ANÁLISE DE REGRESSÃO

Ralph dos Santos Silva

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

Referências (livros)

- *A Modern Approach to Regression with R*, Sheather.
- *An R Companion to Applied Regression*, 3ª edição, Fox e Weisberg.
- ***Applied Regression Analysis,* 3ª edição, Draper e Smith.**
- *Applied Regression Modeling*, 3ª edição, Pardoe.
- *Bayesian and Frequentist Regression Methods*, Wakefield.
- *Introduction to Linear Regression Analysis*, 6ª edição, Montgomery, Peck e Vining.
- *Linear Regression and Correlation: A Beginner's Guide*, Hartshorn.
- *Regression Analysis and Linear Models: Concepts, Applications, and Implementation*, Darlington e Hayes.
- *Regression Analysis by Example*, 5ª edição, Chatterjee e Hadi.
- *Regression Analysis with Python*, Massaron e Boschetti.
- *Regression Analysis with R*, Ciaburro.
- *Regression & Linear Modeling: Best Practices and Modern Methods*, Osborne.
- *Regression Modeling with Actuarial and Financial Applications*, Frees.

Regressão

Em **análise de regressão** estamos interessados em estudar, por exemplo, relações da forma:

$$y = f(\boldsymbol{x}) + \varepsilon,$$

em que

- $\boldsymbol{x}$ é um vetor de variáveis chamadas de *preditoras*, *explicativas* ou *regressoras*;
- y é a variável chamada de *dependente* ou *resposta*;
- a forma funcional $f(\boldsymbol{x})$ depende de quantidades desconhecidas β (*parâmetros*); e
- ε é um ruído aleatório.

Nota: evitaremos a utilização do termo **variáveis independentes** para $\boldsymbol{x}$.

Em geral, a forma funcional de $f(\boldsymbol{x})$ é, por hipótese, conhecida.

Caso contrário, utilizamos polinômios em $\boldsymbol{x}$ para aproximar a verdadeira função $f(\boldsymbol{x})$.

Exemplo 1

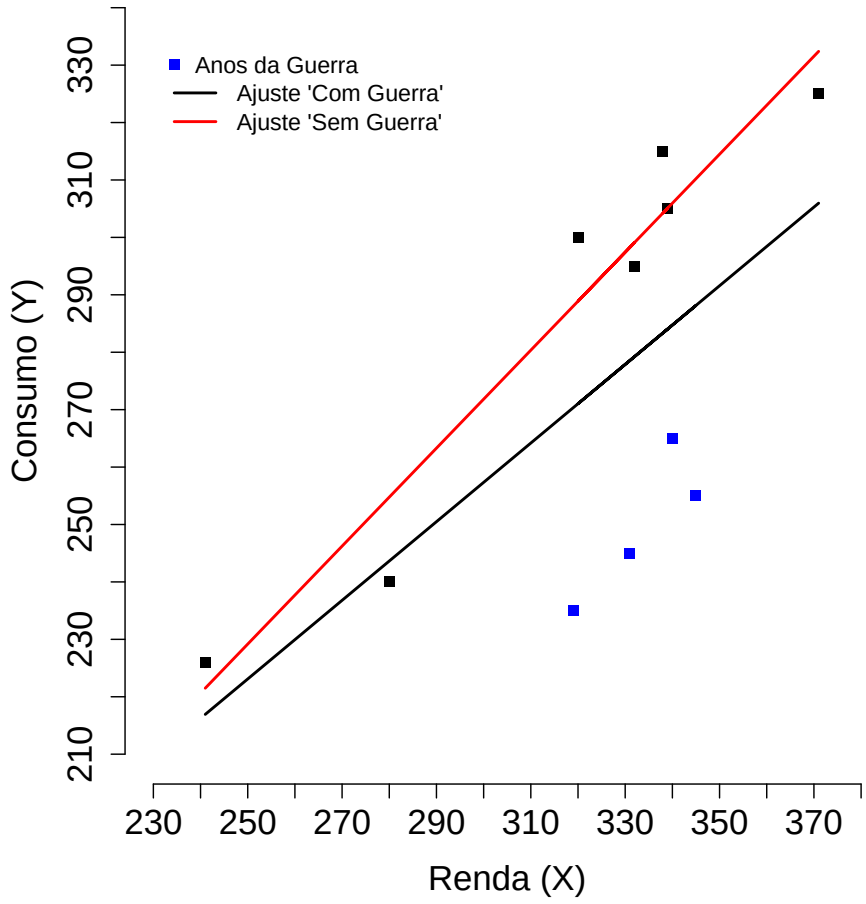


Figura 1: Relação entre a temperatura média em Fahrenheit (x) e libras de vapor por mês (y).

Ver arquivo exemplo_01.r

Regressão linear

O termo linear se refere aos parâmetros do modelo (os $\beta's$). Por exemplo,

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \varepsilon; \\ y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon; \text{ e} \\ y &= \beta_0 + \beta_1 \exp(x_1) + \beta_2 \log(x_2^2) + \varepsilon, \end{aligned}$$

são exemplos de modelos de regressão linear.

Em geral, tratamos ε e y como variáveis aleatórias enquanto $\boldsymbol{x}$ é dado, isto é, a análise é condicional ao conhecimento de $\boldsymbol{x}$ (regressoras).

Por hipótese, temos que **a média de ε é zero**.

Começaremos com o modelo de **regressão linear simples** dado por

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

O objetivo é obter a “melhor” equação da reta que descreve a relação entre x e y.

A ideia de mínimos quadrados

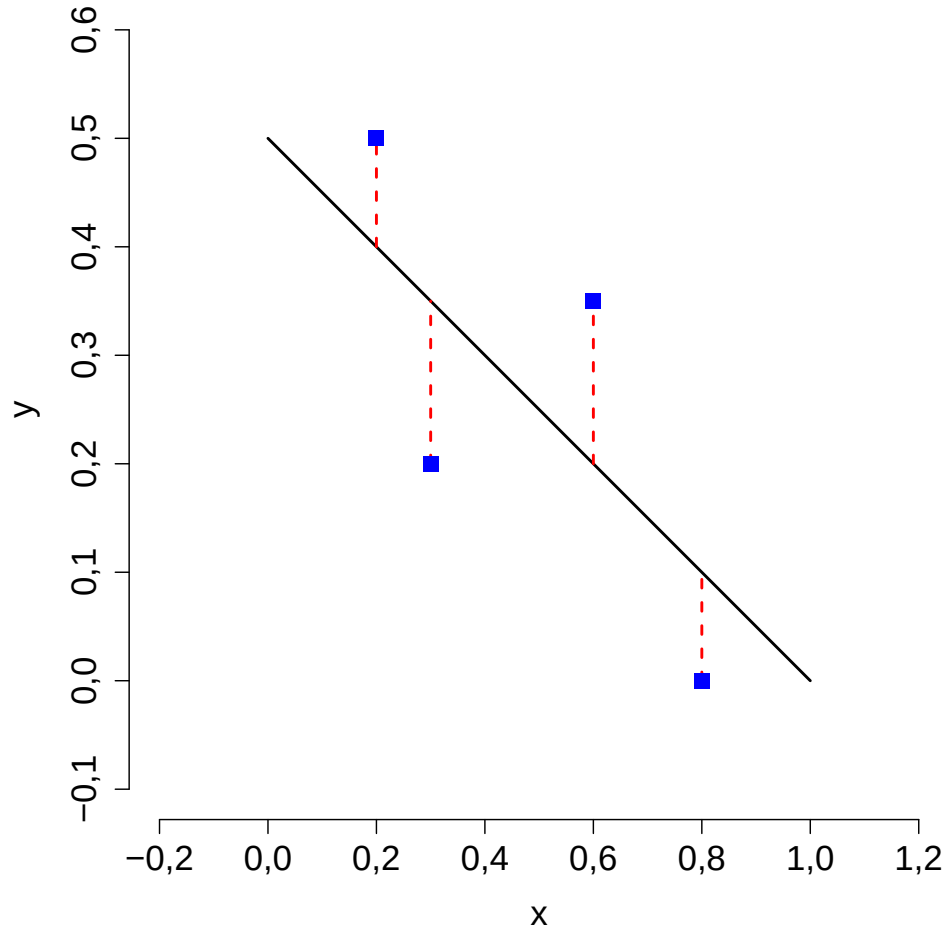


Figura 2: Ilustração da ideia de mínimos quadrados.

A **equação estimada** é denotada por $\widehat{y} = b_0 + b_1 x = \widehat{\beta}_0 + \widehat{\beta}_1 x$, em que $b_0 = \widehat{\beta}_0$ e $b_1 = \widehat{\beta}_1$ são **estimativas pontuais** de β_0 e β_1, respectivamente.

Mínimos quadrados

Suponha que uma amostra aleatória $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$ de tamanho n seja obtida tal que

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Definimos a função **soma de quadrados** como

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Queremos minimizar esta soma de quadrados. Logo,

$$\begin{aligned} \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) , \\ \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) . \end{aligned}$$

Agora, precisamos resolver o sistema de equações dado por

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) &= 0, \\ \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) &= 0. \end{aligned}$$

Assim, temos que

$$\begin{aligned} \sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_i &= 0, \\ \sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 &= 0, \end{aligned}$$

ou

$$\begin{aligned} nb_0 + b_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i, \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i. \end{aligned}$$

Ambas as representações são chamadas de **equações normais** (perpendicular ou ortogonal).

A solução do sistema acima é dado por

$$
\begin{aligned}
b_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} y_i\right]}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left[\sum_{i=1}^{n} x_i\right]^2}{n}} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}, \quad \text{e} \\
b_0 &= \overline{y} - b_1\overline{x},
\end{aligned}
$$

em que

$$
\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \text{e}
$$

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) &= \sum_{i=1}^{n} x_i y_i - \overline{x}\sum_{i=1}^{n} y_i - \overline{y}\sum_{i=1}^{n} x_i + n\overline{x}\,\overline{y} \\
&= \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} \\
&= \sum_{i=1}^{n} x_i y_i - \frac{\left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} y_i\right]}{n}.
\end{aligned}
$$

Utilizaremos as seguintes definições:

$$
\begin{aligned}
S_{xy} &= \sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \sum_{i=1}^{n}(x_i-\overline{x})y_i = \sum_{i=1}^{n}x_i(y_i-\overline{y}) \\
&= \sum_{i=1}^{n}x_iy_i - \frac{\left[\sum_{i=1}^{n}x_i\right]\left[\sum_{i=1}^{n}y_i\right]}{n} = \sum_{i=1}^{n}x_iy_i - n\overline{x}\,\overline{y}, \\
S_{xx} &= \sum_{i=1}^{n}(x_i-\overline{x})^2 = \sum_{i=1}^{n}(x_i-\overline{x})x_i \\
&= \sum_{i=1}^{n}x_i^2 - \frac{\left[\sum_{i=1}^{n}x_i\right]^2}{n} = \sum_{i=1}^{n}x_i^2 - n\overline{x}^2, \quad \text{e} \\
S_{yy} &= \sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum_{i=1}^{n}(y_i-\overline{y})y_i \\
&= \sum_{i=1}^{n}y_i^2 - \frac{\left[\sum_{i=1}^{n}y_i\right]^2}{n} = \sum_{i=1}^{n}y_i^2 - n\overline{y}^2.
\end{aligned}
$$

Assim, temos que $b_1 = \dfrac{S_{xy}}{S_{xx}}$ e $b_0 = \overline{y} - b_1\overline{x}$.

Então, agora é possível escrever a **equação ajustada** ou **predita** como

$$\widehat{y} = b_0 + b_1 x.$$

Agora, substituindo $b_0 = \overline{y} - b_1 \overline{x}$ na equação acima temos que

$$\widehat{y} = \overline{y} + b_1 (x - \overline{x}).$$

Observe que $x = \overline{x} \Rightarrow \widehat{y} = \overline{y}$.

