo de regressão linear clássico se aplica a este modelo transformado.

Por hipótese, $\boldsymbol{\Omega}$ é conhecido, então $\boldsymbol{y}_\star$ e $\boldsymbol{X}_\star$ são os dados observados. Logo,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}_\star^\top \boldsymbol{X}_\star)^{-1}\boldsymbol{X}_\star^\top \boldsymbol{y}_\star = (\boldsymbol{X}^\top \boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{y} = (\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{y}$$

é um estimador eficiente de $\boldsymbol{\beta}$. Este é o estimador de mínimos quadrados generalizados (GLS).

Teorema

1. Se $\mathrm{E}(\boldsymbol{\varepsilon}_\star|\boldsymbol{X}_\star) = \mathbf{0} \Leftrightarrow \mathrm{E}(\boldsymbol{P}\boldsymbol{\varepsilon}|\boldsymbol{P}\boldsymbol{X}) = \mathbf{0} \Leftrightarrow \mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \mathbf{0}$, então $\mathrm{E}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}_\star) = \boldsymbol{\beta}$.
2. O estimador GLS é consistente se $\operatorname{plim}(1/n)\boldsymbol{X}_\star^\top \boldsymbol{X}_\star = \boldsymbol{Q}_\star$, sendo $\boldsymbol{Q}_\star$ uma matriz positiva definida. Logo, $\operatorname{plim}[(1/n)\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{X}]^{-1} = \boldsymbol{Q}_\star^{-1}$.
3. O estimador GLS é normalmente distribuído com média $\boldsymbol{\beta}$ e a variância amostral
$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}_\star) = \sigma^2(\boldsymbol{X}_\star^\top \boldsymbol{X}_\star)^{-1} = \sigma^2(\boldsymbol{X}^\top \boldsymbol{\Omega}\boldsymbol{X})^{-1}.$$
4. O estimador GLS $\widehat{\boldsymbol{\beta}}$ é o estimador linear não tendencioso de variância mínima no modelo de regressão generalizado.

Para testar J restrições lineares, $\boldsymbol{R}\beta = \boldsymbol{q}$, a estatística apropriada é

$$F_{J,n-p} = \frac{(\boldsymbol{R}\widehat{\beta} - \boldsymbol{q})^\top [\boldsymbol{R}\widehat{\sigma}^2(\boldsymbol{X}_\star^\top \boldsymbol{X}_\star)^{-1}\boldsymbol{R}^\top]^{-1}(\boldsymbol{R}\widehat{\beta} - \boldsymbol{q})}{J} = \frac{(\widehat{\varepsilon}_c^\top \widehat{\varepsilon}_c - \widehat{\varepsilon}^\top \widehat{\varepsilon})/J}{\widehat{\sigma}^2},$$

sendo $\widehat{\varepsilon} = \boldsymbol{y}_\star - \boldsymbol{X}_\star\widehat{\beta}$ e $\widehat{\sigma}^2 = \dfrac{\widehat{\varepsilon}^\top\widehat{\varepsilon}}{n-p} = \dfrac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\beta})^\top \Omega^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\beta})}{n-p}$.

Os resíduos do GLS restrito, $\widehat{\varepsilon}_c = \boldsymbol{y}_\star - \boldsymbol{X}_\star\widehat{\beta}_c$, são baseados em

$$\widehat{\beta}_c = \widehat{\beta} - [\boldsymbol{X}^\top\Omega^{-1}\boldsymbol{X}]^{-1}[\boldsymbol{R}(\boldsymbol{X}^\top\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top]^{-1}(\boldsymbol{R}\widehat{\beta} - \boldsymbol{q}).$$

Não existe nenhuma contrapartida precisa para o R^2 no modelo de regressão generalizado.

Se $\boldsymbol{\Omega}$ for desconhecido, então não é possível utilizar o estimador GLS.

Precisamos impor restrições em $\boldsymbol{\Omega}$ para diminuir o número efetivo de parâmetros a estimar.

Heterocedasticidade

Distúrbios da regressão cujas variâncias não são constantes para diferentes observações são heterocedásticos.

Em modelos de regressão heterocedásticos,

$$\text{Var}(\varepsilon_i|\boldsymbol{X}) = \sigma_i^2 = \sigma^2\omega_i, \quad i = 1, 2, \ldots, n \quad \text{ou}$$

$$\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix},$$

sendo $\text{tr}(\boldsymbol{\Omega}) = \sum_{i=1}^{n} \omega_i = n$ para evitar problemas de escala.

Ineficiência de mínimos quadrados

- $\boldsymbol{b}$, o estimador OLS, é ineficiente relativo ao estimador GLS.
- Quanto maior a dispersão de ω_i para diferentes observações, maior a eficiência do GLS sobre OLS.
- Suponha que a forma da heterocedasticidade seja desconhecida.
- O estimador GLS não pode ser usado.
- Se usarmos $\boldsymbol{b}$, então a matriz de covariância $\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ é inapropriada.
- A matriz $\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ é apropriada.
- Temos, usualmente,

$$s^2 = \frac{\boldsymbol{e}^\top\boldsymbol{e}}{n-p} = \frac{\varepsilon^\top\boldsymbol{M}\varepsilon}{n-p} = \frac{\varepsilon^\top\varepsilon}{n-p} - \frac{\varepsilon^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\varepsilon}{n-p},$$

pois $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Isto implica que

$$\mathrm{E}\left(\frac{\varepsilon^\top\varepsilon}{n-p}\middle|\boldsymbol{X}\right) = \frac{\mathrm{tr}(\mathrm{E}(\varepsilon\varepsilon^\top|\boldsymbol{X}))}{n-p} = \frac{n\sigma^2}{n-p} \quad \text{e}$$

$$
\begin{aligned}
\mathrm{E}\left(\frac{\varepsilon^\top \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\varepsilon}{n-p}\Big|\boldsymbol{X}\right) &= \frac{\mathrm{tr}(\mathrm{E}[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\varepsilon\varepsilon^\top\boldsymbol{X}])}{n-p} \\
&= \frac{\mathrm{tr}(\sigma^2(\boldsymbol{X}^\top\boldsymbol{X}/n)^{-1}(\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}/n))}{n-p} \\
&= \frac{\sigma^2}{n-p}\mathrm{tr}\left(\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right)^{-1}\boldsymbol{Q}_n^\star\right),
\end{aligned}
$$

sendo $\boldsymbol{Q}_n^\star = \dfrac{\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}}{n} = \dfrac{1}{n}\sum_{i=1}^{n}\omega_i\boldsymbol{x}_i\boldsymbol{x}_i^\top$.

Se $\boldsymbol{b}$ é consistente, então $\lim_{n\to\infty}\mathrm{E}(s^2) = \sigma^2$.

Se a heterocedasticidade não é correlacionada com as variáveis no modelo, então pelo menos em amostras grandes, os cálculos de mínimos quadrados, apesar de não serem a maneira ótima de utilizar os dados, não levarão a "erros grosseiros".

Observações:

- É verdade que o OLS é ineficiente.
- Se por hipótese da análise - que a heterocedasticidade não é relacionada as variáveis do modelo - é incorreta, então os erros padrões convencionais podem estar longe dos valores apropriados.

Estamos interessados em um estimador para $\boldsymbol{Q}_\star = \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^\top$.

Pode ser mostrado sob condições bem gerais que o estimador

$$\boldsymbol{S}_0 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

tem $\text{plim}\boldsymbol{S}_0 = \text{plim}\boldsymbol{Q}_\star$.

Assim, o estimador consistente de heterocedasticidade de **White** é

$$\widehat{\text{Var}}(\boldsymbol{b}|\boldsymbol{X}) = n(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{S}_0 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

Um dos problemas dos quadrados dos resíduos de OLS é que eles tendem a subestimar os quadrados dos distúrbios verdadeiros.

Davidson e McKinnon propuseram

1. escalar o resultado final por $n/(n-p)$; e
2. usar e_i^2/m_{ii} ao invés de e_i^2, sendo $m_{ii} = 1 - \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i$.

Exemplo - estimadores de White, e de Davidson e McKinnon

Ver arquivo exemplo_19.r

Modelaremos os gastos mensais com cartão de crédito (G):

$$G = \beta_1 + \beta_2 \text{idade} + \beta_3 \text{casa} + \beta_4 \text{renda} + \beta_5 \text{renda}^2 + \varepsilon.$$

A variável “casa” é uma binária indicando casa própria ou não.

Testando a heterocedasticidade

O teste geral de White é dado por $H_0: \ \sigma_i^2 = \sigma^2$ para todo i, e $H_1: \ \text{Não } H_0$.

A dificuldade é em estimar um modelo com n parâmetros.

A matriz de covariância correta para o estimador de mínimos quadrados é

$$\text{Var}(\boldsymbol{b}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$$

que pode ser estimada como anteriormente.

O estimador convencional é $\boldsymbol{V} = s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$.

Se não existir heterocedasticidade, então $\boldsymbol{V}$ dará um estimador consistente de $\text{Var}(\boldsymbol{b}|\boldsymbol{X})$.

Uma versão simples e operacional do teste de White é feita ao obter nR^2 na regressão de e_i^2 sobre uma constante, todas as variáveis em $\boldsymbol{x}_i$ e todos os quadrados e produtos cruzados de $\boldsymbol{x}_i$.

A estatística tem distribuição assintótica qui-quadrada com $r - 1$ graus de liberdade, sendo r o número de regressores na equação, incluindo a constante.

O teste de White não é construtivo. Se H_0 é rejeitada, então o resultado do teste não dá nenhuma indicação do que fazer a seguir.

O teste de **Breusch-Pagan** é baseado nos multiplicadores de Lagrange e tem $H_0: \ \sigma_i^2 = \sigma^2 f(\alpha_0 + \boldsymbol{\alpha}^\top \boldsymbol{z}_i)$, sendo $\boldsymbol{z}_i$ o vetor de variáveis independentes.

O modelo é homecedástico se $\boldsymbol{\alpha} = \mathbf{0}$.

O teste pode ser conduzido com uma regressão simples.

A estatística do teste é

$$LM = \frac{1}{2}\boldsymbol{g}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{g},$$

sendo $\boldsymbol{g}$ o vetor de observações $g_i = e_i^2/(\boldsymbol{e}^\top \boldsymbol{e}/n) - 1$.

Sob H_0, $LM \approx \chi_r^2$, sendo r o número de variáveis em $\boldsymbol{z}$.

Este teste é sensível a hipótese de normalidade.

Koenker e Basset sugeriram utilizar um estimador robusto para a variância de ε_i^2,

$$V = \frac{1}{n}\sum_{i=1}^{n}\left[e_i^2 - \frac{\boldsymbol{e}^\top \boldsymbol{e}}{n}\right]^2.$$

Assim,

$$LM = \frac{1}{V}(\boldsymbol{u} - \overline{u}\mathbf{1})^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top(\boldsymbol{u} - \overline{u}\mathbf{1}),$$

sendo $\boldsymbol{u} = (e_1^2, e_2^2, \ldots, e_n^2)$ e $\overline{u} = \boldsymbol{e}^\top \boldsymbol{e}/n$.

Ver arquivo exemplo_19.r

Mínimos quadrados ponderados

Tendo testado e encontrado evidência de heterocedasticidade, o próximo passo é levar isto em conta na estimação.

O estimador GLS é

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{y}.$$

Suponha que $\text{Var}(\varepsilon_i | \boldsymbol{X}) = \sigma_i^2 = \sigma^2 \omega_i$.

Então, o i-ésimo elemento da diagonal da matriz $\boldsymbol{\Omega}^{-1}$ é $1/\omega_i$.

O estimador GLS é obtido ao regredir

$$\boldsymbol{Py} = \begin{bmatrix} y_1/\sqrt{\omega_1} \\ y_2/\sqrt{\omega_2} \\ \vdots \\ y_n/\sqrt{\omega_n} \end{bmatrix} \quad \text{sobre} \quad \boldsymbol{PX} = \begin{bmatrix} \boldsymbol{x}_1^\top/\sqrt{\omega_1} \\ \boldsymbol{x}_2^\top/\sqrt{\omega_2} \\ \vdots \\ \boldsymbol{x}_n^\top/\sqrt{\omega_n} \end{bmatrix}.$$

Aplicando-se o OLS ao modelo transformado, nós obtemos o estimador de mínimos quadrados ponderados (WLS),

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} w_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right]^{-1} \left[\sum_{i=1}^{n} w_i \boldsymbol{x}_i y_i\right],$$

sendo $w_i = 1/\omega_i$.

Uma especificação comum é que a variância seja proporcional a um dos regressores ou de seu quadrado, por exempo se $\sigma_i^2 = \sigma^2 x_{ik}^2$, então o modelo de regressão transformado para GLS é

$$\frac{y}{x_k} = \beta_k + \beta_1 \frac{x_1}{x_k} + \beta_2 \frac{x_2}{x_k} + \cdots + \frac{\varepsilon}{x_k}.$$

O estimador WLS é consistente seja qual for os pesos utilizados, mas os pesos não podem ser correlacionados com os distúrbios.

Entretanto, a escolha dos pesos deve ser feita com cuidado para não “afetar em demasia” os erros padrões do estimador GLS.

Modelos para dados de painel

Dados que combinam séries temporais e seções cruzadas são comuns em economia.

Temos um modelo de regressão da forma:

$$\begin{aligned} y_{it} &= \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{\top}\boldsymbol{\alpha} + \varepsilon_{it} \\ &= \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + c_i + \varepsilon_{it}. \end{aligned}$$

- $\boldsymbol{x}_{it}$ inclui p regressores, excluindo o termo constante.
- A heterogeneidade ou o efeito do indivíduo é $\boldsymbol{z}_{i}^{\top}\boldsymbol{\alpha}$ com $\boldsymbol{z}_i$ contendo o termo constante e um conjunto de variáveis individuais ou em grupo (raça, sexo, localização, etc.).
- Exogeneidade estrita: $\mathrm{E}(\varepsilon_{it}|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots) = 0$.
- Independência na média: $\mathrm{E}(c_i|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots) = \boldsymbol{\alpha}$.

