Sebenta de Estatística II

Licenciatura em Gestão

Nuno M. Brites

2023/2024

Conteúdo

Todas as informações relacionadas com esta UC encontram-se no Fénix.

Todos os erros e omissões são da minha inteira responsabilidade. Caso detete algum erro ou gralha, muito agradeço que me informe. Sugestões e comentários também serão muito bem-vindos!

Esta sebenta não substitui o livro principal.

Obrigado,

Nuno M. Brites

nbrites@iseg.ulisboa.pt

ISEG, Fevereiro de 2024

Capítulo 1

Estimação pontual

1.1 Introdução

- Considere-se uma amostra casual $(X_1, \ldots, X_n)$ de uma população com função de probabilidade pertencente à família

$$F_\theta = \{f(x \mid \theta) : \theta \in \Theta\}.$$

- A forma funcional de $f(\cdot)$ é conhecida mas os parâmetros θ são desconhecidos. Note-se que θ pode ser um vetor ou apenas um elemento.
- **Problema**: Como usar a informação contida na amostra para "adivinhar" (estimar) o valor do(s) parâmetro(s) desconhecido(s) θ?
- **Ideia**: Fixada a dimensão da amostra, quanto mais precisa a resposta, menor a confiança que nela se deposita.
- A estimação paramétrica vai então desenvolver-se:
  - priveligiando a precisão $\Longrightarrow$ estimação pontual $\Longrightarrow$ estimativas **ou**
  - priveligiando a confiança $\Longrightarrow$ estimação intervalar $\Longrightarrow$ intervalos de confiança
- Parâmetros:
  - Parâmetro multidimensional. Exemplo: suponha que a valorização de um activo financeiro tem distribuição normal com média μ e variância σ^2. Observada uma amostra casual, queremos estimar 2 parâmetros desconhecidos (μ, σ), onde μ representa o retorno médio e σ a volatilidade.
  - Função do(s) parâmetro(s) da distribuição. Exemplo: suponha que a distribuição do número de sinistros originados anualmente por uma apólice de seguro automóvel tem distribuição de Poisson com parâmetro λ. Em vez de estarmos interessados no parâmetro λ (média do fenómeno), podemos estar interessados na probabilidade de se verificarem zero sinistros: $P(X = 0 \mid \lambda) = e^{-\lambda}$. Assim, pretende-se de estimar a função $h(\lambda) = e^{-\lambda}$, que traduz essa probabilidade.
- **Ponto de partida**:
  - amostra casual $(X_1, X_2, \ldots, X_n)$ de uma população $F_\theta = \{f(x \mid \theta) : \theta \in \Theta\}$.
  - apenas θ é desconhecido pois $f(\cdot)$ é conhecida.
  - $\theta \in \Theta$, sendo o espaço do parâmetro, Θ, conhecido (tipicamente $\mathbb{R}_0^+$).

Definição 1.1 (Estimador)**.** Um estimador é uma estatística, $T(X_1, X_2, \ldots, X_n)$, que estima alguma informação sobre a população. Um estimador é uma função que produz estimativas.

Exemplo 1.1. $T(X_1, X_2, \ldots, X_n) = \dfrac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}$.

Definição 1.2 (Estimativa)**.** Valor assumido pelo estimador para a amostra que se observou. Não é uma variável aleatória, nem uma estatística. É um número (real).

Exemplo 1.2. $t = T(x_1, x_2, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$.

Nota . Como encontrar estimadores para determinado parâmetro? Dado um ou mais estimadores, como avaliar a sua qualidade?

1.2 Método dos momentos

- **Ideia**: se soubermos que o parâmetro θ que queremos estimar é a média da população (momento populacional de ordem 1), então podemos usar a média amostral (momento amostral de ordem 1) para estimá-lo: quanto maior a dimensão amostra, mais semelhantes serão os dois.