Os dados de vapor

i	y_i	x_i
1	10,98	35,30
2	11,13	29,70
3	12,51	30,80
4	8,40	58,80
5	9,27	61,40
6	8,73	71,30
7	6,36	74,40
8	8,50	76,70
9	7,82	70,70
10	9,14	57,50
11	8,24	46,40
12	12,19	28,90
13	11,88	28,10
14	9,57	39,10
15	10,94	46,80
16	9,58	48,50
17	10,09	59,30
18	8,11	70,00
19	6,83	70,00
20	8,88	74,50
21	7,68	72,10
22	8,47	58,10
23	8,86	44,60
24	10,36	33,40
25	11,08	28,60

Cálculos para os dados de vapor

$$
\begin{aligned}
n &= 25 \\
\sum_{i=1}^{25} y_i &= 10,98 + 11,13 + \cdots + 11,08 = 235,60 \\
\overline{y} &= \frac{235,60}{25} = 9,424 \\
\sum_{i=1}^{25} x_i &= 35,3 + 29,7 + \cdots + 28,6 = 1.315 \\
\overline{x} &= \frac{1.315}{25} = 52,60 \\
\sum_{i=1}^{25} x_i y_i &= (10,98)(35,3) + (11,13)(29,7) + \cdots + (11,08)(28,6) \\
&= 11.821,432 \\
\sum_{i=1}^{25} x_i^2 &= (35,3)^2 + (29,7)^2 + \cdots + (28,6)^2 = 76.323,42 \\
b_1 &= \frac{11.821,432 - \dfrac{(1315)(235,6)}{25}}{76.323,42 - \dfrac{(1315)^2}{25}} = \frac{-571,128}{7.154,42} = -0,079829
\end{aligned}
$$

A equação ajustada é dada por

$$\begin{aligned}\widehat{y} &= \overline{y} + b_1(x - \overline{x}) = 9,4240 - 0,079829(x - 52,60) \\ &= 13,623 - 0,079829x.\end{aligned}$$

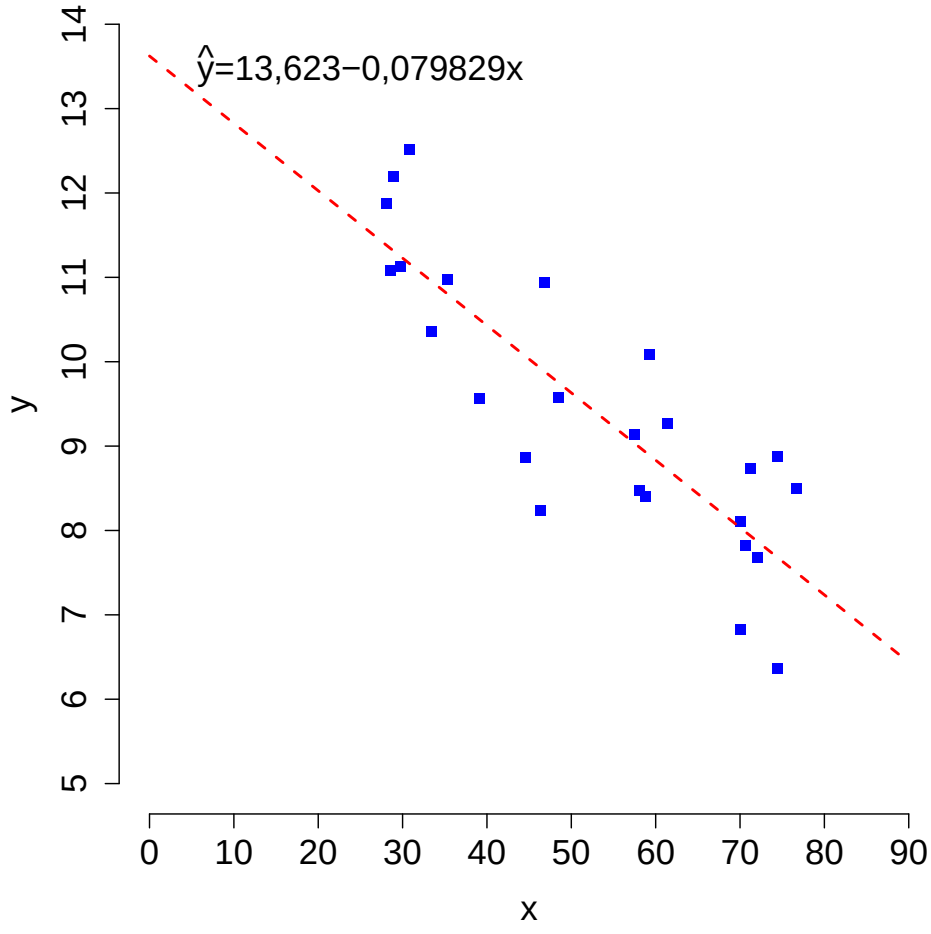


Figura 3: Equação da reta ajustada.

Resíduos

Definimos o **resíduo**, para cada observação, como diferença entre o valor observado y_i e o valor ajustado $\widehat{y}_i$, isto é,

$$e_i = y_i - \widehat{y}_i \quad \text{ou} \quad \widehat{\varepsilon}_i = y_i - \widehat{y}_i.$$

Importante: note que a amostra $(\widehat{\varepsilon}_1, \widehat{\varepsilon}_2, \ldots, \widehat{\varepsilon}_n)$ deve ter comportamento **similar** a uma amostra da distribuição de ε.

Agora, note que $y_i - \widehat{y}_i = (y_i - \overline{y}) - b_1(x_i - \overline{x})$.

Portanto,

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i = \sum_{i=1}^{n} (y_i - \widehat{y}_i) = \sum_{i=1}^{n} (y_i - \overline{y}) - b_1 \sum_{i=1}^{n} (x_i - \overline{x}) = 0.$$

Observações, valores ajustados e resíduos

i	y_i	$\widehat{y}_i$	$y_i - \widehat{y}_i$
1	10,98	10,81	0,17
2	11,13	11,25	-0,12
3	12,51	11,16	1,35
4	8,40	8,93	-0,53
5	9,27	8,72	0,55
6	8,73	7,93	0,80
7	6,36	7,68	-1,32
8	8,50	7,50	1,00
9	7,82	7,98	-0,16
10	9,14	9,03	0,11
11	8,24	9,92	-1,68
12	12,19	11,32	0,87
13	11,88	11,38	0,50
14	9,57	10,50	-0,93
15	10,94	9,89	1,05
16	9,58	9,75	-0,17
17	10,09	8,89	1,20
18	8,11	8,03	0,08
19	6,83	8,03	-1,20
20	8,88	7,68	1,20
21	7,68	7,87	-0,19
22	8,47	8,98	-0,51
23	8,86	10,06	-1,20
24	10,36	10,96	-0,60
25	11,08	11,34	-0,26

Regressão sem a constante (sem intercepto)

Suponha que $\beta_0 = 0$, isto é, a reta passa por $(x, y) = (0, 0)$.

Então, a equação a ser estimada é dada por $y_i = \beta_1 x_i + \varepsilon_i$.

Derivando-se $S(\beta_1)$ em relação a β_1 e igualando-se a zero, temos que

$$\sum_{i=1}^{n} x_i (y_i - \beta_1 x_i) = 0 \quad \Rightarrow \quad b_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

A equação da reta estimada é dada por $\widehat{y} = b_1 x$.

No ponto $x = \overline{x}$ temos que $\widehat{y} = b_1 \overline{x}$, isto é, não resulta em $\overline{y}$.

Além disso, em geral,

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i = \sum_{i=1}^{n} (y_i - \widehat{y}_i) = \sum_{i=1}^{n} (y_i - b_1 x_i) \neq 0.$$

Se $(x, y) = (0, 0)$ for verdade, então $b_0 = 0$. Consequentemente $b_1 = \dfrac{\overline{y}}{\overline{x}}$ que resulta em $\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$.

Centrando os dados

Temos que $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Agora, suponha que $y_i - \overline{y} = (\beta_0 + \beta_1 \overline{x} - \overline{y}) + \beta_1 (x_i - \overline{x}) + \varepsilon_i$.

Reescrevendo, obtemos que $y_i^\star = \beta_0^\star + \beta_1 x_i^\star + \varepsilon_i$, em que

$$\begin{aligned} y_i^\star &= y_i - \overline{y} \\ \beta_0^\star &= \beta_0 + \beta_1 \overline{x} - \overline{y} \\ x_i^\star &= x_i - \overline{x}. \end{aligned}$$

Note que $b_1 = \dfrac{\sum_{i=1}^n x_i^\star y_i^\star}{\sum_{i=1}^n [x_i^\star]^2}$ e $b_0^\star = \overline{y^\star} - b_1 \overline{x^\star} = 0$, pois $\overline{x^\star} = \overline{y^\star} = 0$.

Como isto sempre acontece ($b_0^\star = 0$), o modelo a ser ajustado é dado por

$$y_i - \overline{y} = b_1 (x_i - \overline{x})$$

Perdemos um parâmetro (β_0). Contudo, as quantidades $(y_i - \overline{y})$, para $i = 1, 2, \ldots, n$, representam somente $(n - 1)$ pedaços de informações.

Análise de variância

Temos que avaliar a variação nos dados explicada pela reta de regressão.

Considere que $y_i - \widehat{y}_i = (y_i - \overline{y}) - (\widehat{y}_i - \overline{y})$.

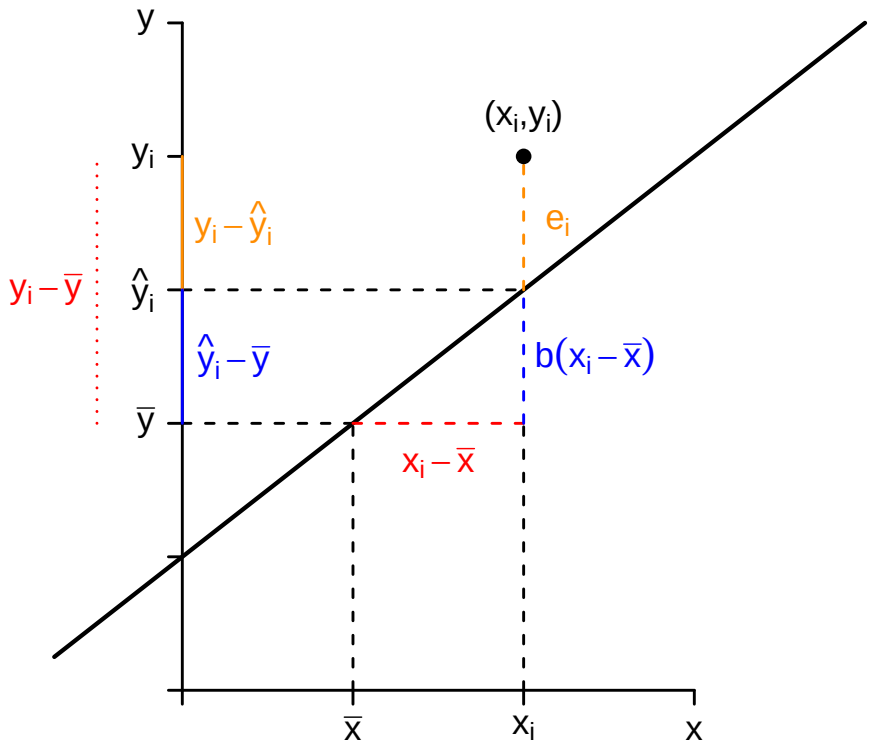


Figura 4: Decomposição de y.

Note que

$$\overline{\widehat{y}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(b_0 + b_1 x_i) = \frac{1}{n}(nb_0 + b_1 n\overline{x}) = b_0 + b_1\overline{x} = \overline{y}.$$

Isto implica novamente que

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}\widehat{\varepsilon}_i = \sum_{i=1}^{n}(y_i - \widehat{y}_i) = n\overline{y} - n\overline{y} = 0.$$

Podemos reescrever a decomposição como

$$(y_i - \overline{y}) = (\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i).$$

Portanto, a soma de quadrados resulta em

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2,$$

pois a soma do produto cruzado (SPC) é igual a zero, isto é,

$$\text{SPC} = 2\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})(y_i - \widehat{y}_i) = 0.$$

Agora, note que **(ver Figura 4)**

$$\widehat{y}_i - \overline{y} = b_1(x_i - \overline{x}) \quad \text{e} \quad y_i - \widehat{y}_i = y_i - \overline{y} - b_1(x_i - \overline{x}).$$

Assim,

$$\text{SPC} = 2\sum_{i=1}^{n} b_1(x_i - \overline{x})[(y_i - \overline{y}) - b_1(x_i - \overline{x})] = 2b_1(S_{xy} - b_1 S_{xx}) = 0,$$

pois

$$b_1 = \frac{S_{xy}}{S_{xx}}.$$

Temos também que

$$\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = \sum_{i=1}^{n} b_1^2 (x_i - \overline{x})^2 = b_1^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}.$$

A soma de quadrados é dada por

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{SQTot}=\underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}_{SQReg}+\underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{SQRes},$$

em que

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

Tabela 1: Análise de variância.

Soma de Quadrados	Fonte	Graus de Liberdade	Média Quadrática
Regressão	$\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2$	1	$MQReg$
Resíduo	$\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2$	$n-2$	s^2
Total	$\sum_{i=1}^{n}(y_i-\overline{y})^2$	$n-1$	s_y^2

Por definição, temos que $MQReg=\dfrac{SQReg}{1}$, $s^2=\dfrac{SQRes}{n-2}$ e $s_y^2=\dfrac{S_{yy}}{(n-1)}$.

Coeficiente de determinação

$$R^2 = \frac{SQReg}{SQTot} = 1 - \frac{SQRes}{SQTot} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

O R^2 mede a proporção da variação total que é explicada pela regressão.

Exemplo 1 (continuação)

Tabela 2: Análise de variância - dados de vapor.

Soma de Quadrados	Fonte	Graus de Liberdade	Média Quadrática	R^2
Regressão	45,59	1	45,59	0,714
Resíduo	18,22	23	0,792	—
Total	63,82	24	2,659	—

Intervalos de confiança e testes de hipóteses para β_0 e β_1

Seja o modelo $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, para $i = 1, 2, \ldots, n$.

Agora, consideraremos as seguintes hipóteses:

HP.1: ε_i é uma variável aleatória com média 0 (zero) e variância constante e desconhecida σ^2_ε. Temos que $\mathrm{E}(\varepsilon_i) = 0$ e $\mathrm{Var}(\varepsilon_i) = \sigma^2_\varepsilon$.

HP.2: ε_i e ε_j são não correlacionados para todo $i \neq j$ tal que $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.

HP.3: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_\varepsilon)$, isto é, ε_i tem distribuição normal.