Estruturas do modelo

1. **Regressão de coleção (*pooled*)**: se $\boldsymbol{z}_i$ contiver somente um termo constante, então OLS resulta em estimativas consistentes e eficientes de α e $\boldsymbol{\beta}$.
2. **Efeitos fixos**: se $\boldsymbol{z}_i$ não é observada, mas correlacionada com $\boldsymbol{x}_{it}$, então o estimador de mínimos quadrados de $\boldsymbol{\beta}$ é tendencioso e inconsistente como consequência da omissão de uma variável.

   O modelo $y_{it} = \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \alpha_i + \varepsilon_{it}$ tem α_i como um termo constante para cada grupo i.

   Efeito fixo está relacionado com a correlação de c_i e $\boldsymbol{x}_{it}$.
3. **Efeitos aleatórios**: se a heterogeneidade individual não observável, entretanto formulada, for por hipótese não correlacionada com as variáveis incluídas, então o modelo pode ser formulado como:

$$\begin{aligned} y_{it} &= \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \mathrm{E}(\boldsymbol{z}_i^\top \boldsymbol{\alpha}) + \{\boldsymbol{z}_i^\top \boldsymbol{\alpha} - \mathrm{E}(\boldsymbol{z}_i^\top \boldsymbol{\alpha})\} + \varepsilon_{it} \\ &= \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \alpha + u_i + \varepsilon_{it}. \end{aligned}$$

4. **Parâmetros aleatórios**: $y_{it} = \boldsymbol{x}_{it}^\top (\boldsymbol{\beta} + \boldsymbol{h}_i) + (\alpha + u_i) + \varepsilon_{it}$.

Modelo de regressão de coleção (*pooled*)

Temos o seguinte modelo e hipóteses:

$$y_{it} = \alpha + \boldsymbol{x}_{it}^{\top}\beta + \varepsilon_{it}, \;\; i = 1, \ldots, n, \;\; t = 1, \ldots, T_i$$
$$\mathrm{E}(\varepsilon_{it} | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{iT_i}) = 0$$
$$\mathrm{Var}(\varepsilon_{it} | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{iT_i}) = \sigma_{\varepsilon}^2$$
$$\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{js} | \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{iT_i}) = 0 \text{ se } i \neq j \text{ ou } t \neq s.$$

Nesta forma, se as outras hipóteses do modelo de regressão linear clássico são satisfeitas, então podemos aplicar os resultados anteriores de mínimos quadrados ordinários.

Estimação por mínimos quadrados

- As hipóteses para a aplicação de mínimos quadrados ordinários são raramente satisfeitas.
- No modelo de efeitos aleatórios, $y_{it} = c_i + \boldsymbol{x}_{it}^\top \beta + \varepsilon_{it}$, sendo $\mathrm{E}(c_i|\boldsymbol{X}_i) = \alpha$, temos

$$\begin{aligned} y_{it} &= \alpha + \boldsymbol{x}_{it}^\top \beta + \varepsilon_{it} + (c_i - \mathrm{E}(c_i|\boldsymbol{X}_i)) \\ &= \alpha + \boldsymbol{x}_{it}^\top \beta + \varepsilon_{it} + u_i \\ &= \alpha + \boldsymbol{x}_{it}^\top \beta + \omega_{it}. \end{aligned}$$

  Deste modo, $\mathrm{E}(\omega_{it}\omega_{is}) = \sigma_u^2$ quando $t \neq s$ (correlacionado).
- O estimador de mínimos quadrados ordinários no modelo de regressão generalizado pode ser consistente.
- Entretanto, o estimador convencional da variância assintótica pode subestimar a verdadeira.

Estimação robusta da matriz de covariância

- Suponha o modelo (agrupando observações)

$$\boldsymbol{y}_i = \boldsymbol{X}_i\beta + \boldsymbol{\omega}_i.$$

- Pode existir heterocedasticidade entre os indivíduos.
- Entretanto, o maior problema de dados de painel é a correlação entre as observações ou a autocorrelação.
- Suponha que $\text{Var}(\boldsymbol{\omega}_i|\boldsymbol{X}_i) = \boldsymbol{\Omega}_i$.
- O estimador de mínimos quadrados ordinários de β é

$$\begin{aligned}
\boldsymbol{b} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \left[\sum_{i=1}^n \boldsymbol{X}_i^\top\boldsymbol{X}_i\right]^{-1}\sum_{i=1}^n \boldsymbol{X}_i^\top\boldsymbol{y}_i \\
&= \left[\sum_{i=1}^n \boldsymbol{X}_i^\top\boldsymbol{X}_i\right]^{-1}\sum_{i=1}^n \boldsymbol{X}_i^\top(\boldsymbol{X}_i\beta + \boldsymbol{\omega}_i) \\
&= \beta + \left[\sum_{i=1}^n \boldsymbol{X}_i^\top\boldsymbol{X}_i\right]^{-1}\sum_{i=1}^n \boldsymbol{X}_i^\top\boldsymbol{\omega}_i.
\end{aligned}$$

A variância assintótica

$$\begin{aligned}\text{Var}(\boldsymbol{b}) &= \text{plim}\left[\sum_{i=1}^{n}\boldsymbol{X}_i^\top\boldsymbol{X}_i\right]^{-1}\text{plim}\left[\sum_{i=1}^{n}\boldsymbol{X}_i^\top\omega_i\omega_i^\top\boldsymbol{X}_i\right]\text{plim}\left[\sum_{i=1}^{n}\boldsymbol{X}_i^\top\boldsymbol{X}_i\right]^{-1}\\ &= \text{plim}\left[\sum_{i=1}^{n}\boldsymbol{X}_i^\top\boldsymbol{X}_i\right]^{-1}\text{plim}\left[\sum_{i=1}^{n}\boldsymbol{X}_i^\top\boldsymbol{\Omega}_i\boldsymbol{X}_i\right]\text{plim}\left[\sum_{i=1}^{n}\boldsymbol{X}_i^\top\boldsymbol{X}_i\right]^{-1}\end{aligned}$$

A matriz no centro deve ser estimada, então

$$\widehat{\text{Var}}(\boldsymbol{b}) = \left[\sum_{i=1}^{n}\boldsymbol{X}_i^\top\boldsymbol{X}_i\right]^{-1}\left[\sum_{i=1}^{n}\boldsymbol{X}_i^\top\widehat{\omega}_i\widehat{\omega}_i^\top\boldsymbol{X}_i\right]\left[\sum_{i=1}^{n}\boldsymbol{X}_i^\top\boldsymbol{X}_i\right]^{-1}.$$

Este estimador da variância assintótica leva mais em conta a correlação entre as observações cruzadas e não a heterocedasticidade.

Exemplo - salários

Trataremos da modelagem dos salários de chefes de família entre 1976 e 1982. Para isto, são utilizadas diversas covariáveis.

Ver arquivo exemplo_20.r

O modelo de efeitos fixos

O modelo de efeitos fixos é dado por

$$y_{it} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}$$

A formulação de efeitos fixos implica que a diferença entre grupos pode ser capturada em diferenças no termo constante.

Cada α_i é tratado como um parâmetro a ser estimador.

Problema: qualquer variável invariante no tempo terá o efeito do termo constante especifíco do indivíduo.

Os coeficientes nas variáveis invariantes no tempo não podem ser estimados.

Esta falta de identificação é o preço que se paga por robustez da especificação de correlação não medida entre os efeitos comuns e as variáveis exógenas.

Estimação por mínimos quadrados

Vamos supor que os dados sejam balanceados e que $\boldsymbol{y}_i$ e $\boldsymbol{X}_i$ os dados agrupados de cada grupo i.

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \mathbf{1}\alpha_i + \boldsymbol{\varepsilon}_i.$$

Temos,

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}$$

ou

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{d}_1 & \boldsymbol{d}_2 & \cdots & \boldsymbol{d}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} + \boldsymbol{\varepsilon},$$

sendo $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{d}_1 & \boldsymbol{d}_2 & \cdots & \boldsymbol{d}_n \end{bmatrix}$ e $\boldsymbol{d}_i$ uma variável indicadora da i-ésima unidade.

Este é um modelo de regressão clássico. Portanto, não é necessário nenhum resultado novo para analizá-lo.

O modelo de efeitos aleatórios

O modelo de efeitos fixos permite que o efeito individual não observado seja correlacionado com as variáveis incluídas.

Uma reformulação do modelo é

$$y_{it} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + (\alpha + u_i) + \varepsilon_{it}$$

com p regressores, incluindo uma constante. Temos,

$$u_i = \boldsymbol{z}_i^{\top}\boldsymbol{\alpha} - \mathrm{E}(\boldsymbol{z}_i^{\top}\boldsymbol{\alpha}) \quad \text{e} \quad \mathrm{E}(\boldsymbol{z}_i^{\top}\boldsymbol{\alpha}) = \boldsymbol{\alpha}$$

Vamos supor exogeneidade estrita:

$$\begin{aligned}
\mathrm{E}(\varepsilon_{it}|\boldsymbol{X}) &= \mathrm{E}(u_i|\boldsymbol{X}) = 0, \\
\mathrm{E}(\varepsilon_{it}^2|\boldsymbol{X}) &= \sigma_\varepsilon^2, \qquad \mathrm{E}(u_i^2|\boldsymbol{X}) = \sigma_u^2, \\
\mathrm{E}(\varepsilon_{it}u_j|\boldsymbol{X}) &= 0, \quad \text{para todo } i,\ t \text{ e } j, \\
\mathrm{E}(\varepsilon_{it}\varepsilon_{js}|\boldsymbol{X}) &= 0, \quad \text{se} \quad t \neq s \quad \text{e} \quad i \neq j, \\
\mathrm{E}(u_i u_j|\boldsymbol{X}) &= 0, \quad \text{se} \quad i \neq j.
\end{aligned}$$

Agruparemos em blocos de T observações, $\boldsymbol{y}_i$, $\boldsymbol{X}_i$, $u_i\mathbf{1}$ e ε_i.

$$\eta_{it} = \varepsilon_{it} + u_i, \quad \boldsymbol{\eta}_i = (\eta_{i1}, \eta_{i2}, \ldots, \eta_{iT})^\top$$
$$\mathrm{E}(\eta_{ij}^2|\boldsymbol{X}) = \sigma_\varepsilon^2 + \sigma_u^2, \quad \mathrm{E}(\eta_{it}\eta_{is}|\boldsymbol{X}) = \sigma_u^2 \quad t \neq s,$$
$$\mathrm{E}(\eta_{it}\eta_{js}|\boldsymbol{X}) = \sigma_u^2 \quad t \neq s \quad \text{para todo } t \text{ e } s \text{ se} \quad i \neq j..$$

Assim,

$$\mathrm{E}(\boldsymbol{\eta}_i\boldsymbol{\eta}_i^\top|\boldsymbol{X}) = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix} = \sigma_\varepsilon^2\boldsymbol{I}_T + \sigma_u^2\mathbf{1}\mathbf{1}^\top.$$

As observações i e j são independentes, logo a matriz de covariância para as nT observações é $\boldsymbol{\Omega} = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}$.

Mínimos quadrados generalizados

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{y} = \left[\sum_{i=1}^{n} \boldsymbol{X}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_i\right]^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_i.$$

Precisamos estimar a matriz $\boldsymbol{\Sigma}$. Uma abordagem para estimar as componentes de variância é a seguinte:

- Estimaremos $\sigma_\eta^2 = \sigma_\varepsilon^2 + \sigma_u^2$ conjuntamente através da variância do modelo de regressão de coleção (*pooled*).
- Estimaremos σ_ε^2 através da variância do modelo de regressão com efeitos fixos.
- Finalmente, estimaremos σ_u^2 pela relação $\sigma_u^2 = \sigma_\eta^2 - \sigma_\varepsilon^2$.

Após estimar a matriz de variância $\boldsymbol{\Sigma}$, estimamos $\boldsymbol{\beta}$ pela equação acima e

$$\widehat{\text{Var}}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \left[\sum_{i=1}^{n} \boldsymbol{X}_i^\top \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}_i\right]^{-1}.$$

Exemplo - salários (revisitado)

Trataremos da modelagem dos salários de chefes de família entre 1976 e 1982. Para isto, são utilizadas diversas covariáveis.

Ver arquivo exemplo_21.r

Regressão não linear e mínimos quadrados não lineares

A forma geral do modelo de regressão não linear é

$$y_i = h(\boldsymbol{x}_i, \boldsymbol{\beta}) + \varepsilon_i.$$

Existem modelos que podem ser linearizados como

$$y = e^{\beta_0} x_1^{\beta_1} x_2^{\beta_2} e^{\varepsilon}.$$

Entretanto, estamos interessados em modelos em que este tipo de transformação não é possível.

Exemplo

Modelo de função de produção da elasticidade constante de substituição:

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln \left(\delta K^{-\rho} + (1 - \delta) L^{-\rho} \right) + \varepsilon.$$

Hipóteses do modelo de regressão não linear

HP.1 Forma funcional: $\mathrm{E}(y_i|\boldsymbol{x}_i) = h(\boldsymbol{x}_i, \boldsymbol{\beta})$, $i = 1, 2, \ldots, n$, sendo $h(\boldsymbol{x}_i, \boldsymbol{\beta})$ uma função diferenciável continuamente de $\boldsymbol{\beta}$.

HP.2 Identificação: o vetor de parâmetros do modelo é identificado (estimável) se existir um parâmetro não nulo $\boldsymbol{\beta}^0 \neq \boldsymbol{\beta}$ tal que $h(\boldsymbol{x}_i, \boldsymbol{\beta}^0) = h(\boldsymbol{x}_i, \boldsymbol{\beta})$ para todo $\boldsymbol{x}_i$.

HP.3 Média zero dos distúrbios: pela Hipótese 1 temos $y_i = h(\boldsymbol{x}_i, \boldsymbol{\beta}) + \varepsilon_i$ sendo $\mathrm{E}(\varepsilon_i|h(\boldsymbol{x}_i, \boldsymbol{\beta})) = 0$.

HP.4 Homocedasticidade e autocorrelação nula:
$\mathrm{E}(\varepsilon_i^2|h(\boldsymbol{x}_j, \boldsymbol{\beta}),\ j = 1, \ldots, n) = \sigma^2 < \infty$ e
$\mathrm{E}(\varepsilon_i\varepsilon_j|h(\boldsymbol{x}_j, \boldsymbol{\beta}),\ j = 1, \ldots, n) = 0$ para todo $i \neq j$.