Iremos considerar uma amostra casual $(X_1, X_2, \ldots, X_n)$ de uma população com fmp/fdp $f(x \mid \theta_1, \theta_2, \ldots, \theta_k)$ com k parâmetros desconhecidos.

Definição 1.3 (Método dos Momentos)**.** Os estimadores do Método dos Momentos podem ser obtidos da seguinte forma. Seja $\mu'_k = E(X^k)$ o momento populacional de ordem k e seja $\mu_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$ o momento amostral de ordem k.

- Passo 1: Determinar o número de parâmetros $(\theta_1, \ldots, \theta_L)$ a estimar, L.
- Passo 2: Determinar μ'_k e igualar a μ_k para $k = 1, \ldots, L$.
- Passo 3: Resolver o sistema com L equações e L incógnitas em ordem a $\theta_1, \ldots, \theta_L$.

A solução do sistema são os estimadores do Método dos Momentos $(\tilde{\theta}_1, \ldots, \tilde{\theta}_L)$.

Quando não houver ambiguidade, estimador e estimativa representam-se indistintamente por $\tilde{\theta}$.

Exercício 1.1. Seja $X \sim B(1, \theta)$ uma população donde se retirou uma amostra casual de dimensão n com o objetivo de estimar θ. Obtenha $\tilde{\theta}$.

```
#



#
```

Exercício 1.2. Seja $X \sim \mathcal{N}(\mu, \sigma^2)$ uma população donde se retirou uma amostra casual de dimensão n com o objetivo de estimar μ e σ^2. Obtenha $\tilde{\mu}$ e $\tilde{\sigma^2}$.

```
#
```

```
#
```

Exercício 1.3. Seja $X \sim U(-\theta, \theta)$ uma população donde se retirou uma amostra casual de dimensão n com o objetivo de estimar θ. Obtenha $\tilde{\theta}$.

```
#
```

```
#
```

Exercício 1.4. Seja $X \sim Exp(\lambda)$ uma população donde se retirou uma amostra casual de dimensão n com o objetivo de estimar λ. Obtenha $\tilde{\lambda}$.

```
#
```

```
#
```

1.3 Método da máxima verosimilhança

Definição 1.4 (Função de verosimilhança)**.** Função de probabilidade conjunta das observações de uma amostra aleatória, considerada como função do parâmetro desconhecido θ.

- $(X_1, X_2, \ldots, X_n)$ amostra casual de uma população com fmp/fdp $f(x \mid \theta)$.
- A fmp/fdp conjunta da amostra aleatória

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta), \qquad (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

representa a probabilidade associada à amostra específica que foi observada $(x_1, x_2, \ldots, x_n)$.

- Fixada a amostra observada, $(x_1, x_2, \ldots, x_n)$, a função $f(x_1, \ldots, x_n \mid \theta)$ pode ser interpretada como função do parâmetro θ e define a **função de verosimilhança**:

$$L(\theta) := L(\theta \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta), \qquad \theta \in \Theta.$$

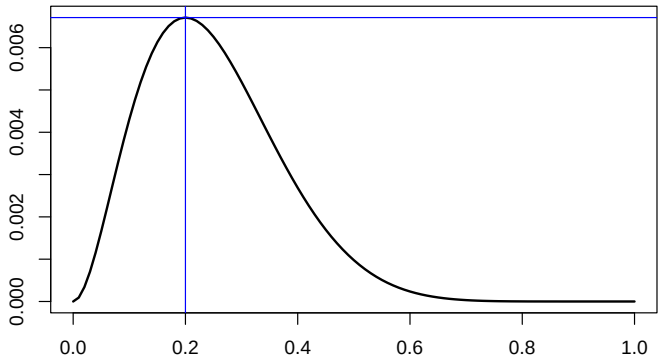
- Para cada valor de θ no espaço do parâmetro, $L(\theta)$ determina a probabilidade de observar a amostra específica $(x_1, x_2, \ldots, x_n)$.

