# Capítulo 2

- 2. Modelo de regressão linear
- 2.1. Motivação e interpretação
- 2.2. O método dos mínimos quadrados
- 2.3. Forma funcional e interpretação dos parâmetros
- 2.4. Estimação da variância da variável residual
- 2.5. Coeficiente de determinação
- 2.6. Hipóteses e propriedades estatísticas do estimador OLS para dados seccionais
- 2.7. Estimação das variâncias
- 2.8. Exemplos com aplicações empíricas

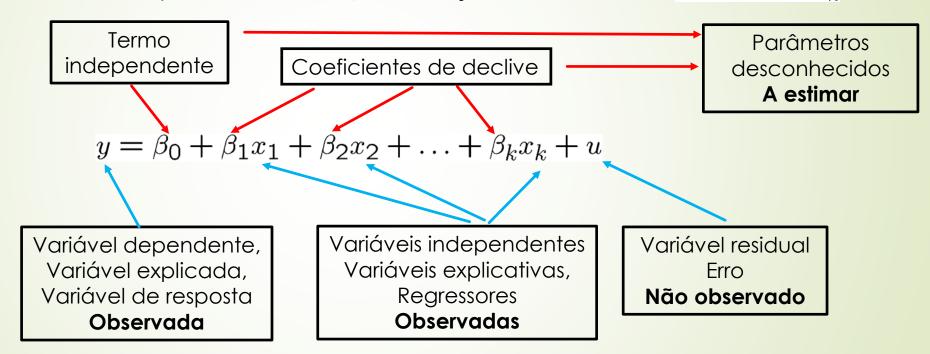
### 2.1. Motivação e interpretação

- Modelo de Regressão Linear Simples
- 1. Fertilidade:  $Filhos = \beta_0 + \beta_1 educ + u$
- 2. Peso de um bebé:  $peso = \beta_0 + \beta_1 cigs + u$
- 3. Salário:  $sal = \beta_0 + \beta_1 educ + u$
- Modelo de Regressão Linear Múltipla: análise ceteris paribus
- 1. Fertilidade:  $Filhos = \beta_0 + \beta_1 educ + \beta_2 idade + \beta_3 rend + u$
- 2. Peso de um bebé:  $peso = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 nconsul + u$
- 3. Salário:  $sal = \beta_0 + \beta_1 educ + \beta_2 exper + u$

### 2.2. O método dos mínimos quadrados

### Formalização do Modelo de Regressão Linear Múltipla

O Modelo explica a variável y em função das variáveis  $x_1, x_2, \ldots, x_k$ 



## 2.2. O método dos mínimos quadrados (OLS)

#### Estimação OLS

Amostra aleatória

$$\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i = 1, \dots n\}$$

• Valores ajustados: considere-se as estimativas  $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ 

$$\hat{y}_{i} = \hat{\beta}_{0} + \hat{\beta}_{1} x_{i1} + \dots + \hat{\beta}_{k} x_{ik}$$

Resíduos

$$\widehat{u}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \ldots - \widehat{\beta}_k x_{ik}$$

Minimização da soma dos quadrados dos resíduos

$$\min \sum_{i=1}^{n} \widehat{u}_{i}^{2} \rightarrow \widehat{\beta}_{0}, \widehat{\beta}_{1}, \widehat{\beta}_{2}, \dots, \widehat{\beta}_{k}$$

# 2.2. O método dos mínimos quadrados (OLS)

Propriedades algébricas dos resíduos OLS

$$\hat{u}_i = y_i - \hat{y}_i$$
 com  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$ 

- 1.  $\sum_{i=1}^{n} \hat{u}_i = 0$
- **2.**  $\sum_{i=1}^{n} x_{ij} \hat{u}_i = 0$
- $\mathbf{3.} \quad \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \ldots + \hat{\beta}_k \bar{x}_k$

As médias observadas na amostra da variável explicada e das variáveis explicativas estão na regressão

# 2.3. Forma funcional e interpretação dos parâmetros

■ Note-se que:  $\beta_j = \frac{\partial y}{\partial x_j}$ 

Mede a variação de yquando  $x_j$  varia de uma unidade mantendo tudo o resto constante