HP.5 Processo de geração de dados: $\boldsymbol{x}_i$ é estritamente exógena ao processo ε_i e tem os dois primeiros momentos bem definidos e convergentes.

HP.6 Modelo de probabilidade: existe uma distribuição de probabilidade bem definida para ε_i.

Condição de primeira ordem para um modelo não linear

Exemplo

Seja o modelo

$$y_i = \beta_1 + \beta_2 \exp\{\beta_3 x_i\} + \varepsilon_i.$$

Então,

$$S(\boldsymbol{b}) = \sum_{i=1}^{n} (y_i - b_1 - b_2 \exp\{b_3 x_i\})^2,$$

tal que

$$\begin{aligned}
\frac{\partial S(\boldsymbol{b})}{\partial b_1} &= -\sum_{i=1}^{n} (y_i - b_1 - b_2 \exp\{b_3 x_i\}) = 0 \\
\frac{\partial S(\boldsymbol{b})}{\partial b_2} &= -\sum_{i=1}^{n} (y_i - b_1 - b_2 \exp\{b_3 x_i\}) \exp\{b_3 x_i\} = 0 \\
\frac{\partial S(\boldsymbol{b})}{\partial b_3} &= -\sum_{i=1}^{n} (y_i - b_1 - b_2 \exp\{b_3 x_i\}) b_2 x_i \exp\{b_3 x_i\} = 0,
\end{aligned}$$

não tem solução analítica.

Definição

Um modelo de regressão não linear é um na qual as condições de primeira ordem de estimação por mínimos quadrados dos parâmetros são funções não lineares dos parâmetros.

Regressão linearizada

Seja o modelo de regressão linear é

$$y = h(\boldsymbol{x}, \boldsymbol{\beta}) + \varepsilon.$$

A expansão de Taylor em torno de $\boldsymbol{\beta}^0$:

$$\begin{aligned} h(\boldsymbol{x}, \boldsymbol{\beta}) &\simeq h(\boldsymbol{x}, \boldsymbol{\beta}^0) + \sum_{k=1}^{K} \frac{\partial h(\boldsymbol{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} (\beta_k - \beta_k^0) \\ &= \left[h(\boldsymbol{x}, \boldsymbol{\beta}^0) - \sum_{k=1}^{K} \beta_k^0 \left(\frac{\partial h(\boldsymbol{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} \right) \right] + \sum_{k=1}^{K} \beta_k \left(\frac{\partial h(\boldsymbol{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} \right). \end{aligned}$$

Seja $x_k^0 = \dfrac{\partial h(\boldsymbol{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0}$. Dado $\boldsymbol{\beta}^0$, temos que x_k^0 é função dos dados.

$$
\begin{aligned}
h(\boldsymbol{x}, \boldsymbol{\beta}) &\simeq \left[h^0 - \sum_{k=1}^{K} x_k^0 \beta_k^0\right] + \sum_{k=1}^{K} x_k^0 \beta_k = h^0 - \boldsymbol{x}^{0\top}\boldsymbol{\beta}^0 + \boldsymbol{x}^{0\top}\boldsymbol{\beta} \Rightarrow \\
y &\simeq h^0 - \boldsymbol{x}^{0\top}\boldsymbol{\beta}^0 + \boldsymbol{x}^{0\top}\boldsymbol{\beta} + \varepsilon \Rightarrow \\
y^0 &= y - h^0 + \boldsymbol{x}^{0\top}\boldsymbol{\beta}^0 = \boldsymbol{x}^{0\top}\boldsymbol{\beta} + \varepsilon^0.
\end{aligned}
$$

ε^0 contem o distúrbio verdadeiro mais o erro de aproximação de Taylor de primeira ordem.

Conhecendo $\boldsymbol{\beta}^0$, podemos calcular y^0 e x^0 e estimar os parâmetros $\boldsymbol{\beta}$ por mínimos quadrados.

Exemplo - regressão linearizada

$$
y_i = \beta_1 + \beta_2 \exp\{\beta_3 x_i\} + \varepsilon_i.
$$

$$
x_1^0 = \frac{\partial h(.)}{\partial \beta_1} = 1, \quad x_2^0 = \frac{\partial h(.)}{\partial \beta_2} = \exp\{\beta_3^0 x\}, \quad \text{e} \quad x_3^0 = \frac{\partial h(.)}{\partial \beta_3} = \beta_2^0 x \exp\{\beta_3^0 x\}.
$$

$$
y^0 = y - h(x, \beta_1^0, \beta_2^0, \beta_3^0) + \beta_1^0 x_1^0 + \beta_2^0 x_2^0 + \beta_3^0 x_3^0.
$$

Sob certas condições, o estimador de mínimos quadrados não linear é consistente.

Além disto, $\boldsymbol{b} \approx \mathcal{N}\left(\beta, \frac{\sigma^2}{n}[\boldsymbol{Q}^0]^{-1}\right)$ sendo $\boldsymbol{Q}^0 = \text{plim}\frac{1}{n}\boldsymbol{X}^{0\top}\boldsymbol{X}^0$ e $\widehat{\text{Var}}(\boldsymbol{b}) = \widehat{\sigma}^2(\boldsymbol{X}^{0\top}\boldsymbol{X}^0)^{-1}$.

Solução iterativa

$$\begin{aligned} b_{t+1} &= \left[\sum_{i=1}^n \boldsymbol{x}_i^0 \boldsymbol{x}_i^{0\top}\right]^{-1}\left[\sum_{i=1}^n \boldsymbol{x}_i^0 (y_i - h_i^0 + \boldsymbol{x}_i^{0\top} b_t)\right] \\ &= b_t + \left[\sum_{i=1}^n \boldsymbol{x}_i^0 \boldsymbol{x}_i^{0\top}\right]^{-1}\left[\sum_{i=1}^n \boldsymbol{x}_i^0 (y_i - h_i^0)\right] \\ &= b_t + (\boldsymbol{X}^{0\top}\boldsymbol{X}^0)^{-1}\boldsymbol{X}^{0\top}\boldsymbol{e}^0 \\ &= b_t + \boldsymbol{\Delta}_t. \end{aligned}$$

Um estimador consistente de σ^2 é

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (y_i - h(\boldsymbol{x}_i, \boldsymbol{b}))^2.$$

Estimação por variáveis instrumentais

- Temos trabalhado com a hipótese de que $\boldsymbol{x}_i$ e ε_i são não correlacionados no modelo de regressão linear.
- Existem aplicações em economia onde isto não é possível.
- Sem esta hipótese, nenhuma das provas de consistência ou não tendenciosidade do estimador de mínimos quadrados permanecem válidos.

Exemplo - modelos de dados de painel dinâmico.

$$y_{it} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \gamma y_{i,t-1} + \varepsilon_{it} + u_i.$$

Temos $y_{i,t-1}$ é correlacionado com o distúrbio $\varepsilon_{it} + u_i$.

- Podemos utilizar o método de variáveis instrumentais.

Hipóteses do modelo

Das hipóteses anteriores, manteremos o resultado

$$\text{plim}(\boldsymbol{X}^\top \boldsymbol{X}/n) = \boldsymbol{Q}_{xx} \quad \text{e} \quad \textbf{HP.3I} : \text{E}(\varepsilon_i|\boldsymbol{x}_i) = \eta_i \quad \Rightarrow \quad \text{E}(\boldsymbol{x}_i\varepsilon_i) = \boldsymbol{\gamma}.$$

Sob certas condições de regularidade, $\text{plim}(1/n)\boldsymbol{X}^\top \boldsymbol{\varepsilon} = \boldsymbol{\gamma}$.

Assim, os regressores $\boldsymbol{X}$ não são mais exógenos.

Agora, suporemos que existe um conjunto de variáveis $\boldsymbol{Z}$ tal que

1 - Exogeneidade: elas são não correlacionadas com os distúrbios; e

2 - Relevância: elas são correlacionadas com as variáveis $\boldsymbol{X}$.

Estas (variáveis *Z*) são chamadas de variáveis instrumentais.

Hipóteses:

HP.7I: $[\boldsymbol{x}_i, \boldsymbol{z}_i, \varepsilon_i]$, $i = 1, \ldots, n$, formam uma sequência iid de variáveis aleatórias.

HP.8Ia: $\mathrm{E}(x_{ik}^2) = \boldsymbol{Q}_{xx,kk} < \infty$, uma constante finita, $k = 1, \ldots, K$.

HP.8Ib: $\mathrm{E}(z_{i\ell}^2) = \boldsymbol{Q}_{zz,\ell\ell} < \infty$, uma constante finita, $\ell = 1, \ldots, L$.

HP.8Ic: $\mathrm{E}(z_{i\ell} x_{ik}) = \boldsymbol{Q}_{zx,\ell k} < \infty$, uma constante finita, $\ell = 1, \ldots, L$ e $k = 1, \ldots, K$.

HP.9I: $\mathrm{E}(\varepsilon_i | \boldsymbol{z}_i) = 0$.

Além disso,

- $\operatorname{plim}(1/n)\boldsymbol{Z}^\top \boldsymbol{Z} = \boldsymbol{Q}_{zz}$,
- $\operatorname{plim}(1/n)\boldsymbol{Z}^\top \boldsymbol{X} = \boldsymbol{Q}_{zx}$ (relevância), e
- $\operatorname{plim}(1/n)\boldsymbol{Z}^\top \varepsilon = \boldsymbol{0}$ (exogeneidade).

Estimação

Mínimos quadrados ordinários

$$\mathrm{E}(\boldsymbol{b}|\boldsymbol{X}) = \boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\eta} \neq \boldsymbol{\beta} \quad \text{(tendencioso)}$$

$$\operatorname{plim}\boldsymbol{b} = \boldsymbol{\beta} + \operatorname{plim}\left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n}\right)^{-1} \operatorname{plim}\left(\frac{\boldsymbol{X}^\top \varepsilon}{n}\right) = \boldsymbol{\beta} + \boldsymbol{Q}_{xx}^{-1}\gamma \neq \boldsymbol{\beta} \text{ (inconsistente).}$$

Estimador por variáveis instrumentais

Temos, $\mathrm{E}(\boldsymbol{z}_i \varepsilon_i) = 0$, então

$$\operatorname{plim}\left(\frac{\boldsymbol{Z}^\top \varepsilon}{n}\right) = \operatorname{plim}\left(\frac{\boldsymbol{Z}^\top \boldsymbol{y}}{n}\right) - \operatorname{plim}\left(\frac{\boldsymbol{Z}^\top \boldsymbol{X}\boldsymbol{\beta}}{n}\right) = \mathbf{0}.$$

Portanto,

$$\operatorname{plim}\left(\frac{\boldsymbol{Z}^\top \boldsymbol{y}}{n}\right) = \left[\operatorname{plim}\left(\frac{\boldsymbol{Z}^\top \boldsymbol{X}}{n}\right)\right]\boldsymbol{\beta} + \operatorname{plim}\left(\frac{\boldsymbol{Z}^\top \varepsilon}{n}\right) = \left[\operatorname{plim}\left(\frac{\boldsymbol{Z}^\top \boldsymbol{X}}{n}\right)\right]\boldsymbol{\beta}.$$

Suponha que Z tenha o mesmo número de variáveis do que X.

O posto de $\boldsymbol{Z}^\top \boldsymbol{X}$ é p, e $\boldsymbol{Z}^\top \boldsymbol{X}$ é uma matriz quadrada. Logo,

$$\left[\text{plim}\left(\frac{\boldsymbol{Z}^\top \boldsymbol{X}}{n}\right)\right]^{-1} \text{plim}\left(\frac{\boldsymbol{Z}^\top \boldsymbol{y}}{n}\right) = \boldsymbol{\beta}.$$

que nos leva ao **estimador por variáveis instrumentais**

$$\boldsymbol{b}_{IV} = (\boldsymbol{Z}^\top \boldsymbol{X})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}.$$

Já provamos que $\boldsymbol{b}_{IV}$ é consistente.

A distribuição assintótica de $\boldsymbol{b}_{IV}$ é dada por

$$\boldsymbol{b}_{IV} \approx \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \boldsymbol{Q}_{zx}^{-1} \boldsymbol{Q}_{zz} \boldsymbol{Q}_{xz}^{-1})$$

sendo $\boldsymbol{Q}_{zx} = \text{plim}\left(\frac{\boldsymbol{Z}^\top \boldsymbol{X}}{n}\right)$ e $\boldsymbol{Q}_{zz} = \text{plim}\left(\frac{\boldsymbol{Z}^\top \boldsymbol{Z}}{n}\right)$.

Estimamos σ^2 por

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{b}_{IV})^2$$

e temos

$$\begin{aligned}
\widehat{\text{Var}}(\boldsymbol{b}_{IV}) &= \frac{1}{n}\left[\left(\frac{\widehat{\varepsilon}^\top\widehat{\varepsilon}}{n}\right)\left(\frac{\boldsymbol{Z}^\top\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{Z}^\top\boldsymbol{Z}}{n}\right)\left(\frac{\boldsymbol{X}^\top\boldsymbol{Z}}{n}\right)^{-1}\right] \\
&= \widehat{\sigma}^2(\boldsymbol{Z}^\top\boldsymbol{X})^{-1}(\boldsymbol{Z}^\top\boldsymbol{Z})(\boldsymbol{X}^\top\boldsymbol{Z})^{-1},
\end{aligned}$$

sendo $\widehat{\varepsilon} = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{Z}^\top\boldsymbol{X})^{-1}\boldsymbol{Z}^\top]\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_{IV}$.