Exercício 1.5. Seja $X \sim B(1, \theta)$ uma população donde se retirou uma amostra casual de dimensão n com o objetivo de estimar θ. Obtenha a função de verossimilhança e observe as seguintes figuras:

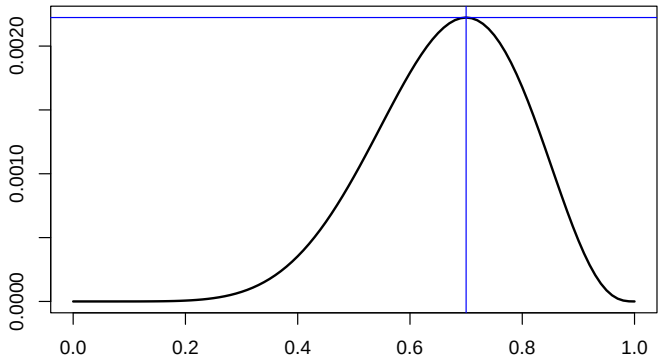
```
#
```

```
#
```

- $n = 10$ e $\sum_{i=1}^{10} x_i = 2 \Longrightarrow L(\theta) = \theta^2(1-\theta)^{10-2} = \theta^2(1-\theta)^8$. **Os valores mais verossímeis de θ situam-se em torno de** 0.2**.**



- $n = 10$ e $\sum_{i=1}^{10} x_i = 7 \Longrightarrow L(\theta) = \theta^7(1-\theta)^{10-7} = \theta^7(1-\theta)^3$. **Os valores mais verossímeis de θ situam-se em torno de** 0.7**.**



Definição 1.5 (Estimador de máxima verosimilhança)**.** Dada a amostra observada $(x_1, \ldots, x_n)$ procura-se uma estimativa $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ tal que

$$L(\hat{\theta} \mid x_1, x_2, \ldots, x_n) \geq L(\theta \mid x_1, x_2, \ldots, x_n), \qquad \theta \in \Theta.$$

A esta estimativa corresponde o estimador $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$.

- Cálculo do máximo: zero da primeira derivada (se existir) e segunda derivada negativa (se existir).
- Geralmente é mais fácil considerar o logaritmo da função de verossimilhança (função log-verosimilhança):

$$l(\theta) = \ln(L(\theta)).$$

- Como o logaritmo é uma função monótona crescente, $l(\theta)$ e $L(\theta)$ têm o mesmo maximizante.
- Em geral, o maximizante é dado por:

$$\frac{dL(\theta)}{d\theta} = 0, \quad \frac{d^2L(\theta)}{d\theta^2} < 0 \quad \text{ou} \quad \frac{dl(\theta)}{d\theta} = 0, \quad \frac{d^2l(\theta)}{d\theta^2} < 0.$$

- Quando não houver ambiguidade, estimador e estimativa representam-se indistintamente por $\widehat{\theta}$ (tal como no método dos momentos).

Nota . Deverá ter especial atenção a:

- O maximizante pode não ser um ponto interior do domínio. Relembrar que uma f.r.v.r. atinge o máximo, caso exista, no interior de um intervalo (a, b) ou na sua fronteira (ver Matemática I).
- O EMV pode não ser único.
- $L(\theta)$ pode ter máximos locais.
- A função de verosimilhança (ou log-verosimilhança) pode não ser diferenciável.