 O MRLM permite manter na análise o valor das outras variáveis explicativas fixo ou constante mesmo que sejam correlacionadas com a variável em análise

Análise ceteris paribus

lacktriangle Tem que se admitir ainda que o erro não varia quando  $x_i$  varia

# 2.3. Forma funcional e interpretação dos parâmetros

**Interpretação** de  $\hat{eta}_i$ 

$$\hat{y}_I = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k \quad \text{e} \quad \hat{y}_F = \hat{\beta}_0 + \hat{\beta}_1 (x_1 + 1) + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

$$\hat{y}_{F} - \hat{y}_{I} = \Delta \hat{y} =$$

$$\hat{\beta}_{0} + \hat{\beta}_{1} x_{1} + \hat{\beta}_{1} + \hat{\beta}_{2} x_{2} + \dots + \hat{\beta}_{k} x - \hat{\beta}_{0} - \hat{\beta}_{1} x_{1} - \hat{\beta}_{2} x_{2} - \dots - \hat{\beta}_{k} x$$

Então

$$\hat{\beta}_j = \Delta \hat{y}$$
 if  $\Delta x_j = 1$  ceteris paribus

Variação total de y

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \dots + \hat{\beta}_k \Delta x_k$$

# 2.3. Forma funcional e interpretação dos parâmetros

- Interpretação de  $\hat{eta}_j$  em termos de efeito parcial depois de removido o efeito das outras variáveis
- Mostra-se que no MRLM  $\hat{eta}_i$  pode ser obtido em 2 passos:
- 1) Regressão de  $X_i$  em termos de todas as outras variáveis explicativas



expurga de  $x_j$  a influência de todas as outras variáveis explicativas

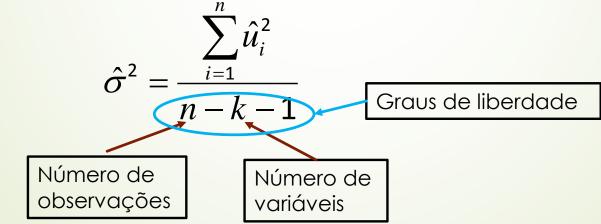
- 2) Regressão de y em função dos resíduos da regressão anterior
- Os resíduos da 1º regressão representam a parte de  $\mathcal{X}_j$  que não depende das outras variáveis explicativas
- O coeficiente de declive da segunda regressão é igual a  $\hat{eta}_j$  e representa assim o impacto de  $x_j$  depois de expurgados (removidos) os efeitos das outras variáveis

### 2.4. Estimação da variância da v. residual

- Estimação da variância da variável residual (erro)
- u é uma variável aleatória que verifica

- 
$$E(u) = 0$$
  
-  $V(u) = \sigma^2$  Parâmetro a estimar

Estimador da variância



# 2.2. O método dos mínimos quadrados (OLS)

- Coeficiente de determinação: medir a qualidade do ajustamento
  - Decomposição da variação total (supondo modelo com termo independente)



lacktriangle Coeficiente de determinação ou  $R^2$ 

$$R^2 = SSE/SST = 1 - SSR/SST$$

### 2.5. Coeficiente de determinação

#### Observações

- $0 \le R^2 \le 1$
- Sempre que se junta uma variável ao modelo o R<sup>2</sup> aumenta mesmo que o poder explicativo desta var. seja irrelevante
- Não usar o R² para comparar a capacidade explicativa de modelos com
  - 1. <u>diferente número</u> de variáveis explicativas
  - 2. variáveis dependentes definidas em escalas muito diferentes Alternativa para  $R^2$  na situação  $1: \longrightarrow \overline{R}^2$  ajustado
- o R<sup>2</sup> é igual <u>ao quadrado do coeficiente de correlação empírica</u> entre os valores observados e os valores ajustados da variável dependente

### 2.5. Coeficiente de determinação

- Observações (continuação)
  - lacktriangle Um  $R^2$  elevado não significa que exista realmente uma relação causal entre as variáveis
  - lacktriangle Um  $R^2$  baixo não é necessariamente um sintoma de que os efeitos parciais das variáveis não sejam estimados com precisão
- lacksquare Cálculo do  $ar{R}^2$