Se Z contiver mais variáveis que X, então podemos ter a projeção das colunas de $\boldsymbol{X}$ no espaço das colunas de $\boldsymbol{Z}$:

$$\widehat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{X}.$$

Neste caso, temos as variáveis instrumentais $\widehat{\boldsymbol{X}}$ e

$$\boldsymbol{b}_{IV} = (\widehat{\boldsymbol{X}}^\top \boldsymbol{X})^{-1}\widehat{\boldsymbol{X}}^\top \boldsymbol{y} = [\boldsymbol{X}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{X}]^{-1}\boldsymbol{X}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{y}.$$

Temos também

$$\widehat{\text{Var}}(\boldsymbol{b}_{IV}) = \widehat{\sigma}^2[\boldsymbol{X}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{X}]^{-1}.$$

Podemos reescrever

$$\boldsymbol{b}_{IV} = (\widehat{\boldsymbol{X}}^\top \boldsymbol{X})^{-1}\widehat{\boldsymbol{X}}^\top \boldsymbol{y} = (\boldsymbol{X}^\top (\boldsymbol{I} - \boldsymbol{M}_z)\boldsymbol{X})^{-1}\boldsymbol{X}^\top (\boldsymbol{I} - \boldsymbol{M}_z)\boldsymbol{y} = (\widehat{\boldsymbol{X}}^\top \widehat{\boldsymbol{X}})^{-1}\widehat{\boldsymbol{X}}^\top \boldsymbol{y}.$$

Isto é chamado de estimador de mínimos quadrados em dois passos (primeiro $\widehat{\boldsymbol{X}}$, depois $\boldsymbol{b}_{IV}$).

O estimador de σ^2 não deve ser baseado em $\widehat{\boldsymbol{X}}$.

O estimador

$$s^2_{IV} = \frac{(\boldsymbol{y} - \widehat{\boldsymbol{X}}\boldsymbol{b}_{IV})^\top (\boldsymbol{y} - \widehat{\boldsymbol{X}}\boldsymbol{b}_{IV})}{n}$$

é inconsistente para σ^2, com ou sem a correção para os graus de liberdade.

Exemplo

No modelo de equilíbrio de mercado, podemos ter a equação

$$\text{Wks}_{it} = \gamma_1 + \gamma_2 \ln \mathit{Wage}_{it} + \gamma_3 \mathit{Ed}_i + \gamma_4 \mathit{Union}_{it} + \gamma_5 \mathit{Fem}_i + u_{it}.$$

Se o número de semanas trabalhadas e o salário aceito são determinados conjuntamente, então $\ln \mathit{Wage}$ e u_{it} nesta equação são correlacionados.

Consideramos dois conjuntos de variáveis instrumentais:

$$\begin{aligned} \boldsymbol{Z}_1 &= [1, \mathit{Ind}_{it}, \mathit{Ed}_i, \mathit{Union}_{it}, \mathit{Fem}_i] \\ \text{e} & \\ \boldsymbol{Z}_2 &= [1, \mathit{Ind}_{it}, \mathit{Ed}_i, \mathit{Union}_{it}, \mathit{Fem}_i, \mathit{SMSA}_{it}]. \end{aligned}$$

Tabela 18: Equação do trabalho estimada.

Variável	OLS		IV com $\boldsymbol{Z}_1$		IV com $\boldsymbol{Z}_2$	
	Estimado	E.P.	Estimado	E.P.	Estimado	E.P.
Constant	44,7665	1,2153	18,8986	13,0590	30,7044	4,9997
ln *Wage*	0,7326	0,1972	5,1828	2,2454	3,1518	0,8572
Education	-0,1532	0,0321	-0,4600	0,1578	-0,3200	0,0661
Union	-1,9960	0,1701	-2,3602	0,2567	-2,1940	0,1860
Female	-1,3498	0,2642	0,6957	1,0650	-0,2378	0,4679

- Notemos o valor baixo para o coeficiente do ln *Wage* no OLS.
- As duas estimativas por variáveis instrumentais são maiores.
- Para as variáveis $\boldsymbol{Z}_2$, utilizamos a estimação em dois passos discutidos anteriormente.

Estimação por máxima verossilhança

A função de verossimilhança é dada por

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = f(y_1, \ldots, y_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

para y_i, $i = 1, \ldots, n$, iid condicional em $\boldsymbol{\theta}$.

O logaritmo da função de verossimilhança é dada por

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \ln L(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\theta}).$$

Suponha que no modelo clássico de regressão os distúrbios sejam normais. Então,

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = -\frac{1}{2} \sum_{i=1}^{n} \left[\ln \sigma^2 + \ln(2\pi) + \frac{1}{\sigma^2}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 \right]$$

sendo $\boldsymbol{X}$ a matriz $n \times p$ dos regressores com a i-ésima linha igual a $\boldsymbol{x}_i$.

Estamos interessados em obter estimativas dos parâmetros, $\boldsymbol{\theta}$.

Definição

O vetor de parâmetros $\boldsymbol{\theta}$ é **identificável** (estimável) se para qualquer outro vetor de parâmetros, $\gamma \neq \boldsymbol{\theta}$, para algum dado $\boldsymbol{y}$, $L(\gamma|\boldsymbol{y}) \neq L(\boldsymbol{\theta}|\boldsymbol{y})$.

O príncipio da máxima verossimilhança

Suponha uma amostra de 10 observações de uma distribuição Poisson: 5, 0, 1, 1, 0, 3, 2, 3, 4 e 1.

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{e^{-100\theta}\theta^{20}}{207.360}.$$

Qual valor mais provável de θ que produziu esta amostra?

A função tem um único máximo em $\theta = 2$, que é a estimativa de máxima verossimilhança.

$$
\begin{aligned}
\ell(\theta|\boldsymbol{y}) &= -n\theta + n\overline{y}\ln\theta - \sum_{i=1}^{n}\ln(y_i!) \\
\frac{\partial\ell(\theta|\boldsymbol{y})}{\partial\theta} &= -n + \frac{n\overline{y}}{\theta} = 0, \quad \Rightarrow \quad \widehat{\theta}_{MV} = \overline{y}_n \\
\frac{\partial^2\ell(\theta|\boldsymbol{y})}{\partial\theta^2} &= -\frac{n\overline{y}}{\theta^2} < 0, \quad \Rightarrow \quad \overline{y}_n \text{ é ponto de máximo.}
\end{aligned}
$$

Chamamos de **equação de verossimilhança** a expressão:

$$
\frac{\partial\ell(\boldsymbol{\theta}|\textbf{dados})}{\partial\boldsymbol{\theta}} = \frac{\partial\ln L(\boldsymbol{\theta}|\textbf{dados})}{\partial\boldsymbol{\theta}} = \mathbf{0}.
$$

Denotaremos:

- $\widehat{\boldsymbol{\theta}}$ como o estimador de máxima verossimilhança;
- $\boldsymbol{\theta}_0$ o valor verdadeiro do vetor de parâmetros; e
- $\boldsymbol{\theta}$ um outro valor possível do vetor parâmetros.

Teorema

Sob regularidade, o EMV tem as seguintes propriedades assintóticas:

M1: Consistência: $\operatorname{plim} \widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

M2: Normalidade: $\widehat{\boldsymbol{\theta}} \approx \mathcal{N}(\boldsymbol{\theta}_0, [\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1})$ sendo

$$\boldsymbol{I}(\boldsymbol{\theta}_0) = \mathrm{E}_0\left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right).$$

M3: Eficiência: $\widehat{\boldsymbol{\theta}}$ é assintoticamente eficiente e atinge o limite inferior de Cramér-Rao.

M4: Invariância: o EMV de $\boldsymbol{\gamma} = \boldsymbol{c}(\boldsymbol{\theta}_0)$ é $\boldsymbol{c}(\widehat{\boldsymbol{\theta}})$ se $\boldsymbol{c}(\boldsymbol{\theta}_0)$ é uma função contínua e continuamente diferenciável (transformações biunívocas).

Se $\boldsymbol{g}_i = \dfrac{\partial \ln f(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ (a função escore) e $\boldsymbol{H}_i = \dfrac{\partial \ln f(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$ (a matriz de informação), então

D1: $\mathrm{E}_0(\boldsymbol{g}_i(\boldsymbol{\theta})) = \mathbf{0}$; e

D2: $\mathrm{Var}_0(\boldsymbol{g}_i(\boldsymbol{\theta})) = \mathrm{E}_0(-\boldsymbol{H}_i(\boldsymbol{\theta}))$.

Teorema

$\mathrm{E}_0\left(\frac{1}{n}\ell(\boldsymbol{\theta}_0)\right) > \mathrm{E}_0\left(\frac{1}{n}\ell(\boldsymbol{\theta})\right)$ para qualquer $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ (incluindo $\widehat{\boldsymbol{\theta}}$).

O valor esperado da log-verossimilhança é maximizado no valor verdadeiro dos parâmetros.

Teorema (Cramér-Rao)

Sob condições de regularidade, a variância assintótica de um estimador, consistente e assintoticamente normal, do vetor de parâmetros $\boldsymbol{\theta}_0$ será sempre no mínimo maior ou igual a

$$[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1} = \left[\mathrm{E}_0\left(-\frac{\partial^2\ell(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0^\top}\right)\right]^{-1} = \left[\mathrm{E}_0\left(\left(\frac{\partial\ell(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0}\right)\left(\frac{\partial\ell(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0}\right)^\top\right)\right]^{-1}.$$

Teste da razão de verossimilhança

Seja $\widehat{\theta}_U$ o EMV sob o modelo irrestrito e $\widehat{\theta}_R$ sob o modelo restrito.

Sejam também $\widehat{L}_U$ e $\widehat{L}_R$ as respectivas funções de verossimilhança avaliadas nestes pontos.

Então, a razão de verossimilhança é

$$\widehat{\lambda} = \frac{\widehat{L}_R}{\widehat{L}_U}, \quad \text{com} \quad 0 \leqslant \lambda \leqslant 1.$$

Teorema
Sob regularidade e sob H_0, a distribuição para amostras grandes de $-2 \ln \widehat{\lambda}$ é qui-quadrada, com os graus de liberdade igual ao número de restrições impostas.

A hipótese nula é rejeitada se este valor ultrapassa o valor crítico apropriado da distribuição qui-quadrada.

Teste de Wald

Se $\boldsymbol{x} \sim \mathcal{N}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então $(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \sim \chi^2_J$.

Suponha que $\widehat{\boldsymbol{\theta}}$ seja a estimativa do vetor de parâmetros sem restrições e que $H_0 : \boldsymbol{c}(\boldsymbol{\theta}) = \boldsymbol{q}$.

Se as restrições são válidas, então $\widehat{\boldsymbol{\theta}}$ deve satisfazê-las. Caso contrário, $\boldsymbol{c}(\boldsymbol{\theta}) - \boldsymbol{q}$ deve ser longe de $\mathbf{0}$.

Teorema

A estatística de Wald é

$$W = (\boldsymbol{c}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{q})^\top \left[\text{Var}(\boldsymbol{c}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{q})\right]^{-1} (\boldsymbol{c}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{q}).$$

Sob H_0, em amostras grandes, W tem uma distribuição qui-quadrada com os graus de liberdade igual ao número de restrições, isto é, o número de equações em $\boldsymbol{c}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{q} = \mathbf{0}$.

Ademais, $\widehat{\text{Var}}(\boldsymbol{c}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{q}) = \widehat{\boldsymbol{C}}\widehat{\text{Var}}(\widehat{\boldsymbol{\theta}})\widehat{\boldsymbol{C}}^\top$ sendo $\widehat{\boldsymbol{C}} = \left[\dfrac{\partial \boldsymbol{c}(\widehat{\boldsymbol{\theta}})}{\partial \widehat{\boldsymbol{\theta}}}\right]$.

Teste do Multiplicador de Lagrange

Suponha que maximizamos a log-verossimilhança sob a restrição $\boldsymbol{c}(\theta) - \boldsymbol{q} = \mathbf{0}$. Seja $\boldsymbol{\lambda}$ o vetor de multiplicadores de Lagrange e definamos a função lagrangeana

$$\ln L^{\star}(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) + \boldsymbol{\lambda}^{\top}(\boldsymbol{c}(\boldsymbol{\theta}) - \boldsymbol{q}).$$

A solução é dada por

$$\begin{aligned} \frac{\partial \ln L^{\star}}{\partial \boldsymbol{\theta}} &= \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \boldsymbol{C}^{\top}\boldsymbol{\lambda} \\ \frac{\partial \ln L^{\star}}{\partial \boldsymbol{\lambda}} &= \boldsymbol{c}(\boldsymbol{\theta}) - \boldsymbol{q} = \mathbf{0}. \end{aligned}$$

Podemos testar $H_0 : \ \boldsymbol{\lambda} = \mathbf{0}$.

Teorema

A estatística de teste do multiplicador de Lagrange é

$$LM = \left(\frac{\partial \ell(\widehat{\boldsymbol{\theta}}_R)}{\partial \widehat{\boldsymbol{\theta}}_R}\right)^{\top} \left[\boldsymbol{I}(\widehat{\boldsymbol{\theta}}_R)\right]^{-1} \left(\frac{\partial \ell(\widehat{\boldsymbol{\theta}}_R)}{\partial \widehat{\boldsymbol{\theta}}_R}\right).$$

Sob H_0, LM tem distribuicção limite como qui-quadrada com graus de liberdade igual ao número de restrições. Todos os termos são calculados no estimador restrito.

Comparando modelos

Duas medidas comuns para modelos não encaixados baseados na mesma lógica são

- Critério de Informação de Akaike: $AIC = -2 \ln L(\widehat{\boldsymbol{\theta}}) + 2p$; e
- Critério de Informação Bayesiana/Schwarz: $BIC = -2 \ln L(\widehat{\boldsymbol{\theta}}) + p \ln n$,

sendo p o número de parâmetros do modelo. Menor AIC (ou BIC), melhor o modelo.

Modelo de regressão linear normal

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i.$$

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\}. \end{aligned}$$

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}.$$

$$\begin{bmatrix} \dfrac{\partial\ell(\boldsymbol{\beta},\sigma^2)}{\partial\boldsymbol{\beta}} \\ \dfrac{\partial\ell(\boldsymbol{\beta},\sigma^2)}{\partial\sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{\sigma^2} \\ -\dfrac{n}{2\sigma^2} + \dfrac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ 0 \end{bmatrix} \quad \Rightarrow$$

$$\widehat{\boldsymbol{\beta}}_{MV} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \quad \text{e} \quad \widehat{\sigma}^2_{MV} = \frac{\boldsymbol{e}^\top\boldsymbol{e}}{n}.$$

Temos,

$$\begin{bmatrix} \dfrac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} & \dfrac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta} \partial \sigma^2} \\ \dfrac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2 \partial \boldsymbol{\beta}} & \dfrac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\dfrac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} & -\dfrac{\boldsymbol{X}^\top \boldsymbol{\varepsilon}}{\sigma^4} \\ -\dfrac{\boldsymbol{\varepsilon}^\top \boldsymbol{X}}{\sigma^4} & \dfrac{n}{2\sigma^4} - \dfrac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{\sigma^6} \end{bmatrix} \quad \Rightarrow$$

$$\left[\boldsymbol{I}(\boldsymbol{\beta}, \sigma^2) \right]^{-1} = \begin{bmatrix} \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \dfrac{2\sigma^4}{n} \end{bmatrix}.$$

Temos, $\sqrt{n}(\widehat{\sigma}^2_{MV} - \sigma^2) \xrightarrow{d} \mathcal{N}(0, 2\sigma^4)$.