Consequentemente, temos que

$$\begin{aligned}
\mathrm{E}(y_i|x_i) &= \beta_0 + \beta_1 x_i; \\
\mathrm{Var}(y_i|x_i) &= \sigma^2_\varepsilon; \\
\mathrm{Cov}(y_i, y_j|x_i, x_j) &= 0, \quad \text{para todo} \quad i \neq j; \ \text{e} \\
(y_i|x_i) &\sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2_\varepsilon).
\end{aligned}$$

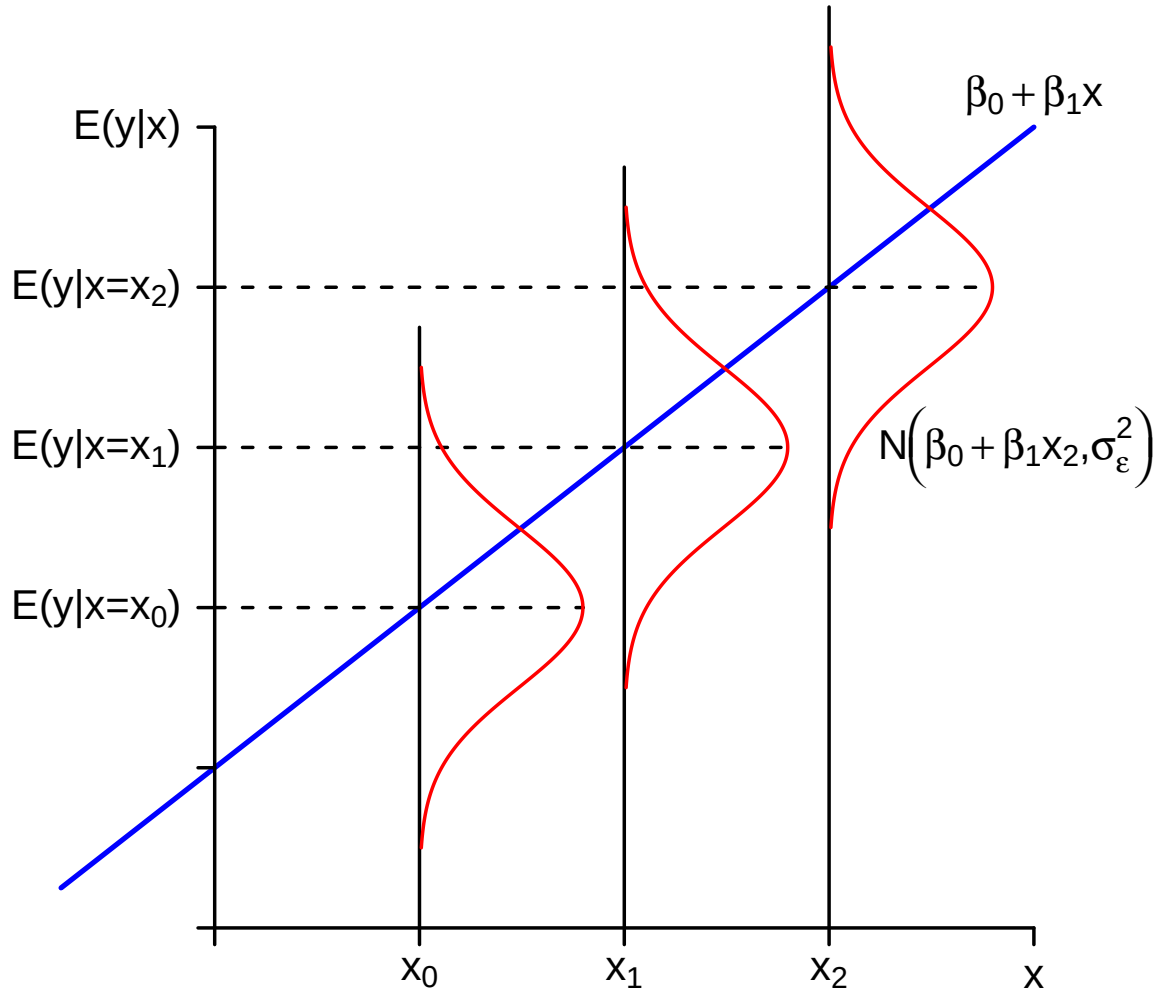


Figura 5: O modelo clássico de regressão com erros normais.

Propriedades de b_1

Temos que

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sum_{i=1}^{n}\omega_i y_i,$$

em que

$$\omega_i = \frac{(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Assim,

$$\begin{aligned}
\mathrm{E}(b_1|\boldsymbol{x}) &= \mathrm{E}\left(\sum_{i=1}^{n}\omega_i y_i \middle| \boldsymbol{x}\right) = \sum_{i=1}^{n}\omega_i \mathrm{E}(y_i|x_i) \\
&= \sum_{i=1}^{n}\omega_i(\beta_0 + \beta_1 x_i) = \frac{\beta_0\sum_{i=1}^{n}(x_i - \overline{x}) + \beta_1\sum_{i=1}^{n}(x_i - \overline{x})x_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\
&= \beta_0\frac{\sum_{i=1}^{n}(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + \beta_1\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \beta_1.
\end{aligned}$$

$$\begin{aligned}\text{Var}(b_1|\boldsymbol{x}) &= \text{Var}\left(\sum_{i=1}^{n}\omega_i y_i\middle|\boldsymbol{x}\right) = \sum_{i=1}^{n}\omega_i^2\text{Var}(y_i|x_i) \\ &= \sum_{i=1}^{n}\omega_i^2\sigma_\varepsilon^2 = \sigma_\varepsilon^2\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^2} = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\sigma_\varepsilon^2}{S_{xx}}.\end{aligned}$$

Portanto,

$$\text{DP}(b_1|\boldsymbol{x}) = \frac{\sigma_\varepsilon}{\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^{1/2}} = \frac{\sigma_\varepsilon}{S_{xx}^{1/2}}.$$

Assumindo que o modelo é o correto, substituímos σ_ε por s e obtemos que

$$\text{ep}(b_1|\boldsymbol{x}) = \widehat{\text{DP}(b_1|\boldsymbol{x})} = \frac{s}{\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^{1/2}} = \frac{s}{S_{xx}^{1/2}}.$$

Se considerarmos **HP.3**, e como $b_1 = \sum_{i=1}^{n}\omega_i y_i$ é uma combinação linear de normais, temos que

$$(b_1|\boldsymbol{x}) \sim \mathcal{N}\left(\beta_1, \frac{\sigma_\varepsilon^2}{S_{xx}}\right).$$

Exemplo 2: distribuição amostral do estimador de mínimos quadrados

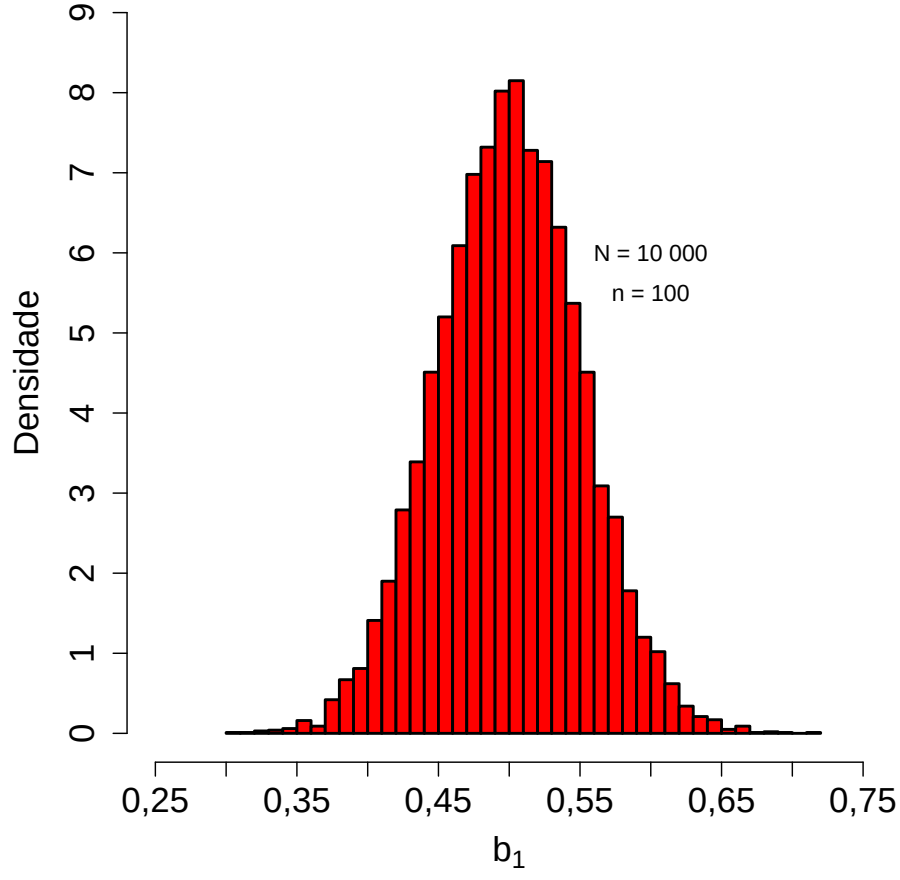
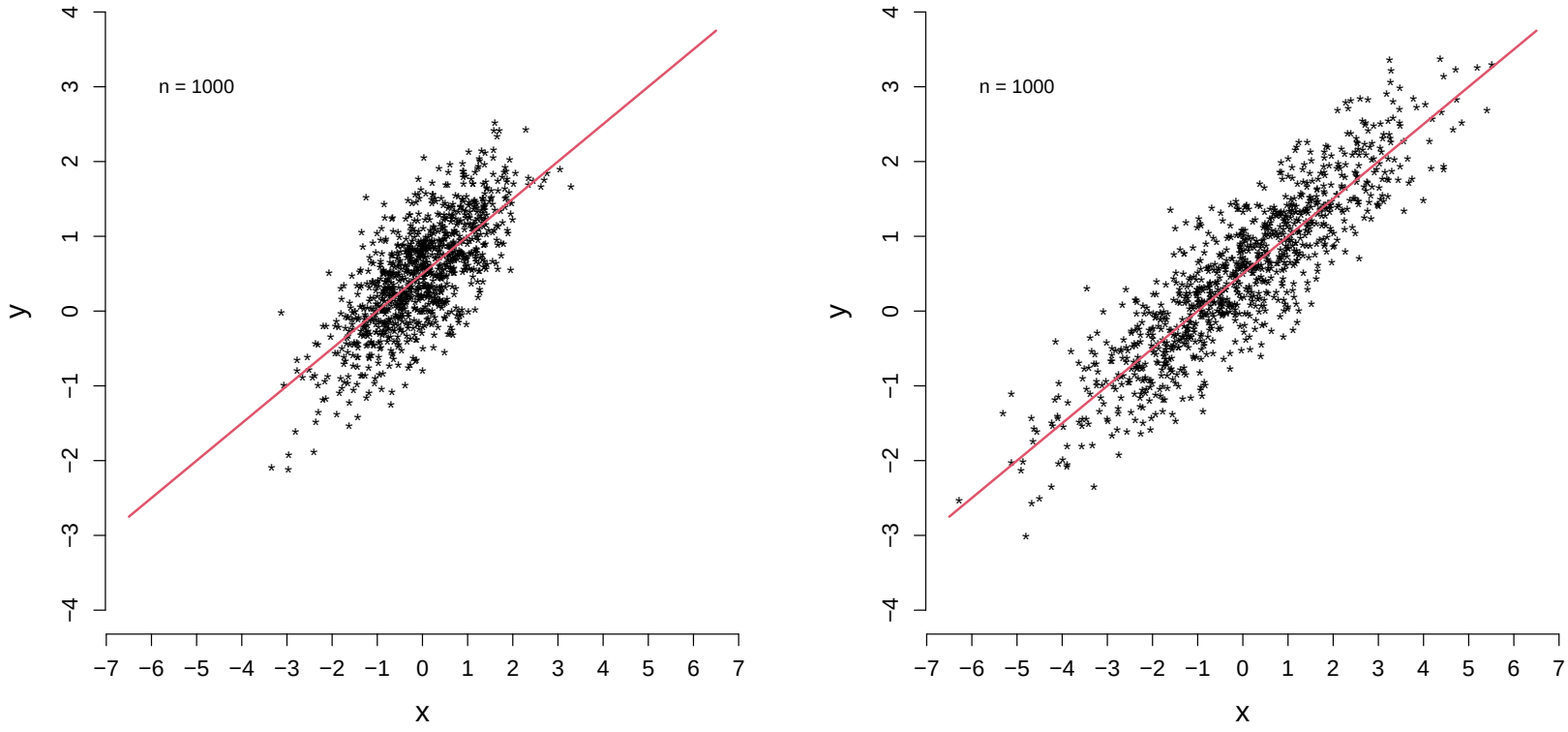


Figura 6: Distribuição amostral do estimador b_1 de β_1 na regressão $y = \beta_0 + \beta_1 x + \varepsilon$.

Ver arquivo exemplo_02.r

Exemplo 3: variância do estimador de mínimos quadrados



(a) Menor variância em x.

(b) Maior variância em x.

Figura 7: Variância do estimador b_1 de β_1 na regressão $y = \beta_0 + \beta_1 x + \varepsilon$ com $S_{xx}(1) = 1.015,76$ e $S_{xx}(2) = 3.657,90$.