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{(SSR/n)}{(SST/n)}$$
 é uma estimativa de 1

 $\frac{1-\frac{\sigma_u^2}{\sigma_y^2}}{1-\frac{\sigma_u^2}{\sigma_y^2}}$ 

$$\bar{R}^2 = 1 - \frac{(SSR/(n-k-1))}{(SST/(n-1))} = adjusted R^2$$

Correção dos graus de liberdade do numerador e do denominador

$$\bar{R}^2 = 1 - (1 - R^2)(n - 1)/(n - k - 1)$$

Pode ser negativo

Hipótese DS.1: Modelo linear nos parâmetros

Na população, a relação entre y e os Coeficientes é linear

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

■ Hipótese DS.2: Amostragem aleatória

$$\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i = 1, \dots n\}$$

O modelo inclui uma variável residual (não observada)

Os dados são constituem uma amostra puramente aleatória da população

 $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + u_i$ 

Cada observação verifica o modelo da população

- Hipótese DS.3: não existe multicolinearidade perfeita entre as variáveis explicativas
  - Nenhuma variável pode ser constante
  - Nenhuma variável pode ser uma combinação linear da outra
- Exemplo de multicolienaridade perfeita

$$voteA = \beta_0 + \beta_1 shareA + \beta_2 shareB + u$$

Votação em A num sistema bipartidário

Proporção de votos em A nas eleições anteriores Proporção de votos em B nas eleições anteriores

shareA+shareB =1

Hipótese DS.4

O valor das variáveis explicativas não deve conter informação sobre a média do erro

$$E(u_i|x_{i1},x_{i2},\ldots,x_{ik})=0$$



$$Corr(u, x_j) = 0 \ \forall j = 1, ..., k$$

- Variáveis explicativas que são correlacionadas com o erro são consideradas endógenas
- Variáveis explicativas que não são correlacionadas com o erro são consideradas exógenas
- A hipótese DS.4 verifica-se quando todas as variáveis são exógenas

#### ■ Teorema 1

Sob as hipóteses DS.1 a DS.4 demonstra-se que o estimador OLS é centrado:

$$E(\hat{\beta}_j) = \beta_j \ j = 1, ..., k$$

► Hipótese DS.5: Homocedasticidade

$$Var(u_i|x_{i1},x_{i2},\ldots,x_{ik}) = \sigma^2$$

O valor das variáveis explicativas não pode conter qualquer informação sobre a variância do erro

A variância do erro não pode depender do valor das variáveis explicativas

#### Variância do estimador OLS

Sob as hipóteses DS.1 a DS.5 então,

Variância do erro

$$Var(\widehat{\beta}_j) = \underbrace{SST_j(1-R_j^2)}, \quad j = 1, \dots, k$$

Variação total na amostra da v. explicativa j

 $R^2$  da regressão da variável explicativa j sobre todas as outras variáveis explicativas (incluindo a constante)

$$\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

#### Componentes da variância OLS

#### 1. <u>Variância do Erro</u>

- Quando maior a variância do erro, menor a precisão com que se estimam os coeficientes (maior a variância do OLS) porque existe mais "distúrbio" no modelo
- A variância do erro não diminui aumentando o nº de observações na amostra

#### 2. <u>Variação total do regressor j</u>

- Quanto maior a variação total de  $x_j$  maior a precisão com que se estima  $\beta_i$
- lacktriangle a variação total de  $x_j$  pode aumentar quando aumenta a dimensão da amostra

- 3. Correlação linear entre os regressores do modelo
  - lacktriangleright O  $R_j^2$  (coeficiente de determinação da regressão de  $x_j$  sobre todas as outras variáveis explicativas) é tanto maior quanto maior for a correlação entre  $x_j$  e as outras variáveis do modelo.
  - lacktriangle Quanto maior for  $R_j^2$  menor a precisão com que se estima $eta_j$  .

