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 $x_1, x_2, \ldots, x_k$

Termo independente

Coeficientes de declive

Parâmetros desconhecidos
A estimar

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

Variável dependente,
Variável explicada,
Variável de resposta
Observada

Variáveis independentes
Variáveis explicativas,
Regressores
Observadas

Variável residual
Erro
Não observado

2.2. O método dos mínimos quadrados (OLS)

- **Estimação OLS**
  - Amostra aleatória

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

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

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$$

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

  - Minimização da soma dos quadrados dos resíduos

$$\min \sum_{i=1}^{n} \hat{u}_i^2 \quad \rightarrow \quad \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$$

2.2. O método dos mínimos quadrados (OLS)

- **Propriedades algébricas dos resíduos OLS**

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

1. $\sum_{i=1}^{n} \widehat{u}_i = 0$

2. $\sum_{i=1}^{n} x_{ij} \widehat{u}_i = 0$

3. $\bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}_1 + \ldots + \widehat{\beta}_k \bar{x}_k$

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: $\beta_j = \dfrac{\partial y}{\partial x_j}$

Mede a variação de y quando 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

- Tem que se admitir ainda que o erro não varia quando x_j varia

2.3. Forma funcional e interpretação dos parâmetros

- **Interpretação de** $\hat{\beta}_j$

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

$$\hat{y}_F - \hat{y}_I = \Delta\hat{y} =$$

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \ldots - \hat{\beta}_k x$$

Então

$$\boxed{\hat{\beta}_j = \Delta\hat{y} \quad \text{if} \quad \Delta x_j = 1 \;\; \textit{ceteris paribus}}$$

- **Variação total de** y

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

2.3. Forma funcional e interpretação dos parâmetros

- **Interpretação de $\hat{\beta}_j$ em termos de efeito parcial depois de removido o efeito das outras variáveis**

- Mostra-se que no MRLM $\hat{\beta}_j$ pode ser obtido em 2 passos:

1) Regressão de x_j em termos de todas as outras variáveis explicativas

$\Downarrow$

$\boxed{\text{expurga de } x_j \text{ a influência de todas as outras variáveis explicativas}}$

2) Regressão de y em função dos resíduos da regressão anterior

- Os resíduos da 1º regressão representam a parte de x_j que não depende das outras variáveis explicativas
- O coeficiente de declive da segunda regressão é igual a $\hat{\beta}_j$ 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

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1}$$

Graus de liberdade

Número de observações

Número de variáveis

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)

$$SST = SSE + SSR$$

Variação Total: soma total dos quadrados

Variação explicada: soma explicada dos quadrados

Variação residual: soma dos quadrados dos resíduos

  - Coeficiente de determinação ou R^2

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

2.5. Coeficiente de determinação

- **Observações**
  - $0 \leq R^2 \leq 1$
  - Sempre que se junta uma variável ao modelo o R^2 aumenta mesmo que o poder explicativo desta var. seja irrelevante
  - Não usar o R^2 para comparar a capacidade explicativa de modelos com
    1. diferente número de variáveis explicativas
    2. variáveis dependentes definidas em escalas muito diferentes

    Alternativa para R^2 na situação **1**: ➡ $\bar{R}^2$ ajustado
  - o R^2 é igual ao quadrado do coeficiente de correlação empírica entre os valores observados e os valores ajustados da variável dependente

2.5. Coeficiente de determinação

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

R^2 da população

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

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

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

2.6. Hipóteses e Propriedades estatísticas

- Hipótese DS.1: Modelo linear nos parâmetros

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

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

O modelo inclui uma variável residual (não observada)

- Hipótese DS.2: Amostragem aleatória

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

Os dados são constituem uma amostra puramente aleatória da população

Cada observação verifica o modelo da população

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

2.6. Hipóteses e Propriedades estatísticas

- 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

$\boxed{\text{shareA+shareB =1}}$

2.6. Hipóteses e Propriedades estatísticas

- Hipótese DS.4

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

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

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

  - 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

2.6. Hipóteses e Propriedades estatísticas

- **Teorema 1**

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

$$E\left(\hat{\beta}_j\right) = \beta_j \quad j = 1,\ldots,k$$

- 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

2.6. Hipóteses e Propriedades estatísticas

- **Variância do estimador OLS**

Sob as hipóteses DS.1 a DS.5 então,

Variância do erro

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \quad j = 1, \ldots, k$$

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)

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

2.6. Hipóteses e Propriedades estatísticas

- **Componentes da variância OLS**

1. Variância do Erro
   - 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. Variação total do regressor *j*
   - Quanto maior a variação total de x_j maior a precisão com que se estima β_j
   - a variação total de x_j pode aumentar quando aumenta a dimensão da amostra

2.6. Hipóteses e Propriedades estatísticas

3. Correlação linear entre os regressores do modelo

- 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.

- Quanto maior for R_j^2 menor a precisão com que se estima β_j .

$\Downarrow$

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 $x_j : R_j \to 1$

2.6. Hipóteses e Propriedades estatísticas

- **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 apenas 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 $\longrightarrow$ **sintomas de multicolinearidade**

2.6. Hipóteses e Propriedades estatísticas

- **Teorema 2: Teorema de Gauss Markov**
  - Sob as hipóteses DS.1 a DS.5 o estimador OLS para os coeficientes do modelo $\beta_j \;\; j=0,1,\ldots,k$ é o **mais eficiente** na classe dos estimadores centrados e lineares em y.

$$Var(\hat{\beta}_j) \leq Var(\tilde{\beta}_j) \quad j=0,1,\ldots,k$$

Para qualquer $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$ que verifique $E(\tilde{\beta}_j) = \beta_j, j=0,\ldots,k$

- **Teorema 3:**
  - Sob as hipóteses DS.1 a DS.5 $E\left(\hat{\sigma}^2\right) = \sigma^2$

2.6. Hipóteses e Propriedades estatísticas

- **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

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

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

- **Omissão de variáveis relevantes**

Verdadeiro modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

x_2 relevante $\Rightarrow \beta_2 \neq 0$

$$y = \alpha_0 + \alpha_1 x_1 + w$$

Modelo especificado: omissão de X_2

2.6. Hipóteses e Propriedades estatísticas

- 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

$$\Rightarrow \quad y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u$$

Substituindo a relação no verdadeiro modelo obtém-se o modelo especificado

$$= \underbrace{(\beta_0 + \beta_2\delta_0)}_{\alpha_0} + \underbrace{(\beta_1 + \beta_2\delta_1)}_{\alpha_1} x_1 + \underbrace{(\beta_2 v + u)}_{w}$$

- Em conclusão: não se consegue estimar β_0, β_1 e β_2.

- Os coeficientes estimados são enviesados para estimar os coeficientes do verdadeiro modelo.

2.7. Estimação das variâncias

- **Desvios padrão e erros padrão de $\hat{\beta}_j$**
  - Desvio padrão de $\hat{\beta}_j$ (desconhecido)

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

  - Erro padrão de $\hat{\beta}_j$

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

Valor dado pelos softwares.