Ver arquivo exemplo_03.r

Intervalo de confiança e teste de hipóteses para β_1

Se considerarmos **HP.3**, então $(b_1|\boldsymbol{x}) \sim \mathcal{N}\left(\beta_1, \frac{\sigma_\varepsilon^2}{S_{xx}}\right)$.

Pode ser mostrado que

$$\text{IC}_{100(1-\alpha)\%}(\beta_1) = \left(b_1 - t_{n-2,1-\alpha/2}\frac{s}{\sqrt{S_{xx}}}, b_1 + t_{n-2,1-\alpha/2}\frac{s}{\sqrt{S_{xx}}}\right).$$

Além disso, para o teste estatístico de hipóteses

$$\begin{cases} H_0: & \beta_1 = \beta_{1,0} \\ H_1: & \beta_1 \neq \beta_{1,0} \end{cases}$$

temos que a estatística do teste é dada por

$$t = \frac{(b_1 - \beta_{1,0})}{\text{ep}(b_1)} = \frac{(b_1 - \beta_{1,0})\sqrt{S_{xx}}}{s}.$$

Se $|t| > t_{n-2,1-\alpha/2}$, então rejeita-se a hipótese nula.

Exemplo 1 (continuação)

O erro padrão de b_1 é dado por $\text{ep}(b_1) = \dfrac{s}{\sqrt{S_{xx}}} = \dfrac{0,8901}{\sqrt{84,58}} = 0,01052.$

Se considerarmos $\alpha = 0,05$, temos que $t_{23;\,0,975} = 2,069$.

Assim,

$$\begin{aligned} \text{IC}_{95\%}(\beta_1) &= (-0,0798 - 2,069 \times 0,01052;\ -0,0798 + 2,069 \times 0,01052) \\ &= (-0,10160;\ -0,05806). \end{aligned}$$

Agora, para o teste

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0, \end{cases}$$

temos que $t = (b_1 - 0)/\text{ep}(b_1) = -0,0798/0,01052 = -7,60$.

Como $|t| = 7,60 > t_{23;\,0,975} = 2,069$, então rejeita-se a H_0, isto é, β_1 é estatisticamente diferente de zero.

Concluímos que a variável x realmente ajuda explicar a variável y através da regressão linear $y = \beta_0 + \beta_1 x + \varepsilon$.

Ver arquivo exemplo_01.r

Propriedades de b_0

Temos que $b_0 = \overline{y} - b_1\overline{x}$. Assim,

$$
\begin{aligned}
\mathrm{E}(b_0|\boldsymbol{x}) &= \mathrm{E}(\overline{y} - b_1\overline{x}|\boldsymbol{x}) = \mathrm{E}(\overline{y}|\boldsymbol{x}) - \mathrm{E}(b_1|\boldsymbol{x})\overline{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(y_i|x_i) - \beta_1\overline{x} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \beta_1\overline{x} = \beta_0 \\
\mathrm{Var}(b_0|\boldsymbol{x}) &= \mathrm{Var}(\overline{y} - b_1\overline{x}|\boldsymbol{x}) \\
&= \mathrm{Var}(\overline{y}|\boldsymbol{x}) + [\overline{x}]^2\mathrm{Var}(b_1|\boldsymbol{x}) - 2\mathrm{Cov}(\overline{y}, b_1|\boldsymbol{x})\overline{x} \\
&= \frac{\sigma_\varepsilon^2}{n} + [\overline{x}]^2\frac{\sigma_\varepsilon^2}{S_{xx}} = \sigma_\varepsilon^2\left[\frac{1}{n} + \frac{[\overline{x}]^2}{S_{xx}}\right] \\
&= \sigma_\varepsilon^2\left[\frac{S_{xx} + n[\overline{x}]^2}{nS_{xx}}\right] = \sigma_\varepsilon^2\left[\frac{\sum_{i=1}^{n}x_i^2}{nS_{xx}}\right], \qquad \text{pois} \\
\mathrm{Cov}(\overline{y}, b_1|\boldsymbol{x}) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n}y_i, \sum_{i=1}^{n}\omega_i y_i\middle|\boldsymbol{x}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}(y_i, \omega_i y_i|\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\omega_i\mathrm{Var}(y_i|\boldsymbol{x}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\omega_i\sigma_\varepsilon^2 = \frac{\sigma_\varepsilon^2}{n}\sum_{i=1}^{n}\frac{(x_i - \overline{x})}{S_{xx}} = 0.
\end{aligned}
$$

Portanto,

$$\mathrm{DP}(b_0|\boldsymbol{x}) = \sigma_\varepsilon \left[\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}} \right]^{1/2}.$$

Assumindo que o modelo é o correto, substituímos σ_ε por s e obtemos que

$$\mathrm{ep}(b_0|\boldsymbol{x}) = \widehat{\mathrm{DP}(b_0|\boldsymbol{x})} = s \left[\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}} \right]^{1/2}.$$

Se considerarmos **HP.3**, e como $b_0 = \overline{y} - b_1\overline{x}$ é uma combinação linear de normais, temos que

$$(b_0|\boldsymbol{x}) \sim \mathcal{N}\left(\beta_0, \sigma_\varepsilon^2 \left[\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}} \right] \right).$$

Intervalo de confiança e teste de hipóteses para β_0

Pode ser mostrado que

$$\text{IC}_{100(1-\alpha)\%}(\beta_0) = \left(b_0 - t_{n-2,1-\alpha/2} s \left[\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}} \right]^{1/2}, b_0 + t_{n-2,1-\alpha/2} s \left[\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}} \right]^{1/2} \right).$$

Além disso, para o teste estatístico de hipóteses

$$\begin{cases} H_0: & \beta_0 = \beta_{0,0} \\ H_1: & \beta_0 \neq \beta_{0,0} \end{cases}$$

temos que a estatística do teste é dada por

$$t = \frac{(b_0 - \beta_{0,0})}{\text{ep}(b_0)} = \frac{(b_0 - \beta_{0,0})\sqrt{nS_{xx}}}{s\sqrt{\sum_{i=1}^{n} x_i^2}}.$$

Se $|t| > t_{n-2,1-\alpha/2}$, então rejeita-se a hipótese nula.

Tabela 3: 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	$< 10^{-15}$

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

Tabela 4: 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	$< 10^{-15}$

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

Propriedades das somas de quadrados

Suponha que $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, 2, \ldots, n$, seja o modelo correto. Assim,

$$
\begin{aligned}
\mathrm{E}(\overline{y}|\boldsymbol{x}) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i \middle| \boldsymbol{x}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(y_i|\boldsymbol{x}) \\
&= \frac{1}{n}\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \overline{x} \\
\mathrm{E}(SQTot|\boldsymbol{x}) &= \mathrm{E}\left(\sum_{i=1}^{n} (y_i - \overline{y})^2 \middle| \boldsymbol{x}\right) = \mathrm{E}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2 \middle| \boldsymbol{x}\right) \\
&= \sum_{i=1}^{n} \mathrm{E}(y_i^2|x_i) - n\mathrm{E}(\overline{y}^2|\boldsymbol{x}) \\
&= \sum_{i=1}^{n} \left[\mathrm{Var}(y_i|x_i) + \mathrm{E}^2(y_i|x_i)\right] - n\left[\mathrm{Var}(\overline{y}|\boldsymbol{x}) + \mathrm{E}^2(\overline{y}|\boldsymbol{x})\right] \\
&= \sum_{i=1}^{n} \left[\sigma_\varepsilon^2 + (\beta_0 + \beta_1 x_i)^2\right] - n\left[\frac{\sigma_\varepsilon^2}{n} + (\beta_0 + \beta_1 \overline{x})^2\right] \\
&= (n-1)\sigma_\varepsilon^2 + \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i)^2 - n(\beta_0 + \beta_1 \overline{x})^2 \\
&= (n-1)\sigma_\varepsilon^2 + \beta_1^2 S_{xx}
\end{aligned}
$$

Continuando, temos que

$$\begin{aligned}
\mathrm{E}(SQReg|\boldsymbol{x}) &= \mathrm{E}\left(\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2\Big|\boldsymbol{x}\right)=\sum_{i=1}^{n}\mathrm{E}((\widehat{y}_i-\overline{y})^2|\boldsymbol{x})\\
&= \sum_{i=1}^{n}\mathrm{E}(b_1^2(x_i-\overline{x})^2|\boldsymbol{x})=\mathrm{E}(b_1^2|\boldsymbol{x})S_{xx}\\
&= \left[\mathrm{Var}(b_1|\boldsymbol{x})+\mathrm{E}^2(b_1|\boldsymbol{x})\right]S_{xx}=\left[\frac{\sigma_\varepsilon^2}{S_{xx}}+\beta_1^2\right]S_{xx}\\
&= \sigma_\varepsilon^2+\beta_1^2S_{xx}.\\
\mathrm{E}(SQRes|\boldsymbol{x}) &= \mathrm{E}(SQTot-SQReg|\boldsymbol{x})=\mathrm{E}(SQTot|\boldsymbol{x})-\mathrm{E}(SQReg|\boldsymbol{x})\\
&= (n-1)\sigma_\varepsilon^2+\beta_1^2S_{xx}-\sigma_\varepsilon^2-\beta_1^2S_{xx}=(n-2)\sigma_\varepsilon^2.
\end{aligned}$$

Consequentemente,

$$\mathrm{E}(MQReg|\boldsymbol{x})=\sigma_\varepsilon^2+\beta_1^2S_{xx}\quad\text{e}\quad\mathrm{E}(s^2|\boldsymbol{x})=\sigma_\varepsilon^2.$$

Se considerarmos **HP.3** e que $\beta_1=0$, então é possível mostrar *MQReg* e s^2 são independentes, e que

$$\frac{MQReg}{\sigma_\varepsilon^2}\sim\chi_1^2\quad\text{e}\quad\frac{(n-2)s^2}{\sigma_\varepsilon^2}\sim\chi_{n-2}^2.$$

Além disso, $\dfrac{MQReg}{s^2}\sim\mathcal{F}_{1,n-2}$.

Teste F para significância da regressão

Para o teste estatístico de hipóteses

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}$$

temos que a estatística do teste (alternativa) é dada por

$$F = \frac{MQReg}{s^2}.$$

Se $F > F_{1,n-2,1-\alpha}$, então rejeita-se a hipótese nula.

Note que os testes T e F são equivalentes (quando temos somente uma variável regressora), pois neste caso $F = T^2 \sim \mathcal{F}_{1,n-2}$.

Exemplo 1 (continuação)

Queremos testar via estatística

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}$$

Temos que

$$F = \frac{MQReg}{s^2} = \frac{45,59}{0,792} = 57,54$$

e $F_{1;\,23;\,0,95} = 4,279$. Como $F > F_{1;\,23;\,0,95}$, então rejeitamos H_0.

Note que $t = -7,586$ e $F = t^2 = 57,54$. Então, para o modelo com uma regressora, os testes t e F geram os mesmos resultados e conclusões.

Tabela 5: Análise de variância - dados de vapor.

Soma de Quadrados	Fonte	Graus de Liberdade	Média Quadrática	F	R^2
Regressão	45,59	1	45,59	57,54	0,714
Resíduo	18,22	23	0,792	—	—
Total	63,82	24	2,659	—	—

Ver arquivo exemplo_01.r

Correlação de Pearson, b_1 e R^2

O coeficiente de correlação de Pearson entre duas variáveis aleatórias é definido por

$$\rho(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(y)\text{Var}(x)}}$$

e mede o grau de relação linear entre estas variáveis.

Dado uma amostra aleatória $((x_1, y_1), \ldots, (x_n, y_n))$, temos a correlação amostral dada por

$$r(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2\right]^{1/2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

Assim, temos que

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}S_{yy}^{1/2}}{S_{xx}^{1/2}S_{xx}^{1/2}S_{yy}^{1/2}} = r(x, y)\frac{S_{yy}^{1/2}}{S_{xx}^{1/2}}.$$

Então, b_1 e $r(x, y)$ têm o mesmo sinal.

Dado que $(\widehat{y}_i - \overline{y}) = b_1(x_i - \overline{x})$, note que

$$
\begin{aligned}
R^2 &= \frac{SQReg}{SQTot} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n} b_1^2(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= \frac{b_1^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2 S_{xx}}{S_{xx}^2 S_{yy}} = \left[\frac{S_{xy}}{S_{xx}^{1/2} S_{yy}^{1/2}}\right]^2 = [r(x,y)]^2; \\
[r(x,y)]^2 &= \frac{\left[\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right]^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= \frac{\left[\sum_{i=1}^{n} b_1(x_i - \overline{x})(y_i - \overline{y})\right]^2}{\sum_{i=1}^{n} b_1^2(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= \frac{\left[\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})(y_i - \overline{y})\right]^2}{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2} = [r(y,\widehat{y})]^2; \quad \text{e} \\
R^2 &= [r(y,\widehat{y})]^2.
\end{aligned}
$$

(É possível testar $H_0 : \rho(x,y) = 0$ contra $H_1 : \rho(x,y) \neq 0$.)

O coeficiente de determinação R^2

Sejam $\mathcal{M}_1$ e $\mathcal{M}_2$ os seguintes modelos de regreessão:

$$\begin{aligned} \mathcal{M}_1 : \quad & y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \\ \mathcal{M}_2 : \quad & y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \end{aligned}$$

com coeficientes de determinação R_1^2 e R_2^2, respectivamente.