Se $H_0 : \ \boldsymbol{R}\beta - \boldsymbol{q} = \boldsymbol{0}$, o teste da razão F,

$$F = \frac{(\boldsymbol{Rb} - \boldsymbol{q})^\top [\boldsymbol{R}s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top](\boldsymbol{Rb} - \boldsymbol{q})}{J} \sim \mathcal{F}_{J,n-p},$$

para qualquer tamanho de amostra se os distúrbios forem normalmente distribuídos.

Os outros testes e suas estatísticas de teste continuam não tendo distribuição exata conhecida para amostras finitas ou pequenas.

Modelo de regressão generalizado

$$
\begin{aligned}
&y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \ldots, n \\
&\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{0} \\
&\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top|\boldsymbol{X}) = \sigma^2\boldsymbol{\Omega}.
\end{aligned}
$$

Por hipótese, teremos $\boldsymbol{\Omega}$ como uma matriz constante conhecida.

$$
\begin{aligned}
\ell(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\ln\left|\boldsymbol{\Omega}\right| - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
\frac{\partial\ell}{\partial\boldsymbol{\beta}} &= \frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2}\boldsymbol{X}_\star^\top(\boldsymbol{y}_\star - \boldsymbol{X}_\star\boldsymbol{\beta}) = \boldsymbol{0} \qquad \boldsymbol{\Omega}^{-1} = \boldsymbol{P}^\top\boldsymbol{P} \\
\frac{\partial\ell}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \qquad \boldsymbol{X}_\star = \boldsymbol{P}\boldsymbol{X} \\
&= -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4}(\boldsymbol{y}_\star - \boldsymbol{X}_\star\boldsymbol{\beta})^\top(\boldsymbol{y}_\star - \boldsymbol{X}_\star\boldsymbol{\beta}) = 0. \qquad \boldsymbol{y}_\star = \boldsymbol{P}\boldsymbol{y} \\
\widehat{\boldsymbol{\beta}}_{MV} &= (\boldsymbol{X}_\star^\top\boldsymbol{X}_\star)^{-1}\boldsymbol{X}_\star^\top\boldsymbol{y} = (\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{y} \quad \text{e} \\
\widehat{\sigma}^2_{MV} &= \frac{1}{n}(\boldsymbol{y}_\star - \boldsymbol{X}_\star\widehat{\boldsymbol{\beta}}_{MV})^\top(\boldsymbol{y}_\star - \boldsymbol{X}_\star\widehat{\boldsymbol{\beta}}_{MV}) \\
&= \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{MV})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{MV}).
\end{aligned}
$$

Temos que $\widehat{\sigma}^2_{MV}$ é tendencioso para σ^2. Para obter um estimador não tendencioso precisamos multiplicá-lo pelo fator $n/(n-p)$.

Se $\boldsymbol{\Omega}$ for desconhecido, então precisamos estimar $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Omega})$ simultaneamente.

Mas $\boldsymbol{\Omega}$ tem $n(n+1)/2 - 1$ parâmetros. Precisamos impor restrições.

Em geral, a estimação conjunta de todos os parâmetros será complicada.

Modelo de heterocedasticidade multiplicativa

Considere um modelo de regressão com variância dada por

$$\sigma_i^2 = \sigma^2 \exp\{\boldsymbol{w}_i^\top \boldsymbol{\alpha}\} = \exp\{\boldsymbol{z}_i^\top \boldsymbol{\gamma}\},$$

sendo $\boldsymbol{z}_i^\top = (1, \boldsymbol{w}_i^\top)$ e $\boldsymbol{\gamma}^\top = (\ln \sigma^2, \boldsymbol{\alpha}^\top)$.

Neste caso, tomamos $\boldsymbol{\Omega} = \text{diag}(\exp\{\boldsymbol{z}_1^\top \boldsymbol{\gamma}\}, \exp\{\boldsymbol{z}_2^\top \boldsymbol{\gamma}\}, \ldots, \exp\{\boldsymbol{z}_n^\top \boldsymbol{\gamma}\})$.

$$
\begin{aligned}
\ell(\boldsymbol{\beta},\boldsymbol{\gamma}) &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\ln\sigma_i^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{\varepsilon_i^2}{\sigma_i^2} \\
&= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\ln \boldsymbol{z}_i^\top\boldsymbol{\gamma} - \frac{1}{2}\sum_{i=1}^{n}\frac{\varepsilon_i^2}{\exp\{\boldsymbol{z}_i^\top\boldsymbol{\gamma}\}} \\
\frac{\partial\ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{n}\boldsymbol{x}_i\frac{\varepsilon_i}{\exp\{\boldsymbol{z}_i^\top\boldsymbol{\gamma}\}} = \boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon} = \mathbf{0} \\
\frac{\partial\ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}} &= \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{z}_i\left(\frac{\varepsilon_i^2}{\exp\{\boldsymbol{z}_i^\top\boldsymbol{\gamma}\}} - 1\right) = \mathbf{0}. \\
\frac{\partial^2\ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} &= -\sum_{i=1}^{n}\frac{1}{\exp\{\boldsymbol{z}_i^\top\boldsymbol{\gamma}\}}\boldsymbol{x}_i\boldsymbol{x}_i^\top = -\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X} \\
\frac{\partial^2\ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^\top} &= -\sum_{i=1}^{n}\frac{\varepsilon_i}{\exp\{\boldsymbol{z}_i^\top\boldsymbol{\gamma}\}}\boldsymbol{x}_i\boldsymbol{z}_i^\top \\
\frac{\partial^2\ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top} &= -\frac{1}{2}\sum_{i=1}^{n}\frac{\varepsilon_i^2}{\exp\{\boldsymbol{z}_i^\top\boldsymbol{\gamma}\}}\boldsymbol{z}_i\boldsymbol{z}_i^\top.
\end{aligned}
$$

Temos,

$$\begin{aligned}
\mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^\top}\right) &= \mathbf{0} \quad \text{porque} \quad \mathrm{E}(\varepsilon_i|\boldsymbol{x}_i,\boldsymbol{z}_i)=0, \quad \text{e} \\
\mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top}\right) &= \frac{1}{2}\boldsymbol{Z}^\top\boldsymbol{Z} \quad \text{porque} \quad \mathrm{E}\left(\frac{\varepsilon_i^2}{\sigma_i^2}\middle|\boldsymbol{x}_i,\boldsymbol{z}_i\right)=1.
\end{aligned}$$

Seja $\boldsymbol{\delta}=(\boldsymbol{\beta},\boldsymbol{\gamma})$. Então,

$$\mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top}\right)=\begin{bmatrix}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X} & \mathbf{0}\\ \mathbf{0} & \frac{1}{2}\boldsymbol{Z}^\top\boldsymbol{Z}\end{bmatrix}=-\overline{\boldsymbol{H}}.$$

O método do escore é um algoritmo para encontrar uma solução:

$$\boldsymbol{\delta}_{t+1}=\boldsymbol{\delta}_t-\overline{\boldsymbol{H}}^{-1}\boldsymbol{g}_t$$

sendo $\boldsymbol{g}=\left(-\frac{\partial \ell(\boldsymbol{\delta})}{\partial\boldsymbol{\beta}^\top},-\frac{\partial \ell(\boldsymbol{\delta})}{\partial\boldsymbol{\gamma}^\top}\right)^\top$.

Como $\overline{\boldsymbol{H}}$ é bloco diagonal, temos

$$\begin{aligned}
\boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + (\boldsymbol{X}^\top \boldsymbol{\Omega}_t^{-1} \boldsymbol{X}) \boldsymbol{X}^\top \boldsymbol{\Omega}_t^{-1} \boldsymbol{\varepsilon}_t \\
&= \boldsymbol{\beta}_t + (\boldsymbol{X}^\top \boldsymbol{\Omega}_t^{-1} \boldsymbol{X}) \boldsymbol{X}^\top \boldsymbol{\Omega}_t^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_t) \\
&= (\boldsymbol{X}^\top \boldsymbol{\Omega}_t^{-1} \boldsymbol{X}) \boldsymbol{X}^\top \boldsymbol{\Omega}_t^{-1} \boldsymbol{y} \\
\boldsymbol{\gamma}_{t+1} &= \boldsymbol{\gamma}_t + \left[2(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \right] \left[\frac{1}{2} \sum_{i=1}^{n} \boldsymbol{z}_i \left(\frac{\varepsilon_{i,t}}{\exp\{\boldsymbol{z}_i^\top \boldsymbol{\gamma}_t\}} - 1 \right) \right].
\end{aligned}$$

Esboço do algoritmo:

1. Estime a variância dos distúrbios σ_i^2 com $\exp\{\boldsymbol{z}_i^\top \boldsymbol{\gamma}\}$.
2. Calcule $\boldsymbol{\beta}_{t+1}$.
3. Calcule $\boldsymbol{\gamma}_{t+1}$.
4. Calcule $\boldsymbol{d}_{t+1} = \boldsymbol{\delta}_{t+1} - \boldsymbol{\delta}_t$. Se $\boldsymbol{d}_{t+1}$ for grande, retorne ao Passo 1.

Temos também que

$$\begin{aligned}
\text{Var}(\widehat{\boldsymbol{\beta}}_{MV}) &= (\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X})^{-1} \\
\text{Var}(\widehat{\boldsymbol{\gamma}}_{MV}) &= 2(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}.
\end{aligned}$$

Testar a hipótese de homocedasticidade, $H_0: \boldsymbol{\alpha} = \mathbf{0}$, usando o teste de Wald,

$$W = \begin{pmatrix} 0 & \widehat{\boldsymbol{\alpha}}^\top \end{pmatrix} \left[\frac{1}{2} \boldsymbol{Z}^\top \boldsymbol{Z} \right]^{-1} \begin{pmatrix} 0 \\ \widehat{\boldsymbol{\alpha}}^\top \end{pmatrix} = 2\widehat{\boldsymbol{\alpha}}^\top (\boldsymbol{Z}_1^\top \widetilde{\boldsymbol{M}} \boldsymbol{Z}_1) \widehat{\boldsymbol{\alpha}} \sim \chi^2_J.$$

Exemplo

$$\ln C_{i,t} = \beta_1 + \beta_2 \ln Q_{i,t} + \beta_3 [\ln Q_{i,t}]^2 + \beta_3 \ln Pf_{i,t} + \beta_5 \text{Loadfactor}_{i,t} + \varepsilon_{i,t},$$

sendo

- $C_{i,t}$ o custo total;
- $Q_{i,t}$ a produção; e
- $Pf_{i,t}$ o preço do combustível.

Os dados são compostos de 90 observações para seis firmas observadas em 15 anos.

Para completar o modelo, temos

$$\sigma^2_{i,t} = \sigma^2 \exp\{\alpha \text{Loadfactor}_{i,t}\} = \exp\{\gamma_1 + \gamma_2 \text{Loadfactor}_{i,t}\}.$$

Tabela 19: Modelo de heterocedasticidade multiplicativa

β_1	β_2	β_3	β_4	γ_2
9,2774	0,91609	0,021643	0,40174	9,78076
(0,20977)	(0,032993)	(0,011017)	(0,016332)	(2,839)

Para testa $H_0: \alpha = 0$, temos $W = (9,78076/2,839)^2 = 11,869 > 3,84$.
Logo, rejeitamos a hipótese nula.

Modelo de fronteira estocástica

$y = f(\boldsymbol{x})$ produção dado os insumos $\boldsymbol{x}$.

Temos que $y \leqslant f(\boldsymbol{x}) \Rightarrow y = h(\boldsymbol{x}, \boldsymbol{\beta}) + u$, u deve ser negativo.

Qualquer distúrbio diferente de zero deve ser interpretado como ineficiência.

Da função Cobb-Douglas, temos

$$\ln y = \beta_1 + \sum_k \beta_k \ln x_k - u, \quad u \geqslant 0.$$

Fronteira estocástica:

$$\begin{aligned} \ln y &= \beta_1 + \sum_k \beta_k \ln x_k - u + \nu, \quad u \geqslant 0, \quad \nu \sim \mathcal{N}(0, \sigma_\nu^2) \\ &= \beta_1 + \sum_k \beta_k \ln x_k + \varepsilon. \end{aligned}$$

A fronteira de qualquer firma particular é $h(\boldsymbol{x}, \boldsymbol{\beta}) + \nu$.

Os dados estão em logaritmos, logo u é a medida da porcentagem na qual a observação particular falha em atingir a fronteira (razão de produção ideal).

Sejam $\varepsilon = \nu - u$, $\lambda = \sigma_u/\sigma_\nu$, $\sigma = (\sigma_u^2 + \sigma_\nu^2)^{1/2}$ e $\Phi(z)$ a f.d.a.