Exercício 1.6. Seja $X \sim B(1, \theta)$ uma população donde se retirou uma amostra casual de dimensão n com o objetivo de estimar θ. Obtenha $\hat{\theta}$.

```
#
```

```
#
```

Exercício 1.7. Seja $(X_1, \dots X_n)$ uma a.c. proveniente de uma população X com função probabilidade $f(x \mid \theta) = \theta x^{\theta-1}$, $0 < x < 1$, $\theta > 0$, e $E(X) = \dfrac{\theta}{1+\theta}$. Determine $\tilde{\theta}$ e $\hat{\theta}$.

```
#
```

```
#
```

Exercício 1.8. $X \sim U(0, \theta)$. Determine $\hat{\theta}$.

```
#
```

```
#
```

1.3.0.1 Propriedade da Invariância dos EMV

Teorema 1.1. *Seja $\hat{\theta}$ um EMV de θ e $h(u)$ uma função biunívoca. Então $h(\hat{\theta})$ é o EMV de $h(\theta)$.*

Exercício 1.9. $X \sim Po(\lambda)$. Determine o EMV de $P(X = 0)$.

```
#
```

```
#
```

1.4 Propriedades dos Estimadores

1.4.1 Estimadores centrados

Definição 1.6 (Estimador centrado)**.** Um estimador $T = T(X_1, X_2, \ldots, X_n)$ de θ é **centrado** (ou não enviesado) se

$$E(T) = \theta, \quad \forall \theta \in \Theta.$$

Nota . O valor esperado do estimador deve ser igual ao verdadeiro valor do parâmetro a estimar, qualquer que seja $\theta \in \Theta$. O conceito de estimador centrado só se aplica quando existe $E(T)$.

Definição 1.7 (Enviesamento)**.** Se $E(T) \neq \theta$, então o estimador diz-se **enviesado** e o seu enviesamento é dado por

$$\text{viés}(T) = E(T) - \theta.$$

Exercício 1.10. $X \sim B(1, \theta)$. Será $\hat{\theta} = \overline{X}$ enviesado para θ?

```
#
```

```
#
```

- O conceito de estimador centrado não permite distinguir estimadores que apresentem uma distribuição por amostragem fortemente concentrada em torno do parâmetro a ser estimado de outros em que a dispersão é claramente superior.
- Na imagem seguinte qual dos dois estimadores será "melhor"?

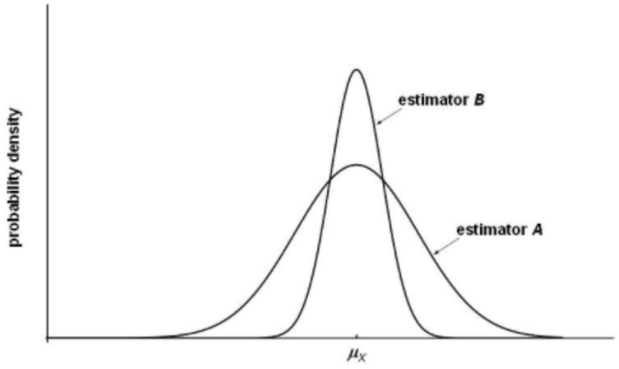


Figura 1.1: Comparação entre dois estimadores centrados.

Definição 1.8 (Eficiência)**.** Sejam T_1 e T_2 dois estimadores centrados para θ. O estimador T_1 é mais eficiente do que T_2 quando

$$Var(T_1) < Var(T_2), \quad \forall\theta \in \Theta.$$

O estimador T^* é o mais eficiente para θ quando a relação se verifica qualquer que seja o outro estimador T centrado para θ.

Nota . Sobre a eficiência dos estimadores:

- A eficiência requer a existência de momentos de segunda ordem dos estimadores.
- A definição de eficiência apresenta dois conceitos diferentes:
  - eficiência relativa: estabelece uma relação entre 2 ou mais estimadores centrados para θ.
  - eficiência absoluta: na classe de todos os estimadores centrados para θ.
- Para obter o estimador mais eficiente, recorre-se à desiguladade de **Fréchet-Cramér-Rao**.