Então, temos que $R_2^2 \geqslant R_1^2$.

O resultado diz que adicionar outra covariável ao modelo tende a melhorar a regressão (aumentar a *SQReg*) e diminuir o resíduo (*SQRes*).

Análise dos resíduos

Podemos realizar inspeções gráficas ou testes de hipóteses.

- Verificar a normalidade;
- Verificar o efeito do tempo se os dados forem ordenados no tempo;
- Verificar a homocedasticidade (variância constante);
- Verificar possíveis transformações dos dados;
- Verificar uma ordem polinomial maior do que a ajustada com os $\boldsymbol{x}$'s;
- Verificar pontos aberrantes ou de alavancagem; e
- Verificar as hipóteses atreladas ao tipo de dado.

Note que

- $\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$;
- $\text{Cov}(x, \widehat{\varepsilon}) = 0$ (por hipótese);
- Se $t = 1, 2, \ldots, n$, então $\text{Cov}(t, \widehat{\varepsilon}) = 0$ (por hipótese);
- $\text{Cov}(\widehat{y}, \widehat{\varepsilon}) = 0$, mas $\text{Cov}(y, \widehat{\varepsilon}) \neq 0$; e
- $\text{Var}(y|\boldsymbol{x}) = \sigma^2$.

Algumas definições (pausa em regressão)

Momentos ordinários:

$$m'_\ell = \mathrm{E}(y^\ell) = \int_{-\infty}^{\infty} y^\ell f(y) dy.$$

Definimos $\mu_y = m'_1 = \mathrm{E}(y)$ como a média.

Momentos centrais:

$$m_\ell = \mathrm{E}\left((y - \mu_y)^\ell\right) = \int_{-\infty}^{\infty} (y - \mu_y)^\ell f(y) dy.$$

Definimos $\sigma_y^2 = m_2 = \mathrm{E}\left((y - \mu_y)^2\right)$ como a variância.

O coeficiente de assimetria é definido por

$$A(y) = \mathrm{E}\left(\frac{[y-\mu_y]^3}{\sigma_y^3}\right)$$

e o coeficiente de curtose por

$$K(y) = \mathrm{E}\left(\frac{[y-\mu_y]^4}{\sigma_y^4}\right).$$

- A quantidade $K(y) - 3$ é chamada de excesso de curtose porque $K(y) = 3$ para a distribuição normal.
- Uma distribuição com excesso de curtose positivo é dita ter caudas pesadas (leptocúrtica).
- Uma distribuição com excesso de curtose negativo é dita ter caudas leves (platicúrtica).

Momentos amostrais

Suponha que $\{y_1, y_2, \ldots, y_n\}$ seja uma amostra aleatória de y com n observações.

- Média amostral: $\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- Variância amostral: $\hat{\sigma}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \hat{\mu}_y)^2$.
- O coeficiente de assimetria amostral: $\widehat{A}(y) = \frac{1}{(n-1)\hat{\sigma}_y^3}\sum_{i=1}^{n} (y_i - \hat{\mu}_y)^3$.
- O coeficiente de curtose amostral: $\widehat{K}(y) = \frac{1}{(n-1)\hat{\sigma}_y^4}\sum_{i=1}^{n} (y_i - \hat{\mu}_y)^4$.

- Sob a hipótese de normalidade (**hipótese de que os dados são provenientes de uma distribuição normal**), temos que

$$\widehat{A}(y) \approx \mathcal{N}(0, 6/n) \quad \text{e} \quad \widehat{K}(y) \approx \mathcal{N}(3, 24/n)$$

  para n "suficientemente grande".
- Estas aproximações para n grande podem ser utilizadas para testar a hipótese de normalidade dos dados.
- Dado uma amostra aleatória $\{y_1, y_2, \ldots, y_n\}$, para testar a assimetria dos retornos, consideramos

$$H_0 : A(y) = 0$$
$$H_1 : A(y) \neq 0.$$

- A estatística da razão t-Student da assimetria amostral é

$$t = \frac{\widehat{A}(y)}{\sqrt{6/n}} \approx \mathcal{N}(0, 1).$$

- Rejeitamos H_0 ao nível α de significância se $|t| > z_{1-\alpha/2}$, em que $z_{1-\alpha/2}$ o percentil $100(1 - \alpha/2)$ da distribuição normal padrão.

- Para testar o excesso de curstose dos retornos, consideramos

$$H_0 : K(y) - 3 = 0$$
$$H_1 : K(y) - 3 \neq 0.$$

- A estatística da razão t-Student da assimetria amostral é

$$t = \frac{\widehat{K}(y) - 3}{\sqrt{24/n}} \approx \mathcal{N}(0, 1).$$

- Rejeitamos H_0 ao nível α de significância se $|t| > z_{1-\alpha/2}$, em que $z_{1-\alpha/2}$ o percentil $100(1 - \alpha/2)$ da distribuição normal padrão.
- Temos também o teste de **Jarque e Bera** para normalidade com

$$JB = \frac{\widehat{A}^2(y)}{6/n} + \frac{(\widehat{K}(y) - 3)^2}{24/n} \approx \chi^2_2 \quad \text{para } n \text{ “grande”}.$$

- Rejeitamos H_0 (normalidade) se $JB > \chi^\star_{1-\alpha}$, em que $\chi^\star_{1-\alpha}$ o percentil $100(1 - \alpha)$ da distribuição χ^2_2.

Função de autocorrelação (pausa em regressão)

- Considere uma sequência de observações $(y_1, y_2, \ldots, y_n)$ equiespaçadas na escala do tempo.
- Estamos interessados na correlação linear entre y_t e y_{t-h} para algum inteiro h.
- O coeficiente de correlação entre y_t e y_{t-h} é chamado de autocorrelação de defasagem (*lag*) h de y_t.

$$\rho(h) = \frac{\text{Cov}(y_t, y_{t-h})}{\sqrt{\text{Var}(y_t)\text{Var}(y_{t-h})}} = \frac{\text{Cov}(y_t, y_{t-h})}{\text{Var}(y_t)} = \frac{\gamma(h)}{\gamma(0)},$$

  pois $\text{Var}(y_t) = \text{Var}(y_{t-h})$ **(hipótese de estacionariedade fraca)**.
- Temos que $\rho(0) = 1$, $\rho(h) = \rho(-h)$ e $-1 \leqslant \rho(h) \leqslant 1$ para todo h.
- Uma sequência de observações **(fracamente estacionária)** y_t não é correlacionada serialmente se, e somente se, $\rho(h) = 0$ para todo $h > 0$.

Estimação da função de autocorrelação

- Para uma dada amostra $\{y_t\}_{t=1}^n$, a autocorrelação de defasagem 1 de y_t é

$$\hat{\rho}(1) = \frac{\sum_{t=2}^{n}(y_t - \overline{y})(y_{t-1} - \overline{y})}{\sum_{t=1}^{n}(y_t - \overline{y})^2}.$$

- Sob algumas condições gerais, $\hat{\rho}(1)$ é um estimador consistente de $\rho(1)$.
- Se $\{y_t\}$ for uma sequência independente e identicamente distribuída (i.i.d.) e $\mathrm{E}(y_t^2) < \infty$, então $\hat{\rho}(1)$ é assintoticamente normal com média 0 e variância $1/n$.
- Para n suficientemente grande, temos

$$\hat{\rho}(1)\sqrt{n} \approx \mathcal{N}(0,1).$$

- Podemos testar $H_0 : \rho(1) = 0$ contra $H_1 : \rho(1) \neq 0$.

Estimação da função de autocorrelação: defasagem $h \geqslant 0$

- A autocorrelação de defasagem h de y_t é definida por

$$\hat{\rho}(h) = \frac{\sum_{t=h+1}^{n}(y_t - \overline{y})(y_{t-h} - \overline{y})}{\sum_{t=1}^{n}(y_t - \overline{y})^2}, \quad \text{para} \quad 0 \leqslant h \leqslant n-1.$$

- Se $\{y_t\}$ for uma sequência i.i.d. com $\mathrm{E}(y_t^2) < \infty$, então $\hat{\rho}(h)$ é assintoticamente normal com média 0 e variância $1/n$ para todo inteiro positivo e fixo h.
- Para n suficientemente grande, temos

$$\hat{\rho}(h)\sqrt{n} \approx \mathcal{N}(0,1).$$

- Podemos testar $H_0 : \rho(h) = 0$ contra $H_1 : \rho(h) \neq 0$ para h fixo.

Predição na média (de volta à regressão)

Dado uma regressão linear simples ajustada como

$$\widehat{y} = b_0 + b_1 x = \overline{y} + b_1(x - \overline{x}),$$

para prever pontualmente o valor $\mathrm{E}(y_0|\boldsymbol{x}, x_0) = \beta_0 + \beta_1 x_0$ no ponto x_0, utilizamos

$$\widehat{y}_0 = b_0 + b_1 x_0 = \overline{y} + b_1(x_0 - \overline{x}).$$

Além disso,

$$\begin{aligned}
\mathrm{Var}(\widehat{y}_0|\boldsymbol{x}, x_0) &= \mathrm{Var}(\overline{y}|\boldsymbol{x}) + (x_0 - \overline{x})^2 \mathrm{Var}(b_1|\boldsymbol{x}) \\
&= \frac{\sigma_\varepsilon^2}{n} + \frac{(x_0 - \overline{x})^2 \sigma_\varepsilon^2}{S_{xx}} = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]; \quad \text{e} \\
\mathrm{DP}(\widehat{y}_0|\boldsymbol{x}, x_0) &= \sigma_\varepsilon \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]^{1/2}.
\end{aligned}$$

Consequentemente, temos que

$$\mathrm{ep}(\widehat{y}_0|\boldsymbol{x}, x_0) = \widehat{\mathrm{DP}}(\widehat{y}_0|\boldsymbol{x}, x_0) = s \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]^{1/2}.$$

Temos também que

$$\begin{aligned} \mathrm{E}(\widehat{y}_0|\boldsymbol{x}, x_0) &= \mathrm{E}(\overline{y}|\boldsymbol{x}) + (x_0 - \overline{x})\mathrm{E}(b_1|\boldsymbol{x}) \\ &= \beta_0 + \beta_1\overline{x} + (x_0 - \overline{x})\beta_1 = \beta_0 + \beta_1 x_0, \end{aligned}$$

e, sob **HP.3**, $\widehat{y}_0 = \overline{y} + b_1(x_0 - \overline{x})$ é uma combinação linear de normais. Logo,

$$(\widehat{y}_0|\boldsymbol{x}, x_0) \sim \mathcal{N}\left(\beta_0 + \beta_1 x_0, \sigma^2_\varepsilon \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]\right)$$

é possível mostrar que s^2 e $\widehat{y}_0$ são independentes. Daí, temos que

$$\mathrm{IC}_{100(1-\alpha)\%}\left(\mathrm{E}(y_0|\boldsymbol{x}, x_0)\right) = (\widehat{y}_0 - t_{n-2;\,1-\alpha/2} \times \mathrm{ep}(\widehat{y}_0);\ \widehat{y}_0 + t_{n-2;\,1-\alpha/2} \times \mathrm{ep}(\widehat{y}_0)).$$

Ver arquivo exemplo_01.r

Predição para observações

Dado uma regressão linear simples $y = \beta_0 + \beta_1 x + \varepsilon$ e ajustada como

$$\widehat{y} = b_0 + b_1 x = \overline{y} + b_1(x - \overline{x}),$$

para prever pontualmente o valor $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$ no ponto x_0, utilizamos

$$\widehat{y}_0 = b_0 + b_1 x_0 = \overline{y} + b_1(x_0 - \overline{x}).$$

Assim, o erro de previsão é dado por $\widehat{\varepsilon}_0 = y_0 - \widehat{y}_0$ e consequentemente

$$\begin{aligned}
\text{Var}(\widehat{\varepsilon_0}|\boldsymbol{x}, x_0) &= \text{Var}(y_0|\boldsymbol{x}, x_0) + \text{Var}(\widehat{y}_0|\boldsymbol{x}, x_0) \\
&= \sigma_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]; \quad \text{e} \\
\text{DP}\,(\widehat{\varepsilon_0}|\boldsymbol{x}, x_0) &= \sigma_\varepsilon \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]^{1/2}.
\end{aligned}$$

Consequentemente, o erro padrão da previsão é dado por

$$\text{ep}(\widehat{\varepsilon}_0|\boldsymbol{x}, x_0) = \widehat{\text{DP}}(\widehat{\varepsilon}_0|\boldsymbol{x}, x_0) = s\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]^{1/2}.$$

Sob **HP.3**, $\widehat{y}_0$ é combinação linear de normais. Logo,

$$(\widehat{\varepsilon}_0|\boldsymbol{x}, x_0) \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]\right)$$

Daí, temos que

$$\text{IC}_{100(1-\alpha)\%}(y_0|\boldsymbol{x}, x_0) = (\widehat{y}_0 - t_{n-2;\,1-\alpha/2} \times \text{ep}(\widehat{\varepsilon}_0|\boldsymbol{x}, x_0),\ \widehat{y}_0 - t_{n-2;\,1-\alpha/2} \times \text{ep}(\widehat{\varepsilon}_0|\boldsymbol{x}, x_0)).$$

Ver arquivo exemplo_01.r

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

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

em que

- y é a variável dependente;
- $(x_1, x_2, \ldots, x_p)^\top$ é o vetor de variáveis regressoras (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

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

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 regressoras do modelo;

HP.3: Exogeneidade das variáveis regressoras:
$\mathrm{E}(\varepsilon_i | x_{i1}, x_{i2}, \ldots, x_{ip}) = 0 \quad \Rightarrow \quad \mathrm{Cov}(\boldsymbol{x}_i, \varepsilon_i) = \mathbf{0}$
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 regressoras;
- $\boldsymbol{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kn})^\top$ é a amostra da k-ésima variável regressora;
- $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ é a amostra 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 regressoras.