Para o modelo normal truncada,

$$\ln h(\varepsilon_i|\boldsymbol{\beta}, \lambda, \sigma) = \left[-\ln \sigma + \frac{1}{2}\ln\frac{2}{\pi} - \frac{1}{2}\left(\frac{\varepsilon_i}{\sigma}\right)^2 + \ln \Phi\left(\frac{-\varepsilon_i \lambda}{\sigma}\right)\right],$$

$$\mathrm{E}(\varepsilon) = -\sigma_u \left(\frac{2}{\pi}\right)^{1/2},$$

e para o modelo exponencial

$$\ln h(\varepsilon_i|\boldsymbol{\beta}, \lambda, \sigma) = \left[\ln\theta + \frac{1}{2}\theta^2\sigma_\nu^2 + \theta\varepsilon_i + \ln\Phi\left(\frac{-\varepsilon_i}{\sigma_\nu} - \theta\sigma\right)\right], \qquad \mathrm{E}(\varepsilon) = -\frac{1}{\theta}.$$

Fazemos $\beta_0 = \beta_1 + \mathrm{E}(\varepsilon)$ e $\varepsilon^\star = \varepsilon - \mathrm{E}(\varepsilon)$, para obter

$$\ln y = \beta_0 + \sum\nolimits_k \beta_k \ln x_k + \varepsilon^\star.$$

Para o modelo normal truncada, temos

$$\ell(\boldsymbol{\beta}, \sigma, \lambda) = -n \ln \sigma + \frac{n}{2} \ln \frac{2}{\pi} - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{\varepsilon_i}{\sigma} \right)^2 + \sum_{i=1}^{n} \ln \Phi \left(-\frac{-\varepsilon_i \lambda}{\sigma} \right).$$

Seja $\theta = 1/\sigma$ e $\boldsymbol{\gamma} = (1/\sigma)\boldsymbol{\beta}$. Então,

$$\begin{aligned} \ell(\boldsymbol{\beta}, \sigma, \lambda) &= n \ln \theta + \frac{n}{2} \ln \frac{2}{\pi} - \frac{1}{2} \sum_{i=1}^{n} \left(\theta y_i - \boldsymbol{x}_i^\top \boldsymbol{\gamma} \right)^2 \\ &+ \sum_{i=1}^{n} \ln \Phi \left(-\lambda (\theta y_i - \boldsymbol{x}_i^\top \boldsymbol{\gamma}) \right). \end{aligned}$$

Definimos:

- $\alpha_i = \varepsilon_i / \sigma = \theta y_i - \boldsymbol{x}_i^\top \boldsymbol{\gamma}$
- $\delta(y_i, \boldsymbol{x}_i, \lambda, \theta, \boldsymbol{\gamma}) = \dfrac{\phi(-\lambda \alpha_i)}{\Phi(-\lambda \alpha_i)} = \delta_i$
- $\boldsymbol{z}_i = (\boldsymbol{x}_i^\top, -y_i)^\top$, com dimensão $(p+1) \times 1$
- $\boldsymbol{t}_i = (\boldsymbol{0}^\top, 1/\theta)^\top$, com dimensão $(p+1) \times 1$.

As equações de verossimilhança são

$$\begin{aligned}
\frac{\partial \ell(\gamma, \theta, \lambda)}{\partial (\gamma^\top, \theta)^\top} &= \sum_{i=1}^{n} \boldsymbol{t}_i + \sum_{i=1}^{n} \alpha_i \boldsymbol{z}_i + \lambda \sum_{i=1}^{n} \delta_i \boldsymbol{z}_i = \boldsymbol{0} \\
\frac{\partial \ell(\gamma, \theta, \lambda)}{\partial \lambda} &= -\sum_{i=1}^{n} \delta_i \alpha_i = 0.
\end{aligned}$$

Temos também

$$\boldsymbol{H}(\gamma, \sigma, \lambda) = \sum\nolimits_{i=1}^{n} \left\{ \begin{bmatrix} (\lambda^2 \Delta_i - 1)\boldsymbol{z}_i \boldsymbol{z}_i^\top & (\delta_i - \lambda \alpha_i \Delta_i)\boldsymbol{z}_i \\ (\delta_i - \lambda \alpha_i \Delta_i)\boldsymbol{z}_i^\top & \alpha_i^2 \Delta_i \end{bmatrix} - \begin{bmatrix} \boldsymbol{t}_i \boldsymbol{t}_i^\top & \boldsymbol{0} \\ \boldsymbol{0}^\top & 0 \end{bmatrix} \right\}.$$

Segue-se que $\widehat{\text{Var}}\left((\widehat{\gamma}^\top, \widehat{\theta}, \widehat{\lambda})^\top\right) = \left[\boldsymbol{H}(\widehat{\gamma}^\top, \widehat{\theta}, \widehat{\lambda})\right]^{-1}$,

$\widehat{\sigma}_{MV} = 1/\widehat{\theta}_{MV}$ e $\widehat{\boldsymbol{\beta}}_{MV} = \dfrac{\widehat{\gamma}_{MV}}{\widehat{\theta}_{MV}}$.

Recorremos ao método delta para estimar as variâncias.

$$\boldsymbol{G} = \begin{bmatrix} \partial\widehat{\boldsymbol{\beta}}/\partial\widehat{\boldsymbol{\gamma}} & \partial\widehat{\boldsymbol{\beta}}/\partial\widehat{\theta} & \partial\widehat{\boldsymbol{\beta}}/\partial\widehat{\lambda} \\ \partial\widehat{\sigma}/\partial\widehat{\boldsymbol{\gamma}} & \partial\widehat{\sigma}/\partial\widehat{\theta} & \partial\widehat{\sigma}/\partial\widehat{\lambda} \\ \partial\widehat{\lambda}/\partial\widehat{\boldsymbol{\gamma}} & \partial\widehat{\lambda}/\partial\widehat{\theta} & \partial\widehat{\lambda}/\partial\widehat{\lambda} \end{bmatrix} = \begin{bmatrix} -(1/\widehat{\theta})\boldsymbol{I} & -(1/\widehat{\theta}^2)\widehat{\boldsymbol{\gamma}} & \mathbf{0} \\ \mathbf{0}^\top & -(1/\widehat{\theta}^2) & 0 \\ \mathbf{0}^\top & 0 & 1 \end{bmatrix}.$$

$\widehat{\text{Var}}\left((\widehat{\boldsymbol{\beta}}, \widehat{\sigma}, \widehat{\lambda})^\top\right) = \boldsymbol{G}\left[\boldsymbol{H}(\widehat{\boldsymbol{\gamma}}^\top, \widehat{\theta}, \widehat{\lambda})\right]^{-1}\boldsymbol{G}^\top$. Também podemos estimar $\widehat{\sigma}_\nu^2 = \dfrac{\widehat{\sigma}^2}{1+\widehat{\lambda}^2}$ e $\widehat{\sigma}_u^2 = \dfrac{\widehat{\sigma}^2\widehat{\lambda}^2}{1+\widehat{\lambda}^2}$.

Em geral, estamos interessados nas eficiências u_i. Mas obtemos somente $\varepsilon = \boldsymbol{y} - \boldsymbol{x}^\top\boldsymbol{\beta}$.

Podemos, entretanto, utilizar a seguinte aproximação:

$$\text{E}(u|\varepsilon) = \frac{\sigma\lambda}{1+\lambda^2}\left[\frac{\phi(z)}{1-\Phi(z)} - z\right], \qquad z = \frac{\varepsilon\lambda}{\sigma},$$

para o modelo normal truncada e

$$\mathrm{E}(u|\varepsilon) = z + \sigma_\nu \frac{\phi(z/\sigma_\nu)}{\Phi(z/\sigma_\nu)}, \qquad z = -\varepsilon - \theta\sigma_\nu^2,$$

para o modelo exponencial.

Podemos estimar estas quantidades com as estimativas de máxima verossimilhança.

Estamos interessados também na variância relativa. Para o modelo normal truncada,

$$\mathrm{Var}(\varepsilon) = \mathrm{Var}(u) + \mathrm{Var}(\nu) = (1 - 2\pi)\sigma_u^2 + \sigma_\nu^2$$

e para o modelo exponencial

$$\mathrm{Var}(\varepsilon) = \frac{1}{\theta^2} + \sigma_\nu^2.$$

Mais uma vez, podemos utilizar as estimativas de máxima verossimilhança para estimar estas quantidades.

O melhor neste caso seria aplicar a estimação bayesiana dos parâmetros do modelo, inclusive a estimação de todas eficiências u_i.

Estimação e inferência bayesiana

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{L(\boldsymbol{\theta}|\boldsymbol{y})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \\ &\propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y})p(\boldsymbol{\theta}), \end{aligned}$$

sendo

- $p(\boldsymbol{\theta}|\boldsymbol{y})$ a distribuição a posteriori;
- $p(\boldsymbol{\theta})$ a distribuição a priori;
- $p(\boldsymbol{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y})$ a função de verossimilhança; e
- $p(\boldsymbol{y})$ a distribuição marginal dos dados ou a verossimilhança marginal.

Temos,

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Além disso, $p(\boldsymbol{y}) < \infty$ para existência da distribuição a posteriori.

Análise bayesiana do modelo de regressão clássico

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

Logo,

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\}.$$

Seja $d = n - p$ os graus de liberdade e

$$\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} - \boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{b}) = \boldsymbol{e} - \boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{b}).$$

Então,

$$-\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \left(-\frac{ds^2}{2} \right) \sigma^{-2} + \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{b})^\top [\sigma^{-2} \boldsymbol{X}^\top \boldsymbol{X}] (\boldsymbol{\beta} - \boldsymbol{b}).$$

Logo,

$$
\begin{aligned}
L(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) &= (2\pi)^{-d/2} (\sigma^2)^{-d/2} \exp\left\{ -\frac{d s^2}{2\sigma^2} \right\} \\
&\quad \times (2\pi)^{-p/2} (\sigma^2)^{-p/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{b})^\top [\sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}]^{-1} (\boldsymbol{\beta} - \boldsymbol{b}) \right\} \\
&\propto \frac{(\nu s^2)^{\nu-1}}{\Gamma(\nu - 1)} (\sigma^2)^{-\nu} \exp\left\{ -\frac{\nu s^2}{\sigma^2} \right\} \\
&\quad \times (2\pi)^{-p/2} \left| \sigma^2 \boldsymbol{\Delta} \right|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{b})^\top [\sigma^2 \boldsymbol{\Delta}]^{-1} (\boldsymbol{\beta} - \boldsymbol{b}) \right\},
\end{aligned}
$$

sendo $n/2 = d/2 + p/2$, $\nu = d/2$ e $\boldsymbol{\Delta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

A verossimilhança é proporcional ao produto de uma inversa gama (σ^2) com parâmetros $\delta = \nu - 1$ e $\lambda = \nu s^2$, e uma normal p dimensional $(\boldsymbol{\beta} | \sigma^2)$ com média $\boldsymbol{b}$ e matriz de covariância $\sigma^2 \boldsymbol{\Delta}$.

Podemos considerar uma priori não informativa,

$$
p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}.
$$

$$
\begin{aligned}
p(\beta, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) &\propto L(\beta, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) p(\beta, \sigma^2) \\
&\propto \frac{(\nu s^2)^\nu}{\Gamma(\nu)} (\sigma^2)^{-(\nu+1)} \exp\left\{ -\frac{\nu s^2}{\sigma^2} \right\} \\
&\times (2\pi)^{-p/2} \left| \sigma^2 \boldsymbol{\Delta} \right|^{-1/2} \exp\left\{ -\frac{1}{2} (\beta - \boldsymbol{b})^\top [\sigma^2 \boldsymbol{\Delta}]^{-1} (\beta - \boldsymbol{b}) \right\}.
\end{aligned}
$$

Agora seja $A(\beta) = (\beta - \boldsymbol{b})^\top \boldsymbol{\Delta}^{-1} (\beta - \boldsymbol{b})/2$, então

$$
\begin{aligned}
p(\beta | \boldsymbol{y}, \boldsymbol{X}) &= \int p(\beta, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) d\sigma^2 \propto \int_0^\infty (\sigma^2)^{-(n/2+1)} \exp\left\{ -\frac{A(\beta) + \nu s^2}{\sigma^2} \right\} d\sigma^2 \\
&= \frac{\Gamma(n/2)}{[A(\beta) + \nu s^2]^{n/2}} \propto \Gamma(n/2) \left[1 + \frac{1}{2\nu s^2} A(\beta) \right]^{-((n-p)+p)/2} \\
&= \Gamma(n/2) \left[1 + \frac{1}{2\nu} (\beta - \boldsymbol{b})^\top [s^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}]^{-1} (\beta - \boldsymbol{b}) \right]^{-((n-p)+p)/2}.
\end{aligned}
$$

Assim,

$$
(\beta | \boldsymbol{y}, \boldsymbol{X}) \sim t_{n-p} \left(\boldsymbol{b}, s^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \right) \qquad (p \text{ dimensional}).
$$

É fácil mostrar que

$$(\sigma^2|\boldsymbol{y},\boldsymbol{X}) \sim \mathcal{IG}(\nu, \nu s^2).$$

Logo,

$$\begin{aligned} \mathrm{E}(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) &= \boldsymbol{b}, \quad \text{para} \quad n-p>1, \\ \mathrm{E}(\sigma^2|\boldsymbol{y},\boldsymbol{X}) &= \frac{\nu}{\nu-1}s^2, \quad \text{para} \quad \nu>1 \ (n-p>2). \end{aligned}$$

Podemos considerar também uma priori informativa. Por exemplo,

$$(\boldsymbol{\beta}|\sigma^2) \sim \mathcal{N}_p(\boldsymbol{b}_0, \sigma^2\boldsymbol{B}_0) \quad \text{e} \quad \sigma^2 \sim \mathcal{IG}(m, m\sigma_0^2) \quad \text{(priori normal gama)}.$$

$$\begin{aligned} p(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X}) &\propto (\sigma^2)^{-(\nu+m+1)} \exp\left\{-\frac{1}{\sigma^2}(m\sigma_0^2+\nu s^2)\right\} \\ &\times \left|\boldsymbol{\Sigma}\right|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{a})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{a})\right\}, \end{aligned}$$

sendo

$$\begin{aligned} \boldsymbol{\Sigma} &= \left[\sigma^{-2}B_0^{-1}+\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X}\right]^{-1} = \sigma^2\left[B_0^{-1}+\boldsymbol{X}^\top\boldsymbol{X}\right]^{-1} \quad \text{e} \\ \boldsymbol{a} &= \boldsymbol{\Sigma}\left(\sigma^{-2}\boldsymbol{B}_0^{-1}\boldsymbol{b}_0+\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{b}\right). \end{aligned}$$

Assim,

$$\begin{aligned}(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) &\sim t_{n-p}(\boldsymbol{a},\boldsymbol{\Sigma}^{\star}) \quad \text{e}\\ (\sigma^2|\boldsymbol{y},\boldsymbol{X}) &\sim \mathcal{IG}(\nu+m, m\sigma_0^2+\nu\boldsymbol{s}^2).\end{aligned}$$

Exercício

Encontre a expressão exata de $\boldsymbol{\Sigma}^{\star}$.