Evitar de introduzir no modelo variáveis explicativas muito correlacionadas

• Quando a correlação entre as variáveis é muito elevada tem-se um problema de multicolinearidade, i.e. existe  $x_j:R_j\to 1$ 

#### O problema da Multicolienaridade

- Quando as variáveis estão muito correlacionadas torna-se difícil discriminar o efeito de cada uma delas.
- O problema da falta de precisão na estimação provocado pela multicolienaridade afeta a estimação dos efeitos parciais <u>apenas</u> das variáveis que têm multicolinearidade e não de todas as variáveis do modelo.
- A multicolinearidade não representa uma violação da hipótese DS.3
- Pode ser identificada através dos fatores VIF (variance inflation factor)

$$VIF_j = 1/(1 - R_j^2)$$

VIF > 10 sintomas de multicolinearidade

#### Teorema 2: Teorema de Gauss Markov

Sob as hipóteses DS.1 a DS.5 o estimador OLS para os coeficientes do modelo  $\beta_j$  j=0,1,...,k é o **mais eficiente** na classe dos estimadores centrados e lineares em y.

$$Var(\widehat{\beta}_j) \leq Var(\widetilde{\beta}_j) \quad j = 0, 1, \dots, k$$

Para qualquer 
$$ilde{eta}_j = \sum_{i=1}^n w_{ij} y_i$$
 que verifique  $E( ilde{eta}_j) = eta_j, j = 0, \dots, k$ 

#### ■ Teorema 3:

Sob as hipóteses DS.1 a DS.5  $E(\hat{\sigma}^2) = \sigma^2$ 

#### Inclusão de variáveis irrelevantes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$
  $x_3$  irrelevante  $\Rightarrow \beta_3 = 0$ 

como OLS centrado então

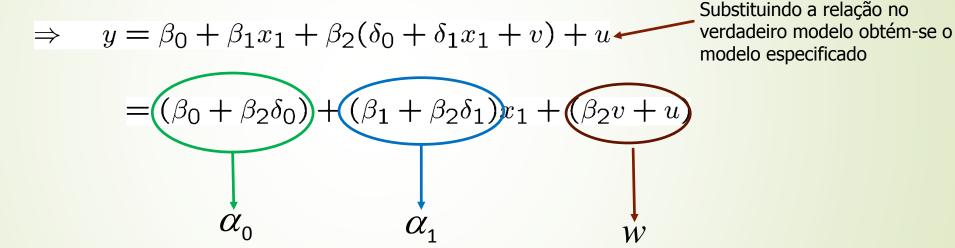
$$E(\hat{\beta}_3) = 0$$

- No entanto a inclusão de variáveis irrelevantes pode levar a que as variâncias sejam maiores.
- Omissão de variáveis relevantes Verdadeiro modelo

$$y=\beta_0+\beta_1x_1+\beta_2x_2+u$$
 
$$x_2 \text{ relevante } \Rightarrow \beta_2 \neq 0$$
 
$$y=\alpha_0+\alpha_1x_1+w \text{ Modelo especificado: omissão de } x_2$$

Enviesamento por omissão de variáveis

$$x_2 = \delta_0 + \delta_1 x_1 + v$$
 Se  $x_2$  estiver correlacionado com  $x_1$  pode assumir-se que verificam uma relação linear



- lacktriangle Em conclusão: não se consegue estimar  $eta_0,eta_1$  e  $eta_2$ .
- Os coeficientes estimados s\u00e3o enviesados para estimar os coeficientes do verdadeiro modelo.

### 2.7. Estimação das variâncias

- lacktriangle Desvios padrão de  $\hat{eta}_{j}$ 
  - Desvio padrão de  $\hat{eta}_j$  (desconhecido)

$$sd(\widehat{\beta}_j) = \sqrt{Var(\widehat{\beta}_j)} = \sqrt{\sigma^2/\left[SST_j(1-R_j^2)\right]}$$

lacksquare Erro padrão de  $\hat{eta}_j$ 

$$se(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\beta}_j)} = \sqrt{\widehat{\sigma}^2 / \left[ SST_j(1 - R_j^2) \right]}$$

Valor dado pelos softwares.