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

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

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.

Distúrbios do modelo de regresssão

Temos:

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

$$\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \begin{bmatrix} \mathrm{E}(\varepsilon_1|\boldsymbol{x}_1) \\ \mathrm{E}(\varepsilon_2|\boldsymbol{x}_2) \\ \vdots \\ \mathrm{E}(\varepsilon_n|\boldsymbol{x}_n) \end{bmatrix} = \mathbf{0};$$

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

  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 $\mathrm{Cov}(\varepsilon_i, \boldsymbol{x}_i) = \mathbf{0}$.
- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \Rightarrow \quad \mathrm{E}(\boldsymbol{y}|\boldsymbol{X}) = \mathrm{E}(\boldsymbol{X}\boldsymbol{\beta}|\boldsymbol{X}) + \mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{0} = \boldsymbol{X}\boldsymbol{\beta}$.

Distúrbios esféricos

Temos:

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

$$\begin{aligned}
\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top|\boldsymbol{X}) &= \begin{bmatrix} \text{E}(\varepsilon_1\varepsilon_1|\boldsymbol{x}_1, \boldsymbol{x}_1) & \text{E}(\varepsilon_1\varepsilon_2|\boldsymbol{x}_1, \boldsymbol{x}_2) & \cdots & \text{E}(\varepsilon_1\varepsilon_n|\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \text{E}(\varepsilon_2\varepsilon_1|\boldsymbol{x}_2, \boldsymbol{x}_1) & \text{E}(\varepsilon_2\varepsilon_2|\boldsymbol{x}_2, \boldsymbol{x}_2) & \cdots & \text{E}(\varepsilon_2\varepsilon_n|\boldsymbol{x}_2, \boldsymbol{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ \text{E}(\varepsilon_n\varepsilon_1|\boldsymbol{x}_n, \boldsymbol{x}_1) & \text{E}(\varepsilon_n\varepsilon_2|\boldsymbol{x}_n, \boldsymbol{x}_2) & \cdots & \text{E}(\varepsilon_n\varepsilon_n|\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\boldsymbol{I};
\end{aligned}$$

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

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 6: 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_07.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

$$\boldsymbol{e} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{y}^{\star} \quad \Rightarrow \quad \boldsymbol{z}^{\star\top}\boldsymbol{e} = \boldsymbol{z}^{\star\top}\boldsymbol{y}^{\star} = c(\boldsymbol{z}^{\star\top}\boldsymbol{z}^{\star}),$$

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)

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

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 7: 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_07.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

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

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 8: 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	$\boldsymbol{y}^\top \boldsymbol{y} - n\overline{y}^2$	$n-1$	$S_{yy}/(n-1) = s_y^2$

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

Exemplo

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

Tabela 9: 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_07.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 $\boldsymbol{z}^\star = \boldsymbol{M}\boldsymbol{z}$, $\boldsymbol{y}^\star = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{e} \Rightarrow \boldsymbol{e}^\top \boldsymbol{e} = \boldsymbol{y}^{\star\top} \boldsymbol{y}^\star$, 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

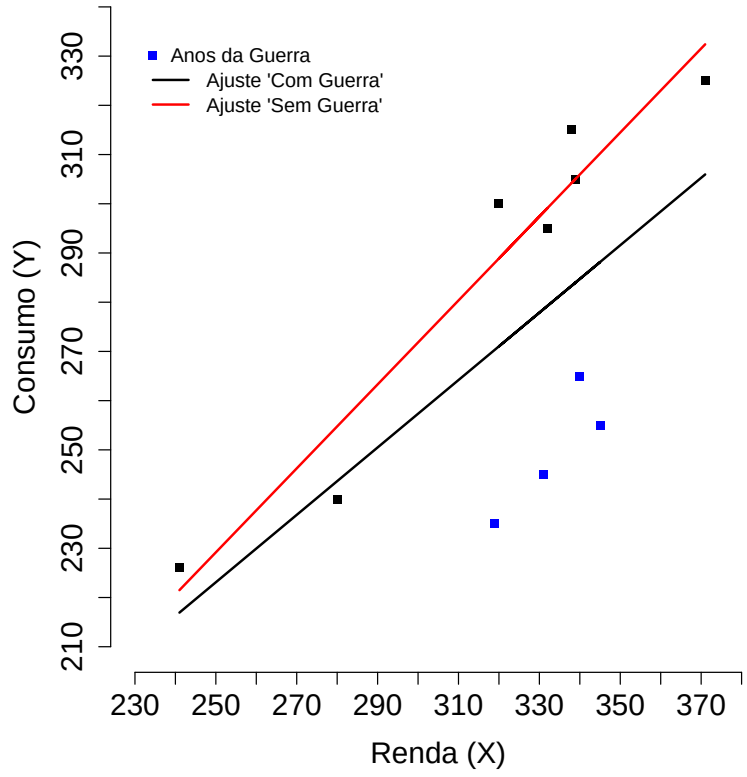


Figura 8: Dados de consumo e renda de 1940-1950 dos EUA. Fonte: *Economic Report of the President, U.S. Government Printing Office, Washington, D.C., 1983.*

Ver arquivo exemplo_08.r

Exemplo (continuação) - ajuste de uma função consumo

Tabela 10: 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 11: 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 covariável é adicionada ao modelo.
- Na verdade, $\overline{R}^2$ **pode até ser negativo.**

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

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

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 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 $\boldsymbol{X}_1^\top \boldsymbol{X}_2 = \mathbf{0}$).

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

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

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 $\boldsymbol{M}$ (simétrica e idempotente) 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.

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 simétrica e idempotente é igual ao seu traço (um número inteiro não negativo).

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 simétrica e 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 ($\boldsymbol{A}$ simétrica e 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 simétrica e 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 $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$.

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,

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

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

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

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 12: 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_09.r

Tabela 13: 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 14: 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):

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

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.

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

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

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

Os modelos em $\mathcal{M}_1$ e $\mathcal{M}_2$ 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

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

uma forma quadrática simétrica e 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 simétricas e 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_10.r

Tabela 15: 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_10.r

Tabela 16: 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_10.r

O intervalo 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

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

- 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 $\boldsymbol{\lambda}_\star$ é irrestrito, utilizamos $2\boldsymbol{\lambda}_\star$ 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:

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

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 $\mathcal{F}$:

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

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.

O modelo de Cobb-Douglas é obtido com $\beta_4 = \beta_5 = \beta_6 = 0$.

Ver arquivo exemplo_11.r

Tabela 17: Resumo do ajuste do modelo translog.

	Estimativa	Erro padrão	Estat. t	Valor p
Constante	0,9442	2,9108	0,324	0,7489
ln TR	3,6136	1,5481	2,334	0,0296
ln CP	-1,8931	1,0163	-1,863	0,0765
$(\ln \text{TR})^2/2$	-0,9641	0,7074	-1,363	0,1874
$(\ln \text{CP})^2/2$	0,0853	0,2926	0,291	0,7735
$(\ln \text{TR})(\ln \text{CP})$	0,3124	0,4389	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**.

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_12.r

Tabela 18: Resumo do ajuste do modelo de rendimentos.

Parâmetro	Estimativa	Erro padrão	Estat. t	Valor p
β_1	3,2401	1,7674	1,833	0,0675
β_2	0,2006	0,0839	2,392	0,0172
β_3	-0,0023	0,0010	-2,345	0,0195
β_3	0,0675	0,0253	2,672	0,0078
β_5	-0,3512	0,1475	-2,380	0,0177

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

O valor $-0,3512$ 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**).

$1 - \exp(-0,3512) = 0,2962 \simeq 1/3$ (mulheres com filhos recebem quase 1/3 menos).

Possíveis aplicações de variáveis dicotômicas

- Estudo do efeito de tratamentos em algum tipo de resposta.
- 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_13.r

Forma funcional e mudança de estrutura (continuação)

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

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

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.

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** (ver próxima página).

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

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.

Método delta

Caso univariado: Seja X_n uma sequência de variáveis aleatórias tal que

$$X_n \xrightarrow{P} \theta \quad \text{e} \quad \sqrt{n}(X_n - \theta) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

então

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{D} \mathcal{N}(0, \sigma^2[g'(\theta)]^2)$$

desde que $g'(\theta)$ exista e seja diferente de zero.

Caso multivariado: Seja $\boldsymbol{X}_n$ uma sequência de vetores aleatórios tal que

$$\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{\theta} \quad \text{e} \quad \sqrt{n}(\boldsymbol{X}_n - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma})$$

então

$$\sqrt{n}(\boldsymbol{h}(\boldsymbol{X}_n) - \boldsymbol{h}(\boldsymbol{\theta})) \xrightarrow{D} \mathcal{N}_k(\mathbf{0}, \boldsymbol{\nabla h}(\boldsymbol{\theta})^\top \boldsymbol{\Sigma} \boldsymbol{\nabla h}(\boldsymbol{\theta}))$$

desde que $\boldsymbol{\nabla h}(\boldsymbol{\theta})$ exista e seja bem definido.

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,

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

o estimador de mínimos quadrados ordinários (OLS) é

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

- 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. Então, 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_14.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_14.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.

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

$$\boldsymbol{I}(\boldsymbol{\theta}_0) = \mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \Big| \theta_0 \right).$$

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: $\mathrm{E}(\boldsymbol{g}_i(\boldsymbol{\theta})|\theta_0) = \mathbf{0}$; e

D2: $\mathrm{Var}(\boldsymbol{g}_i(\boldsymbol{\theta})|\theta_0) = \mathrm{E}(-\boldsymbol{H}_i(\boldsymbol{\theta})|\theta_0)$.

Teorema

$\mathrm{E}\left(\frac{1}{n}\ell(\boldsymbol{\theta}_0)\middle|\theta_0\right) > \mathrm{E}\left(\frac{1}{n}\ell(\boldsymbol{\theta})\middle|\theta_0\right)$ 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

$$\begin{aligned}
[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1} &= \left[\mathrm{E}\left(-\frac{\partial^2\ell(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0^\top}\middle|\theta_0\right)\right]^{-1} \\
&= \left[\mathrm{E}\left(\left(\frac{\partial\ell(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0}\right)\left(\frac{\partial\ell(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0}\right)^\top\middle|\theta_0\right)\right]^{-1}.
\end{aligned}$$

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.

Comparando modelos

Duas medidas comuns para modelos não encaixados baseados na mesma lógica são

- Critério de Informação de Akaike: $AIC = -2 \ln L(\widehat{\theta}) + 2p$; e
- Critério de Informação Bayesiana/Schwarz: $BIC = -2 \ln L(\widehat{\theta}) + p \ln n$,

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, $\sqrt{n}(\widehat{\sigma}^2_{MV} - \sigma^2) \xrightarrow{d} \mathcal{N}(0, 2\sigma^4)$.

Teste F

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

$$
\begin{aligned}
y_i &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \ldots, n \\
\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) &= \mathbf{0} \\
\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top|\boldsymbol{X}) &= \sigma^2\boldsymbol{\Omega}.
\end{aligned}
$$

Por hipótese, teremos $\boldsymbol{\Omega}$ como uma matriz constante conhecida.

$$
\begin{aligned}
\ell(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\ln\left|\boldsymbol{\Omega}\right| - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
\frac{\partial\ell}{\partial\boldsymbol{\beta}} &= \frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2}\boldsymbol{X}_\star^\top(\boldsymbol{y}_\star - \boldsymbol{X}_\star\boldsymbol{\beta}) = \mathbf{0} \qquad \boldsymbol{\Omega}^{-1} = \boldsymbol{P}^\top\boldsymbol{P} \\
\frac{\partial\ell}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \qquad \boldsymbol{X}_\star = \boldsymbol{P}\boldsymbol{X} \\
&= -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4}(\boldsymbol{y}_\star - \boldsymbol{X}_\star\boldsymbol{\beta})^\top(\boldsymbol{y}_\star - \boldsymbol{X}_\star\boldsymbol{\beta}) = 0. \qquad \boldsymbol{y}_\star = \boldsymbol{P}\boldsymbol{y} \\
\widehat{\boldsymbol{\beta}}_{MV} &= (\boldsymbol{X}_\star^\top\boldsymbol{X}_\star)^{-1}\boldsymbol{X}_\star^\top\boldsymbol{y} = (\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{y} \quad \text{e} \\
\widehat{\sigma}^2_{MV} &= \frac{1}{n}(\boldsymbol{y}_\star - \boldsymbol{X}_\star\widehat{\boldsymbol{\beta}}_{MV})^\top(\boldsymbol{y}_\star - \boldsymbol{X}_\star\widehat{\boldsymbol{\beta}}_{MV}) \\
&= \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{MV})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{MV}).
\end{aligned}
$$