Em problemas mais complicados, teremos que recorrer a métodos numéricos, como a amostragem de Gibbs.

Para utilizar a amostragem de Gibbs no modelo de regressão com priori não informativa, temos

$$\begin{aligned}(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}) &\sim \mathcal{N}_p(\boldsymbol{b},\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}) \quad \text{e}\\ (\sigma^2|\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X}) &\sim \mathcal{IG}(n/2, C(\boldsymbol{\beta})),\end{aligned}$$

sendo $C(\boldsymbol{\beta}) = \nu\boldsymbol{s}^2 + (\boldsymbol{\beta}-\boldsymbol{b})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{b})$.

Modelo probit

Considere o modelo probit

$$\begin{aligned} y_i^\star &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,1), \\ y_i &= 1 \quad \text{if} \quad y_i^\star > 0, \quad \text{caso contrário} \quad y_i = 0. \end{aligned}$$

Temos também que

$$Pr(y_i = 1|\boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \quad \text{e} \quad Pr(y_i = 0|\boldsymbol{x}_i) = 1 - \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

A função de verossimilhança

$$L(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right]^{y_i} \left[1 - \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right]^{1-y_i}$$

e teremos a prior como $p(\boldsymbol{\beta}) \propto 1$. Assim,

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) = \frac{\prod_{i=1}^{n} \left[\Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right]^{y_i} \left[1 - \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right]^{1-y_i}}{\int \prod_{i=1}^{n} \left[\Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right]^{y_i} \left[1 - \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right]^{1-y_i} d\boldsymbol{\beta}}.$$

A média a posteriori é dada por

$$\mathrm{E}(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \frac{\displaystyle\int \boldsymbol{\beta} \prod_{i=1}^{n} \left[\Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right]^{y_i} \left[1-\Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right]^{1-y_i} d\boldsymbol{\beta}}{\displaystyle\int \prod_{i=1}^{n} \left[\Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right]^{y_i} \left[1-\Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right]^{1-y_i} d\boldsymbol{\beta}}.$$

O cálculo das integrais acima é complicado.

Consideraremos a técnica conhecida como variáveis auxiliares.

Trataremos $y_i^\star$ como quantidades não observadas a serem estimadas junto com $\boldsymbol{\beta}$.

Primeiro, temos que $p(\boldsymbol{\beta}|\boldsymbol{y}^\star,\boldsymbol{y},\boldsymbol{X}) = p(\boldsymbol{\beta}|\boldsymbol{y}^\star,\boldsymbol{X})$ e também

$$(\boldsymbol{\beta}|\boldsymbol{y}^\star,\boldsymbol{y},\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{b}^\star,(\boldsymbol{X}^\top\boldsymbol{X})^{-1}) \quad \text{sendo} \quad \boldsymbol{b}^\star = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}^\star.$$

Agora, sabemos que $(y_i^\star|\boldsymbol{\beta}, \boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, 1)$. Mas y_i tem informação sobre $y_i^\star$. Logo,

$$(y_i^\star|y_i = 1, \boldsymbol{\beta}, \boldsymbol{x}_i) \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, 1) I(0, +\infty)$$
$$(y_i^\star|y_i = 0, \boldsymbol{\beta}, \boldsymbol{x}_i) \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, 1) I(-\infty, 0).$$

Com estas equações podemos utilizar a amostragem de Gibbs.

Portanto, $\mathrm{E}(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})$ e $\mathrm{E}(\boldsymbol{y}^\star|\boldsymbol{y}, \boldsymbol{X})$ pode ser estimado com a amostra da posteriori.

Exemplo

Temos o modelo,

$$Pr(GRADE_i = 1|\boldsymbol{\beta}, \boldsymbol{x}_i) = \Phi(\beta_1 + \beta_2 GPA_i + \beta_3 TUCE_i + \beta_4 PSI_i),$$

sendo

- GPA_i a média dos pontos do estudante;
- $GRADE_i$ uma variável dicotômica indicando se a nota do estudante no curso de macroeconomia intermediário foi mais alta que nos cursos iniciais;
- PSI_i uma variável dicotômica indicando se o indivíduo participou do *Personalized System Instruction*; e
- $TUCE_i$ a nota do estudante em um teste de economia.

Tabela 20: Estimação do modelo probit para as notas.

Variável	EMV		Bayesiano	
	Estimativa	E.P.	Média	D.P.
Constante	-7,4523	2,5425	-8,6286	2,7995
GPA	1,6258	0,6939	1,8754	0,7668
TUCE	0,0517	0,0839	0,0628	0,0870
PSI	1,4263	0,5950	1,6072	0,6257

Estimação consistente: o método dos momentos

- Momentos populacionais e amostrais.
- Os momentos populacionais dependem dos parâmetros que estamos interessados em estimar.
- A amostra contém informações em forma de estatísticas como momentos.
- Se o modelo tiver p parâmetros, igualamos os p momentos populacionais aos p momentos amostrais correspondentes.
- Então, escrevemos os parâmetros como funções dos momentos amostrais (das estatísticas).
- Os momentos serão consistentes pela lei dos grandes números.
- Os momentos serão assintoticamente normais pelo Teorema de Lindeberg-Levy.
- Aplicação do teorema de Slutsky garante a propriedade dos estimadores dos parâmetros.

Seja $m_k(.)$ uma função contínua e diferenciável que não depende do tamanho de amostra n e seja

$$\overline{m}_k = \frac{1}{n}\sum_{i=1}^{n} m_k(y_i), \qquad k = 1, \ldots, p.$$

Segue-se que

$$\text{plim } \overline{m}_k = \mathrm{E}(m_k(y_i)) = \mu(\theta_1, \ldots, \theta_p).$$

As p equações dos momentos são

$$\begin{aligned}
\overline{m}_{n,1}(\boldsymbol{\theta}) &= \overline{m}_1 - \mu_1(\theta_1, \ldots, \theta_p) = 0 \\
\overline{m}_{n,2}(\boldsymbol{\theta}) &= \overline{m}_2 - \mu_2(\theta_1, \ldots, \theta_p) = 0 \\
&\vdots \quad \vdots \qquad \vdots \quad \vdots \\
\overline{m}_{n,p}(\boldsymbol{\theta}) &= \overline{m}_p - \mu_p(\theta_1, \ldots, \theta_p) = 0.
\end{aligned}$$

- Na maioria dos casos os estimadores pelo método dos momentos não são eficientes.
- A exceção é na amostragem aleatória das distribuições na família exponencial.

Definição (família exponencial)

Uma distribuição pertence a família exponencial se tem o logaritmo da função de verossimilhança dada por

$$\ln L(\theta|\textbf{dados}) = a(\textbf{dados}) + b(\theta) + \sum_{k=1}^{p} c_k(\textbf{dados}) s_k(\theta),$$

sendo $a(.)$, $b(.)$, $c_k(.)$ e $s_k(.)$ funções.

Lembremos que $c_k(\textbf{dados})$ é uma estatística suficiente.

Em geral, consideramos os momentos como

$$\overline{m}_k = \frac{1}{n}\sum_{i=1}^{n} m_k(\boldsymbol{y}_i), \quad k = 1, \ldots, p.$$

Um estimador apropriado da matriz de covariância assintótica de $\overline{\boldsymbol{m}} = (\overline{m}_1, \ldots, \overline{m}_p)$ é

$$\frac{1}{n}F_{jk} = \frac{1}{n}\left\{\frac{1}{n}\sum_{i=1}^{n}[m_j(\boldsymbol{y}_i) - \overline{m}_j][m_k(\boldsymbol{y}_i) - \overline{m}_k]\right\}, \quad j, k = 1, 2, \ldots, p.$$

Agora, seja $\overline{\boldsymbol{G}}(\boldsymbol{\theta})$ a matriz $p \times p$ cuja k-ésima linha é o vetor de derivadas parciais

$$\overline{\boldsymbol{G}}_k^{\top} = \frac{\partial \overline{m}_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\top}}.$$

Da expansão de Taylor em torno do valor verdadeiro $\boldsymbol{\theta}_0$, temos

$$\mathbf{0} \simeq \overline{\boldsymbol{m}}(\boldsymbol{\theta}_0) + \overline{\boldsymbol{G}}^{\top}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Portanto,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \simeq -[\overline{\boldsymbol{G}}^{\top}(\boldsymbol{\theta}_0)]^{-1}\sqrt{n}[\overline{\boldsymbol{m}}(\boldsymbol{\theta}_0)].$$

Usando esta aproximação, pode-se provar que

$$\widehat{\boldsymbol{\theta}} \approx \mathcal{N}(\boldsymbol{\theta}_0, \boldsymbol{\Delta}),$$

sendo $\boldsymbol{\Delta} = \frac{1}{2}\{-[\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1}\}\boldsymbol{\Phi}\{-[\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1}\}$.

Podemos estimar a variância assintótica $\boldsymbol{\Delta}$ por

$$\widehat{\text{Var}}(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\left[\overline{\boldsymbol{G}}^{\top}(\widehat{\boldsymbol{\theta}})]\boldsymbol{F}^{-1}\overline{\boldsymbol{G}}(\widehat{\boldsymbol{\theta}})\right]^{-1}.$$

GMM sob a condição de ortogonalidade

Considere o estimador OLS no modelo de regressão clássico. Por hipótese,

$$\mathrm{E}(\boldsymbol{x}_i \varepsilon_i) = \mathrm{E}(\boldsymbol{x}_i (y_i - \boldsymbol{x}_i^\top \beta)) = \mathbf{0}.$$

Na amostra, temos

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \widehat{\varepsilon}_i = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i^\top \widehat{\beta}) = \mathbf{0}.$$

Podemos ver que o estimador OLS é um estimador do método dos momentos.

Para o estimador de variáveis instrumentais,

$$\text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\varepsilon_i\right) = \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})\right).$$

Na amostra, temos

$$\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{Z}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}^\top\widehat{\boldsymbol{\varepsilon}}\right) = \frac{1}{n}\widehat{\boldsymbol{X}}^\top\widehat{\boldsymbol{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{x}}_i\widehat{\varepsilon}_i = \boldsymbol{0}.$$

Podemos ver que este estimador é um estimador do método dos momentos.

Estimadores de máxima verossimilhança são obtidos ao igualar a derivada da log-verossimilhança a zero.

A função de verossimilhança ponderada é

$$\frac{1}{n} \ln L(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}).$$

Sabemos que

$$\mathrm{E}\left(\frac{\partial \ln f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) = \mathbf{0}.$$

Na amostra, temos

$$\frac{1}{n} \frac{\ln L(\widehat{\boldsymbol{\theta}}|\boldsymbol{y}, \boldsymbol{X})}{\partial \widehat{\boldsymbol{\theta}}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})}{\partial \widehat{\boldsymbol{\theta}}} = \mathbf{0}.$$

Podemos também ver que este estimador é um estimador do método dos momentos.

Generalizando o método dos momentos

- Existem casos em que temos mais equações de momentos do que parâmetros. Logo, o sistema é sobre identificado.
- Para utilizar toda informação na amostra é necessário criar um critério para evitar os confitos que podem existir em um sistema sobre identificado.
- Suponha que o modelo envolva p parâmetros, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$, e que a teoria fornece um conjunto de $L > p$ condições de momentos,

$$\mathrm{E}(m_\ell(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta})) = \mathrm{E}(m_{i\ell}(\boldsymbol{\theta})) = 0,$$

  sendo y_i, $\boldsymbol{x}_i$ e $\boldsymbol{z}_i$ variáveis que aparecem no modelo.
- Denotaremos os correspondentes momentos amostrais por

$$\overline{m}_\ell(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_\ell(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_{i\ell}(\boldsymbol{\theta}).$$

- A menos que as equações sejam funcionalmente dependentes, o sistema de L equações em p parâmetros desconhecidos,

$$\overline{m}_\ell(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_\ell(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}) = 0, \quad \ell = 1, \ldots, L,$$

  não terá uma solução única.
- Uma das possibilidades é minimizar a soma de quadrados,

$$q = \sum_{\ell=1}^{L} \overline{m}_\ell^2 = \overline{\boldsymbol{m}}(\boldsymbol{\theta})^\top \overline{\boldsymbol{m}}(\boldsymbol{\theta}).$$

- Pode ser mostrado que plim $\overline{\boldsymbol{m}}(\boldsymbol{\theta}) = \mathrm{E}(\overline{\boldsymbol{m}}(\boldsymbol{\theta})) = \mathbf{0}$, isto é, o minimizador de q é um estimador consistente de $\boldsymbol{\theta}$.
- Podemos também utilizar mínimos quadrados ponderados,

$$q = \overline{\boldsymbol{m}}(\boldsymbol{\theta})^\top \boldsymbol{W}_n \overline{\boldsymbol{m}}(\boldsymbol{\theta}),$$

  sendo $\boldsymbol{W}_n$ qualquer matriz positiva definida que depende dos dados mas não dos parâmetros. Supomos também que plim$\boldsymbol{W}_n = \boldsymbol{W}$, uma matriz positiva definida.

- Podemos utilizar uma matriz $\boldsymbol{W}_n$ como diagonal com elementos que são os inversos das variâncias individuais dos momentos,

$$w_{\ell\ell} = \frac{1}{\text{Var}(\sqrt{n}\overline{m}_\ell)} = \frac{1}{\phi_{\ell\ell}}.$$

- O estimador de mínimos quadrados ponderados deve minimizar

$$q = \overline{\boldsymbol{m}}(\boldsymbol{\theta})^\top \boldsymbol{\Phi}^{-1} \overline{\boldsymbol{m}}(\boldsymbol{\theta}).$$

- Em geral, os L elementos de $\overline{\boldsymbol{m}}$ podem ser correlacionados.
- Para utilizar mínimos quadrados generalizados, podemos definir a matriz,

$$\boldsymbol{W} = \left[\text{Var}(\sqrt{n}\overline{\boldsymbol{m}})\right]^{-1} = \boldsymbol{\Phi}^{-1}.$$

- Se $\boldsymbol{W}_n$ é uma matriz positiva definida e se plim $\overline{\boldsymbol{m}}(\boldsymbol{\theta}) = \boldsymbol{0}$, então o estimador GMM é consistente.