Teorema 1.2 (Desigualdade de Fréchet-Cramér-Rao (limite inferior para a variância de um estimador))**.** *Seja $(X_1, X_2, \ldots, X_n)$ uma amostra casual de uma população com fmp/fdp $f(x \mid \theta)$, satisfazendo certas condições de regularidade, e seja $T = T(X_1, X_2, \ldots, X_n)$ um estimador centrado para θ. Então,*

$$Var(T) \geq \frac{1}{n\mathcal{I}(\theta)},$$

onde

$$\mathcal{I}(\theta) = E\left[\left(\frac{d\ln(f(X \mid \theta))}{d\theta}\right)^2\right] = -E\left[\frac{d^2\ln(f(X \mid \theta))}{d\theta^2}\right]$$

é a quantidade de ***informação do Fisher****.*

Nota . No cálculo da quantidade de informação do Fisher, geralmente a segunda igualdade é mais fácil de calcular. Para determinadas distribuições a expressão $\mathcal{I}(\theta)$ é conhecida:

Exercício 1.11. Dado $f(x \mid \theta) = \theta x^{\theta-1}$, $0 < x < 1$, $\theta > 0$, mostre que $\mathcal{I}(\theta) = \theta^{-2}$.

Distribuição	Quantidade de informação de Fisher
$X \sim B(n;\theta)$ (n conhecido)	$\mathcal{I}(\theta) = n/[\theta(1-\theta)]$
$X \sim Po(\lambda)$	$\mathcal{I}(\theta) = 1/\lambda$
$X \sim N(\mu,\sigma^2)$ (σ^2 conhecido)	$\mathcal{I}(\theta) = 1/\sigma^2$
$X \sim N(\mu,\sigma^2)$ (μ conhecido)	$\mathcal{I}(\theta) = 1/(2\sigma^4)$
$X \sim G(\alpha,\lambda)$ (α conhecido)	$\mathcal{I}(\theta) = \alpha/\lambda^2$

Figura 1.2: Quantidade de informação de Fisher.

```
#
```

```
#
```

Nota . Conhecendo o limite inferior dado pela desigualdade de Fréchet-Cramér-Rao, compara-se a variância do estimador em análise com este limite:

- caso sejam iguais, não existe nenhum outro estimador centrado de variância inferior e o estimador em análise é o mais eficiente;
- caso sejam diferentes: o quociente

$$\frac{[n\mathcal{I}(\theta)]^{-1}}{Var(T)}$$

fornece uma indicação sobre a eficiência relativa estimador T face ao estimador (caso este exista) de variância igual ao limite inferior da desigualdade de Fréchet-Cramér-Rao.

- Eficiência está associada ao conceito de estimador centrado.
- O que fazer quando se quer comparar estimadores enviesados?

Definição 1.9 (Erro Quadrático Médio)**.** Seja $T = T(X_1, X_2, \ldots, X_n)$ um estimador para θ. Define-se **Erro Quadrático Médio (**EQM**)** como

$$EQM(T) = E\left[(T-\theta)^2\right] = Var(T) + \underbrace{(E(T)-\theta)}_{\text{viés}(T)}{}^2.$$

Nota . Relativamente ao EQM:

- O EQM pondera variância e enviesamento.
- Para estimadores centrados, EQM significa variância.
- O estimador T_1 é "melhor" do que T_2 se

$$EQM(T_1) < EQM(T_2),\ \forall\theta \in \Theta.$$

- O estimador T_1 é o "melhor" estimador se o seu EQM for menor ou igual ao EQM de qualquer outro estimador para θ.
- Em geral, o EQM depende de θ.

1.4.2 Consistência

Definição 1.10 (Estimador consistente)**.** O estimador $T_n = T(X_1, X_2, \ldots, X_n)$ diz-se consistente em média quadrática se

$$\lim_{n\to+\infty} E[(T-\theta)^2] = 0.$$

Nota . Condição **necessária e suficiente** para que o estimador T_n seja consistente em média quadrática

$$\lim_{n\to+\infty} E(T_n) = \theta \quad e \quad \lim_{n\to+\infty} Var(T_n) = 0.$$

Definição 1.11 (Estimador consistente ou simplesmente consistente)**.** O estimador $T_n = T(X_1, X_2, \ldots, X_n)$ diz-se consistente ou simplesmente consistente se

$$\forall\epsilon > 0, \quad \lim_{n\to+\infty} P(\theta-\epsilon < T_n < \theta+\epsilon) = 1, \quad \forall\theta \in \Theta.$$

Nota . Consistência em média quadrática $\Longrightarrow$ consistência simples.