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 $\boldsymbol{\Omega} = \mathrm{diag}(\exp(\boldsymbol{z}_1^\top \boldsymbol{\gamma}), \exp(\boldsymbol{z}_2^\top \boldsymbol{\gamma}), \ldots, \exp(\boldsymbol{z}_n^\top \boldsymbol{\gamma}))$.

$$
\begin{aligned}
\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\ln\sigma_i^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{\varepsilon_i^2}{\sigma_i^2} \\
&= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{z}_i^\top\boldsymbol{\gamma} - \frac{1}{2}\sum_{i=1}^{n}\frac{\varepsilon_i^2}{\exp(\boldsymbol{z}_i^\top\boldsymbol{\gamma})} \\
\frac{\partial\ell(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{n}\boldsymbol{x}_i\frac{\varepsilon_i}{\exp(\boldsymbol{z}_i^\top\boldsymbol{\gamma})} = \boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon} = \mathbf{0} \\
\frac{\partial\ell(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}} &= \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{z}_i\left(\frac{\varepsilon_i^2}{\exp(\boldsymbol{z}_i^\top\boldsymbol{\gamma})} - 1\right) = \mathbf{0}. \\
\frac{\partial^2\ell(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} &= -\sum_{i=1}^{n}\frac{1}{\exp(\boldsymbol{z}_i^\top\boldsymbol{\gamma})}\boldsymbol{x}_i\boldsymbol{x}_i^\top = -\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X} \\
\frac{\partial^2\ell(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^\top} &= -\sum_{i=1}^{n}\frac{\varepsilon_i}{\exp(\boldsymbol{z}_i^\top\boldsymbol{\gamma})}\boldsymbol{x}_i\boldsymbol{z}_i^\top \\
\frac{\partial^2\ell(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top} &= -\frac{1}{2}\sum_{i=1}^{n}\frac{\varepsilon_i^2}{\exp(\boldsymbol{z}_i^\top\boldsymbol{\gamma})}\boldsymbol{z}_i\boldsymbol{z}_i^\top.
\end{aligned}
$$

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,

$$\mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top}\right) = \begin{bmatrix} \boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\boldsymbol{Z}^\top\boldsymbol{Z} \end{bmatrix} = \boldsymbol{I}(\boldsymbol{\delta}).$$

O método do escore é um algoritmo para encontrar uma solução:

$$\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t + [\boldsymbol{I}(\boldsymbol{\delta}_t)]^{-1}\,\nabla\ell(\boldsymbol{\delta}_t)$$

sendo $\nabla\ell(\boldsymbol{\delta}) = \left(\frac{\partial\ell(\boldsymbol{\delta})}{\partial\boldsymbol{\beta}^\top}, \frac{\partial\ell(\boldsymbol{\delta})}{\partial\boldsymbol{\gamma}^\top}\right)^\top$.

Como $\boldsymbol{I}(\boldsymbol{\delta})$ é bloco diagonal, temos

$$\begin{aligned}
\boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + (\boldsymbol{X}^\top \boldsymbol{\Omega}_t^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega}_t^{-1} \boldsymbol{\varepsilon}_t \\
&= \boldsymbol{\beta}_t + (\boldsymbol{X}^\top \boldsymbol{\Omega}_t^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega}_t^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_t) \\
&= (\boldsymbol{X}^\top \boldsymbol{\Omega}_t^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega}_t^{-1} \boldsymbol{y} \\
\boldsymbol{\gamma}_{t+1} &= \boldsymbol{\gamma}_t + \left[2(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \right] \left[\frac{1}{2} \sum_{i=1}^{n} \boldsymbol{z}_i \left(\frac{\varepsilon_{i,t}}{\exp(\boldsymbol{z}_i^\top \boldsymbol{\gamma}_t)} - 1 \right) \right].
\end{aligned}$$

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 $\boldsymbol{d}_{t+1} > \epsilon$, retorne ao Passo 1.

Temos também que

$$\begin{aligned}
\text{Var}(\widehat{\boldsymbol{\beta}}_{MV}) &= (\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X})^{-1} \\
\text{Var}(\widehat{\boldsymbol{\gamma}}_{MV}) &= 2(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}.
\end{aligned}$$

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

$$\begin{aligned}(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}) &\sim \mathcal{N}_p(\boldsymbol{b},\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}) \quad \text{e}\\ (\sigma^2|\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X}) &\sim \mathcal{IG}(n/2, C(\boldsymbol{\beta})),\end{aligned}$$

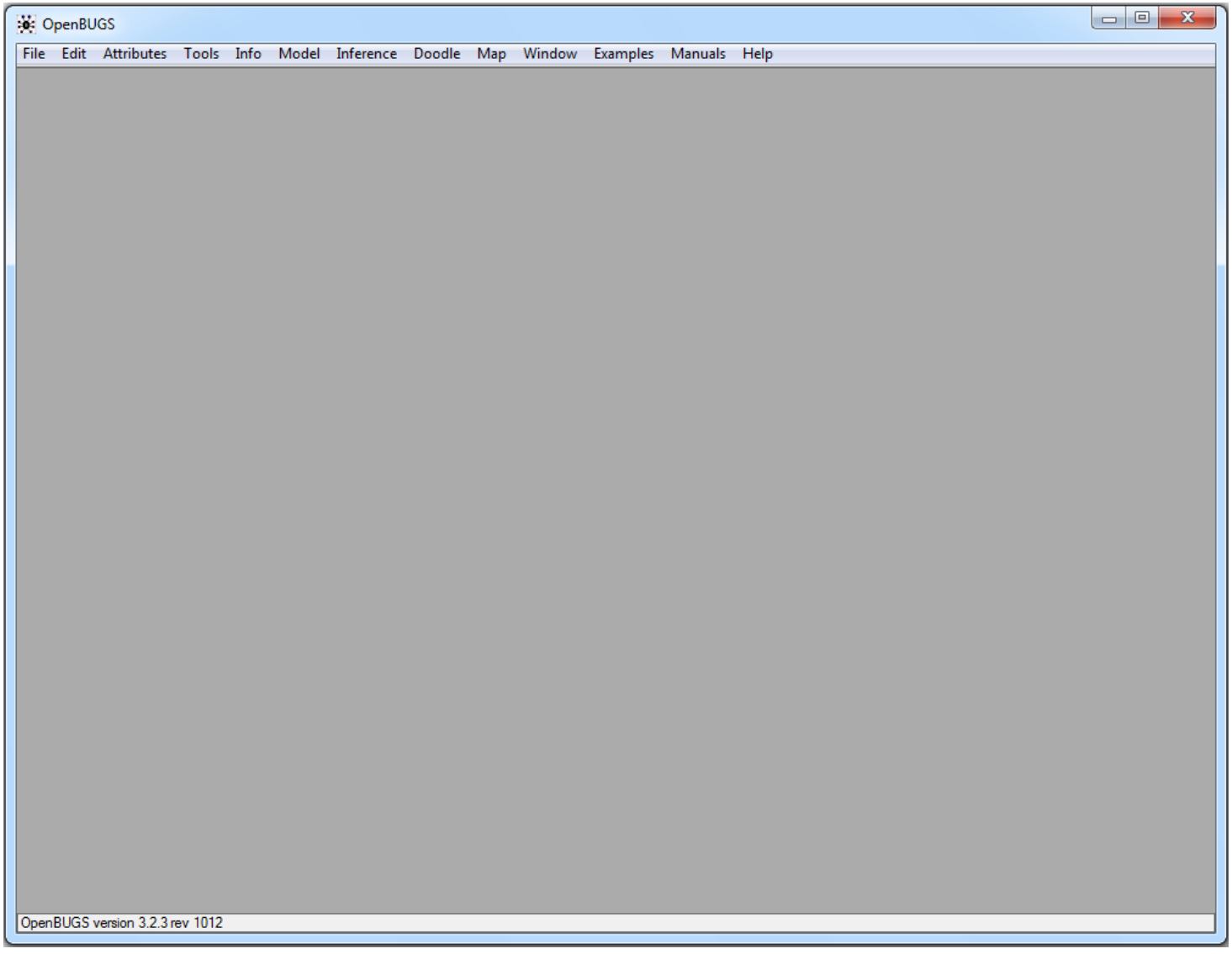
sendo $C(\boldsymbol{\beta}) = \nu\boldsymbol{s}^2 + (\boldsymbol{\beta}-\boldsymbol{b})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{b})$.

O OpenBUGS

- Bastante amigável desde o ponto de vista do usuário;
- **OpenBUGS** disponível no sítio the-bugs-project-openbugs;
- É uma ferramenta ideal para a modelagem;
- Permite mudar a especificação da distribuição a priori de forma simples;
- Em algumas aplicações a convergência pode ser lenta.

Antes, utilizamos o **WinBUGS** disponível the-bugs-project-winbugs;

O entorno do OpenBUGS



Exemplo

Considere o modelo de regressão linear múltiplo dado por:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_p x_{pt} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \quad t = 1, 2, \ldots, n.$$

A distribuição a posteriori é dada por:

$$f(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} f(\boldsymbol{\beta}, \sigma^2),$$

em que $f(\boldsymbol{\beta}, \sigma^2)$ é a distribuição a priori, $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_p})$, $\mathbf{x_i} = (x_{i1}, \ldots, x_{in})^\top$, e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$.

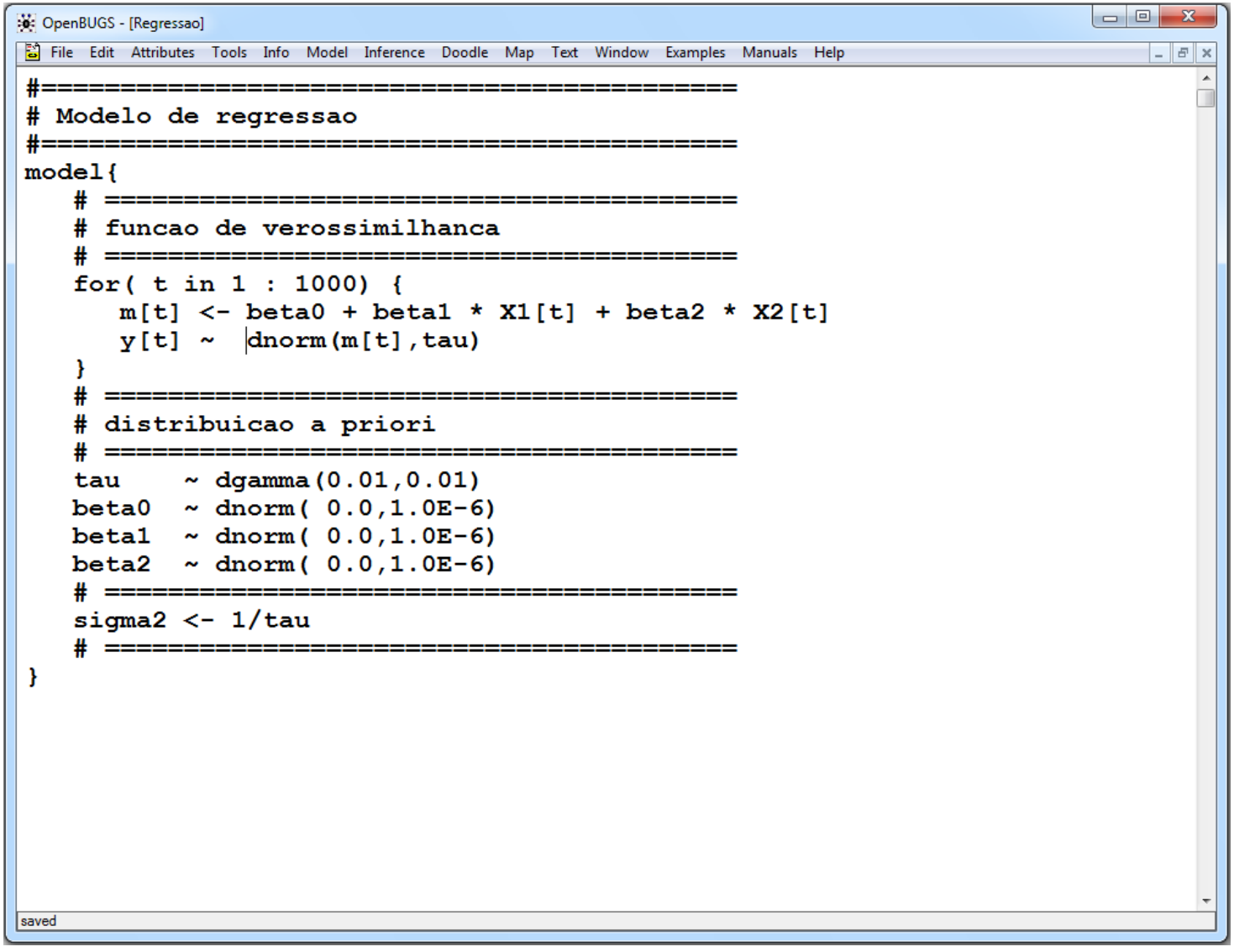
Dados artificiais gerados do seguinte modelo:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t, \quad t = 1, 2, \ldots, n,$$

com $\beta_0 = 1,0$,$\beta_1 = 5,0$, $\beta_2 = -8,0$, $\sigma^2 = 0,25$ e $n = 1000$.