- A matriz de covariância assintótica do estimador GMM é

$$\boldsymbol{V}_{GMM} = \frac{1}{n}\left[\boldsymbol{\Gamma}^\top \boldsymbol{W}\boldsymbol{\Gamma}\right]^{-1} = \frac{1}{n}\left[\boldsymbol{\Gamma}^\top \boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}\right]^{-1},$$

  sendo $\boldsymbol{\Gamma}$ a matriz de derivadas com a j-ésima linha igual a

$$\boldsymbol{\Gamma}^j = \text{plim}\ \frac{\partial \overline{m}_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top},$$

  e $\boldsymbol{\Phi} = \text{Var}(\sqrt{n}\overline{\boldsymbol{m}})$.

- Por hipótese, temos que

$$\sqrt{n}\overline{\boldsymbol{m}}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi}).$$

Teorema

A distribuição assintótica para o estimador GMM é

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{GMM} &\xrightarrow{p} \boldsymbol{\theta} \\ \widehat{\boldsymbol{\theta}}_{GMM} &\approx \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{V}_{GMM}). \end{aligned}$$

Estimação GMM em modelos econométricos

Consideraremos o modelo de regressão com variáveis instrumentais,

$$\begin{aligned} y_i &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i \\ \mathrm{E}(\boldsymbol{z}_i \varepsilon_i) &= \mathbf{0}, \end{aligned}$$

para p variáveis em $\boldsymbol{x}_i$ e um conjunto de $L \geqslant p$ variáveis instrumentais $\boldsymbol{z}_i$.

O resultado da regressão generalizada aparece se $\boldsymbol{z}_i = \boldsymbol{x}_i$ e a regressão clássica ao adicionar que $\boldsymbol{\Omega} = \boldsymbol{I}$.

Consideraremos três casos:

- Regressão clássica: $\mathrm{Var}(\varepsilon_i | \boldsymbol{X}, \boldsymbol{Z}) = \sigma^2$,
- Heterocedasticidade: $\mathrm{Var}(\varepsilon_i | \boldsymbol{X}, \boldsymbol{Z}) = \sigma_i^2$,
- Regressão generalizada: $\mathrm{Cov}(\varepsilon_t, \varepsilon_s | \boldsymbol{X}, \boldsymbol{Z}) = \sigma^2 \omega_{ts}$.

A hipótese $\mathrm{E}(\boldsymbol{z}_i \varepsilon_i) = 0$ implica a condição de ortogonalidade

$$\mathrm{Cov}(\boldsymbol{z}_i, \varepsilon_i) = \mathbf{0} \quad \text{ou} \quad \mathrm{E}(\boldsymbol{z}_i (y_i - \boldsymbol{x}_i^\top \beta)) = \mathbf{0}.$$

As equações de momentos da população são

$$\mathrm{E}\left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_i (y_i - \boldsymbol{x}_i^\top \beta) \right) = \mathrm{E}(\overline{\boldsymbol{m}}(\beta)) = \mathbf{0}.$$

As equações de momentos da amostra são

$$\left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_i (y_i - \boldsymbol{x}_i^\top \widehat{\beta}) \right] = \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{m}_i (\widehat{\beta}) \right] = \overline{\boldsymbol{m}}(\widehat{\beta}) = \mathbf{0}.$$

Para os casos em que o modelo é identificável, podemos escrever

$$\overline{\boldsymbol{m}}(\widehat{\beta}) = \left(\frac{1}{n} \boldsymbol{z}^\top \boldsymbol{y} \right) - \left(\frac{1}{n} \boldsymbol{Z}^\top \boldsymbol{X} \right) \widehat{\beta}.$$

Se o modelo for exatamente identificável ($L = p$), então pode ser mostrado que

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top \boldsymbol{X})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}.$$

Se o modelo for sobre identificável ($L > p$), então o sistema $\overline{\boldsymbol{m}}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ não tem uma solução única.

Podemos escolher o critério de mínimos quadrados,

$$\min_{\widehat{\boldsymbol{\beta}}} q = \min_{\widehat{\boldsymbol{\beta}}} \overline{\boldsymbol{m}}(\widehat{\boldsymbol{\beta}})^\top \overline{\boldsymbol{m}}(\widehat{\boldsymbol{\beta}}).$$

Neste caso, as condições de primeira ordem são

$$\begin{aligned} \frac{\partial q}{\partial \boldsymbol{\beta}} &= 2 \left(\frac{\partial \overline{\boldsymbol{m}}(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \right) \overline{\boldsymbol{m}}(\widehat{\boldsymbol{\beta}}) = 2 \overline{\boldsymbol{G}}(\widehat{\boldsymbol{\beta}})^\top \overline{\boldsymbol{m}}(\widehat{\boldsymbol{\beta}}) \\ &= 2 \left(\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{Z} \right) \left(\frac{1}{n} \boldsymbol{Z}^\top \boldsymbol{y} - \frac{1}{n} \boldsymbol{Z}^\top \boldsymbol{X} \widehat{\boldsymbol{\beta}} \right) = \mathbf{0}, \end{aligned}$$

cuja solução é

$$\widehat{\boldsymbol{\beta}} = [(\boldsymbol{X}^\top \boldsymbol{Z})(\boldsymbol{Z}^\top \boldsymbol{X})]^{-1} (\boldsymbol{X}^\top \boldsymbol{Z})(\boldsymbol{Z}^\top \boldsymbol{y}).$$

Hipóteses sobre estimador GMM:

- Convergência dos momentos: plim $\overline{\boldsymbol{m}}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$.
- Identificação para estimação GMM: a matriz $L \times p$

$$\boldsymbol{\Gamma}(\boldsymbol{\beta}) = \mathrm{E}(\overline{\boldsymbol{G}}(\boldsymbol{\beta})) = \operatorname{plim} \overline{\boldsymbol{G}}(\boldsymbol{\beta}) = \operatorname{plim} \frac{\partial \overline{\boldsymbol{m}}}{\partial \boldsymbol{\beta}^\top} = \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \overline{\boldsymbol{m}}_i}{\partial \boldsymbol{\beta}^\top}$$

  deve ter posto de linha igual a p.
- Os momentos amostrais convergem para uma distribuição normal.

É possível mostrar que sob certas condições o estimador GMM é consistente.

A matriz de covariância assintótica de $\widehat{\boldsymbol{\beta}}$ é dada por

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}[\boldsymbol{\Gamma}^\top \boldsymbol{\Gamma}]^{-1} \boldsymbol{\Gamma}^\top \mathrm{Var}(\sqrt{n}\overline{\boldsymbol{m}}(\boldsymbol{\beta}))\boldsymbol{\Gamma}[\boldsymbol{\Gamma}^\top \boldsymbol{\Gamma}]^{-1}.$$

Para o modelo de regressão em questão, temos

$$\begin{aligned} \overline{\boldsymbol{m}} &= \frac{1}{n}(\boldsymbol{Z}^\top \boldsymbol{y} - \boldsymbol{Z}^\top \boldsymbol{Z}\beta), \\ \overline{\boldsymbol{G}}(\beta) &= \frac{1}{n}\boldsymbol{Z}^\top \boldsymbol{X}, \\ \boldsymbol{\Gamma}(\beta) &= \boldsymbol{Q}_{\boldsymbol{ZX}}. \end{aligned}$$

Precisamos, agora, determinar a forma de

$$\boldsymbol{V} = \text{Var}(\sqrt{n}\overline{\boldsymbol{m}}(\beta)).$$

Dado a forma de $\overline{\boldsymbol{m}}$ acima, temos

$$\boldsymbol{V} = \frac{1}{n}\text{Var}\left(\sum_{i=1}^{n} \boldsymbol{z}_i \varepsilon_i\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \sigma^2 \omega_{ij} \boldsymbol{z}_i \boldsymbol{z}_j^\top = \sigma^2 \frac{\boldsymbol{Z}^\top \boldsymbol{\Omega} \boldsymbol{Z}}{n}$$

para o caso mais geral.

Agora, precisamos estimar a variância assintótica. Para isto, precisamos estimar $\boldsymbol{V}$.

O estimador de $\boldsymbol{V}$ é dado por

- Regressão clássica:

$$\widehat{\boldsymbol{V}} = \frac{(\boldsymbol{e}^\top \boldsymbol{e}/n)}{n} \sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^\top = \frac{(\boldsymbol{e}^\top \boldsymbol{e}/n)}{n} \boldsymbol{Z}^\top \boldsymbol{Z}.$$

- Regressão heterocedástica:

$$\widehat{\boldsymbol{V}} = \frac{1}{n} \sum_{i=1}^{n} e_i^2 \boldsymbol{z}_i \boldsymbol{z}_i^\top.$$

- Regressão generalizada:

$$\begin{aligned} \widehat{\boldsymbol{V}} &= \frac{1}{n} \left[\sum_{t=1}^{n} e_t^2 \boldsymbol{z}_t \boldsymbol{z}_t^\top \right. \\ &+ \left. \sum_{\ell=1}^{n} \left(1 - \frac{\ell}{p+1} \right) \sum_{t=\ell+1}^{n} e_t e_{t-\ell} (\boldsymbol{z}_t \boldsymbol{z}_{t-\ell} + \boldsymbol{z}_{t-\ell} \boldsymbol{z}_t^\top) \right]. \end{aligned}$$

Entretanto, se os elementos de $\overline{\boldsymbol{m}}(\beta)$ forem não correlacionados, podemos utilizar

$$\widehat{\boldsymbol{V}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{m}_i(\widehat{\beta}) \boldsymbol{m}_i(\widehat{\beta})^\top.$$

Assim, estimamos a variância assintótica por

$$\begin{aligned}\widehat{\text{Var}}(\widehat{\beta}) &= \frac{1}{n}\left[\overline{\boldsymbol{G}}(\widehat{\beta})^\top \overline{\boldsymbol{G}}(\widehat{\beta})\right]^{-1} \overline{\boldsymbol{G}}(\widehat{\beta})^\top \widehat{\boldsymbol{V}}\overline{\boldsymbol{G}}(\widehat{\beta}) \left[\overline{\boldsymbol{G}}(\widehat{\beta})^\top \overline{\boldsymbol{G}}(\widehat{\beta})\right]^{-1} \\ &= n\left[(\boldsymbol{X}^\top \boldsymbol{Z})(\boldsymbol{Z}^\top \boldsymbol{X})\right]^{-1} (\boldsymbol{X}^\top \boldsymbol{Z})\widehat{\boldsymbol{V}}(\boldsymbol{Z}^\top \boldsymbol{X}) \left[(\boldsymbol{X}^\top \boldsymbol{Z})(\boldsymbol{Z}^\top \boldsymbol{X})\right]^{-1}.\end{aligned}$$

Agora, queremos utilizar o critério de mínimos quadrados ponderados

$$\min_{\widehat{\beta}} q = \min_{\widehat{\beta}} \overline{\boldsymbol{m}}(\widehat{\beta})^\top \boldsymbol{W}\overline{\boldsymbol{m}}(\widehat{\beta})$$

sendo $\boldsymbol{W}$ qualquer matriz positiva definida de ponderação. Assim, temos

$$\widehat{\text{Var}}(\widehat{\beta}) = \frac{1}{n}\left[\overline{\boldsymbol{G}}^\top \boldsymbol{W}\overline{\boldsymbol{G}}\right]^{-1} \overline{\boldsymbol{G}}^\top \boldsymbol{W}\widehat{\boldsymbol{V}}\boldsymbol{W}\overline{\boldsymbol{G}} \left[\overline{\boldsymbol{G}}^\top \boldsymbol{W}\overline{\boldsymbol{G}}\right]^{-1}.$$

No caso anterior, tínhamos $\boldsymbol{W} = \boldsymbol{I}$.

Se $L = p$, então $\boldsymbol{W}$ é irrelevante para a solução. Então, baseado na simplicidade, o valor ótimo de $\boldsymbol{W}$ é $\boldsymbol{I}$.

No caso de sobre identificação o valor ótimo de W é V^{-1}. O estimador resultante será o mais eficiente. Logo,

$$\widehat{\beta}_{GMM} = [(\boldsymbol{X}^\top \boldsymbol{Z})\widehat{\boldsymbol{V}}^{-1}(\boldsymbol{Z}^\top \boldsymbol{X})]^{-1}(\boldsymbol{X}^\top \boldsymbol{Z})\widehat{\boldsymbol{V}}^{-1}(\boldsymbol{Z}^\top \boldsymbol{y})$$

e

$$\text{Var}(\widehat{\beta}_{GMM}) = \frac{1}{n}[\overline{\boldsymbol{G}}^\top \boldsymbol{V}^{-1}\overline{\boldsymbol{G}}]^{-1} = [(\boldsymbol{X}^\top \boldsymbol{Z})\boldsymbol{V}^{-1}(\boldsymbol{Z}^\top \boldsymbol{X})]^{-1}.$$

Concluindo, o estimador GMM é dado pela solução de

$$\min_{\beta} q = \min_{\beta} \overline{\boldsymbol{m}}(\beta)^\top [\text{Var}(\sqrt{n}\overline{\boldsymbol{m}}(\beta))]^{-1}\overline{\boldsymbol{m}}(\beta).$$

que sugere que a matriz de pesos é uma função de β - o vetor de parâmetros que estamos interessados em estimar.

A estimação GMM é, em geral, dada em dois passos. No Passo 1 estimamos $\boldsymbol{V}$ e no Passo 2 estimamos β utilizando a inversa de $\widehat{\boldsymbol{V}}$.

Passo 1: Utilize $\boldsymbol{W} = \boldsymbol{I}$ para obter um estimador consistente de $\boldsymbol{\beta}$. Então, estime $\boldsymbol{V}$ com

$$\widehat{\boldsymbol{V}} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{z}_i \boldsymbol{z}_i^\top$$

no caso de heterocedasticidade (o estimador de White) ou, para o caso mais geral, o estimador de Newey-West.

Passo 2: Utilize $\boldsymbol{W} = \widehat{\boldsymbol{V}}^{-1}$ para calcular o estimador GMM de $\boldsymbol{\beta}$.