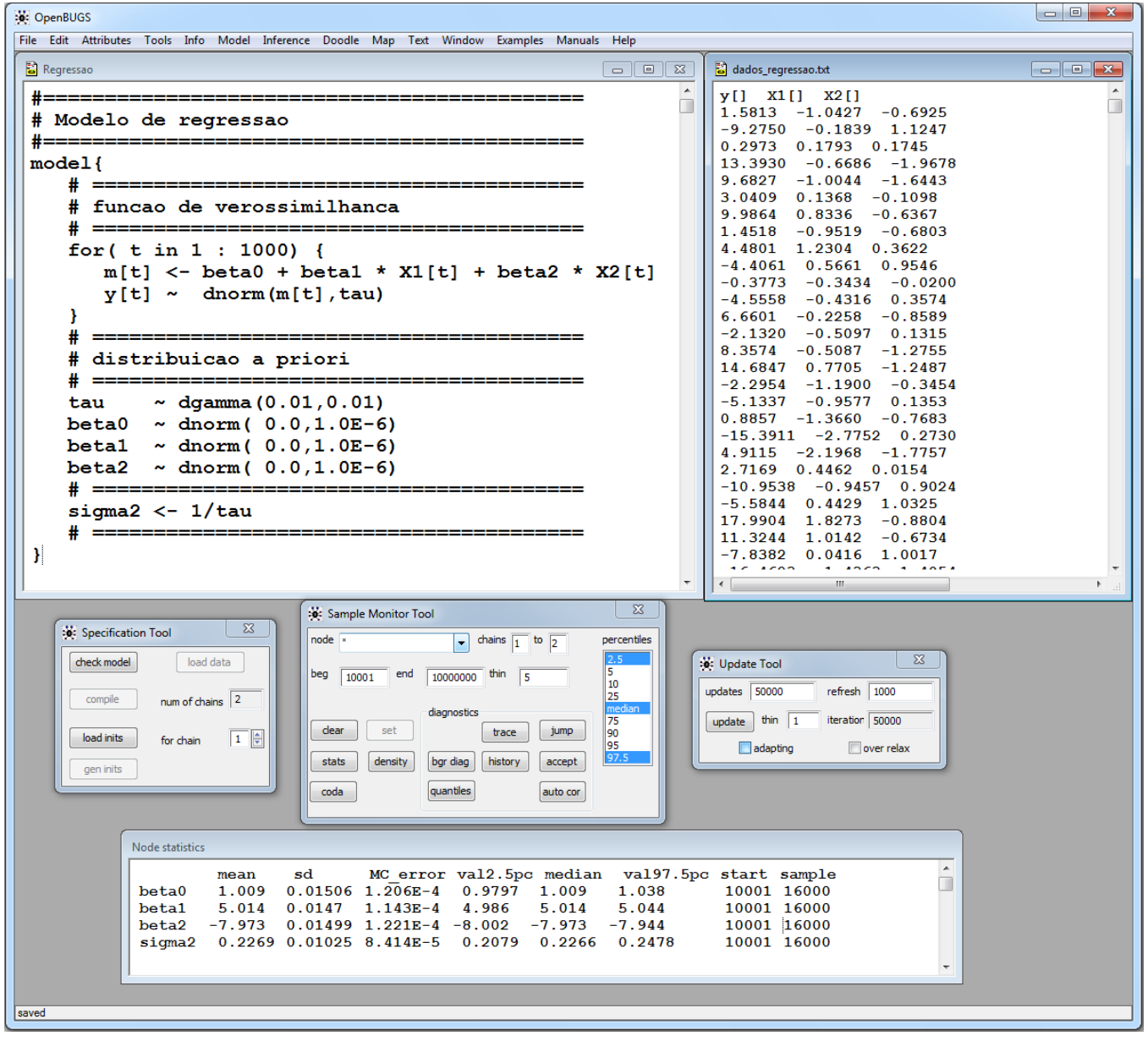
O OpenBUGS trabalha com a precisão $\phi = 1/\sigma^2$.

O modelo em código BUGS

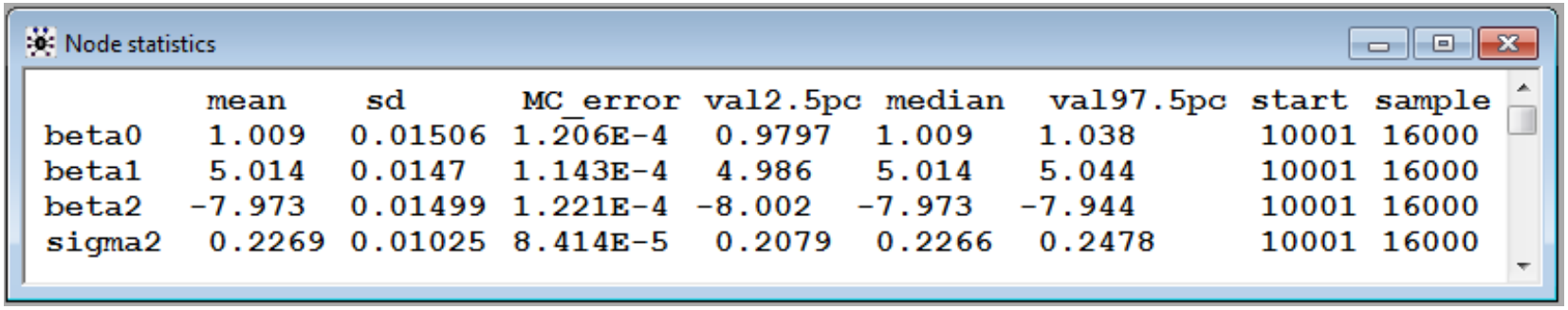


Opções básicas do OpenBUGS

1. ***Model*→ *Specification...***
2. ***check model***
   Verificar por *model is syntactically correct*
3. ***load data***
   Verificar por *data loaded*
4. ***compile***
   Verificar por *model compiled*
5. ***load inits*** ou ***gen inits***
   Verificar por *initial values generated, model initialized*
6. ***Inference*→ *Samples...***
7. ***Model*→ *Update...***

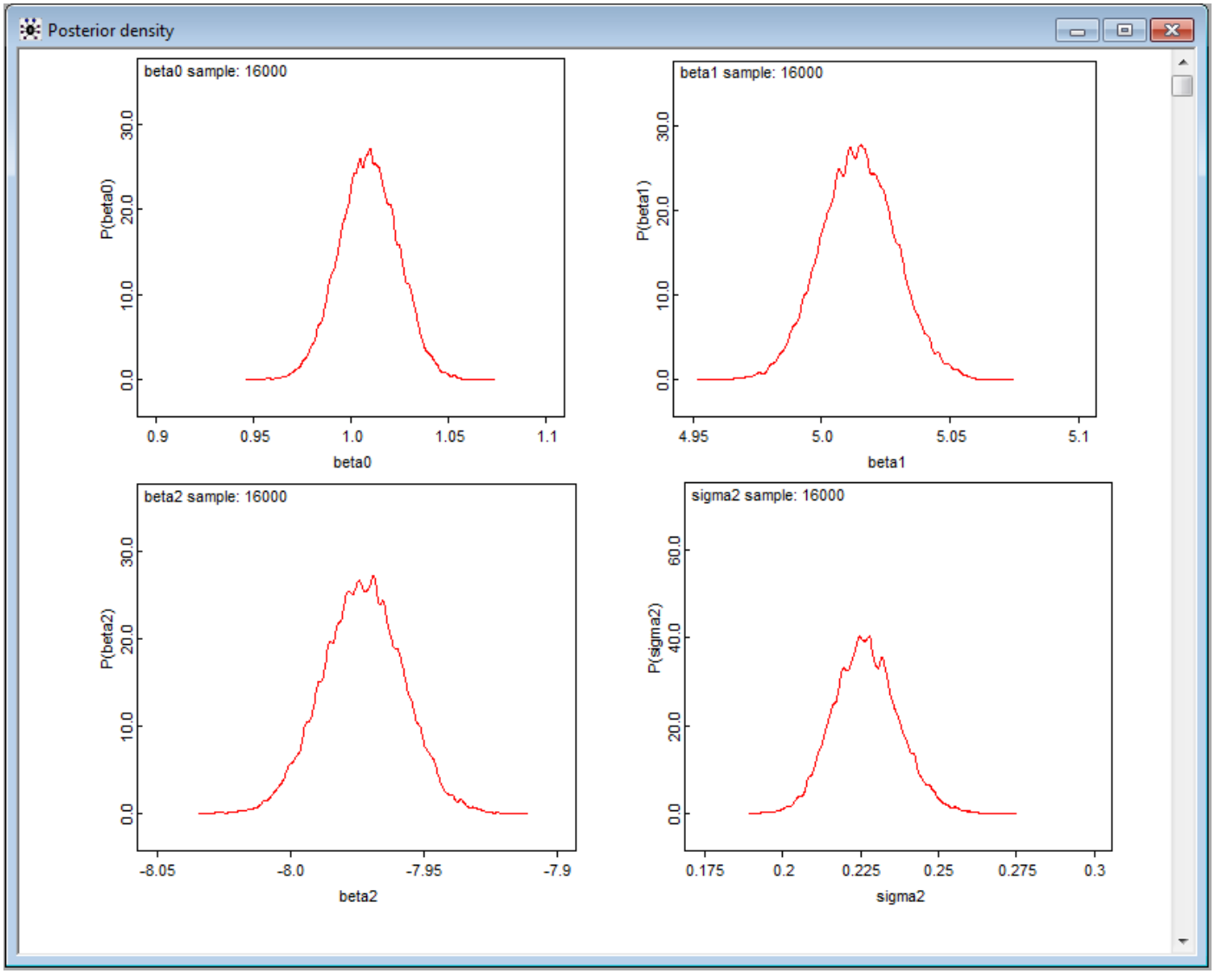

OpenBUGS
File Edit Attributes Tools Info Model Inference Doodle Map Text Window Examples Manuals Help
Regressao
#===
Modelo de regressao
#===
model{
# ===
# funcao de verossimilhanca
# ===
for(t in 1 : 1000) {
m[t] <- beta0 + beta1 * X1[t] + beta2 * X2[t]
y[t] ~ dnorm(m[t],tau)
}
# ===
# distribuicao a priori
# ===
tau ~ dgamma(0.01,0.01)
beta0 ~ dnorm(0.0,1.0E-6)
beta1 ~ dnorm(0.0,1.0E-6)
beta2 ~ dnorm(0.0,1.0E-6)
# ===
sigma2 <- 1/tau
# ===
}
dados_regressao.txt
y[] X1[] X2[]
1.5813 -1.0427 -0.6925
-9.2750 -0.1839 1.1247
0.2973 0.1793 0.1745
13.3930 -0.6686 -1.9678
9.6827 -1.0044 -1.6443
3.0409 0.1368 -0.1098
9.9864 0.8336 -0.6367
1.4518 -0.9519 -0.6803
4.4801 1.2304 0.3622
-4.4061 0.5661 0.9546
-0.3773 -0.3434 -0.0200
-4.5558 -0.4316 0.3574
6.6601 -0.2258 -0.8589
-2.1320 -0.5097 0.1315
8.3574 -0.5087 -1.2755
14.6847 0.7705 -1.2487
-2.2954 -1.1900 -0.3454
-5.1337 -0.9577 0.1353
0.8857 -1.3660 -0.7683
-15.3911 -2.7752 0.2730
4.9115 -2.1968 -1.7757
2.7169 0.4462 0.0154
-10.9538 -0.9457 0.9024
-5.5844 0.4429 1.0325
17.9904 1.8273 -0.8804
11.3244 1.0142 -0.6734
-7.8382 0.0416 1.0017
Specification Tool
check model
load data
compile
num of chains 2
load inits
for chain 1
gen inits
Sample Monitor Tool
node *
chains 1 to 2
beg 10001
end 10000000
thin 5
percentiles
2.5
5
10
25
median
75
90
95
97.5
diagnostics
clear
set
trace
jump
stats
density
bgr diag
history
accept
coda
quantiles
auto cor
Update Tool
updates 50000
refresh 1000
update
thin 1
iteration 50000
adapting
over relax
Node statistics
mean sd MC_error val2.5pc median val97.5pc start sample
beta0 1.009 0.01506 1.206E-4 0.9797 1.009 1.038 10001 16000
beta1 5.014 0.0147 1.143E-4 4.986 5.014 5.044 10001 16000
beta2 -7.973 0.01499 1.221E-4 -8.002 -7.973 -7.944 10001 16000
sigma2 0.2269 0.01025 8.414E-5 0.2079 0.2266 0.2478 10001 16000
saved

Resultados



Parâmetro	Média	D.P.	2,5%	Mediana	97,5%
β_0	1,009	0,0151	0,980	1,009	1,039
β_1	5,014	0,0147	4,986	5,014	5,043
β_2	-7,973	0,0150	-8,002	-7,973	-7,944
σ^2	0,227	0,0103	0,208	0,227	0,248

Resultados


Posterior density
beta0 sample: 16000
P(beta0)
0.0
10.0
20.0
30.0
0.9
0.95
1.0
1.05
1.1
beta0
beta1 sample: 16000
P(beta1)
0.0
10.0
20.0
30.0
4.95
5.0
5.05
5.1
beta1
beta2 sample: 16000
P(beta2)
0.0
10.0
20.0
30.0
-8.05
-8.0
-7.95
-7.9
beta2
sigma2 sample: 16000
P(sigma2)
0.0
20.0
40.0
60.0
0.175
0.2
0.225
0.25
0.275
0.3
sigma2