Propriedades dos estimadores obtidos pelo métodos dos momentos:

- Não são únicos, pois podem ser obtidos à custa de momentos de diferentes ordens.
- O estimador obtido pode não ser admissível (ex: estimador negativo quando o parâmetro é positivo).
- Não gozam da propriedade de invariância.
- Em condições gerais:
  - são consistentes;
  - Possuem distribuição aproximadamente normal quando a dimensão da amostra é muito grande (distribuição assintótica)

Propriedades dos estimadores obtidos pelo método da máxima verosimilhança:

- Não são necessariamente centrados.
- Em condições muito gerais são consistentes.
- Demonstra-se que se existir estimador mais eficiente (na ótica do teorema de Fréchet-Cramér-Rao) ele é solução da equação $\frac{dL(\theta)}{d\theta} = 0$ e portanto estimador de máxima verosimilhança.
- Verificadas certas condições de regularidade, os estimadores de máxima verosimilhança seguem assintoticamente uma distribuição normal. Caso haja apenas um parâmetro desconhecido, tem-se

$$\sqrt{n\mathcal{I}(\theta)}(\hat{\theta}-\theta) \overset{\text{a}}{\sim} \mathcal{N}(0,1).$$

Exercício 1.12. Seja

$$T = \frac{(n-1)X_1 + X_n}{n}$$

um estimador de μ, onde $(X_1, \ldots, X_n)$ representa uma amostra casual de tamanho n proveniente de uma população normal com média μ e variância σ^2. Verifique se T é centrado para μ e se é um estimador consistente em média quadrática para μ.

```
#




#
```

Capítulo 2

Estimação intervalar

2.1 Intervalos de confiança

- Nesta secção, em vez de se propor uma estimativa isolada $\hat{\theta}$ para θ, propõe-se um intervalo (t_1, t_2) ao qual se associa um determinado "nível de confiança".
- Em muitos casos, o intervalo é da forma

$$\left(\hat{\theta} - \delta, \hat{\theta} + \delta\right),$$

onde δ pode ser considerado como uma medida de precisão ou medida de erro inerente à estimativa $\hat{\theta}$.

Definição 2.1 (Intervalo aleatório)**.** Sejam $T_1 = T_1(X_1, X_2, \ldots, X_n)$ e $T_2 = T_2(X_1, X_2, \ldots, X_n)$, $T_1 < T_2$, duas estatísticas tais que

$$P(T_1 < \theta < T_2) = 1 - \alpha, \quad 0 < \alpha < 1.$$

Então, (T_1, T_2) é um **intervalo aleatório** para θ com probabilidade $1 - \alpha$.

Definição 2.2 (Intervalo de confiança)**.** Seja $(x_1, x_2, \ldots, x_n)$ um valor observado da amostra $(X_1, X_2, \ldots, X_n)$ e $t_1 = T_1(x_1, x_2, \ldots, x_n)$, $t_2 = T_2(x_1, x_2, \ldots, x_n)$ os valores observados de T_1 e T_2 para essa amostra observada.

Então, (t_1, t_2) é um **intervalo de confiança** a $(1 - \alpha)100\%$ para θ.

2.2 Variáveis fulcrais

Definição 2.3 (Variável fulcral)**.** Uma **variável fulcral**, $Z(X_1, X_2, \ldots, X_n, \theta)$:

- é uma função da amostra casual;
- é uma função do parâmetro θ;
- tem fmp/fdp $f(z)$ conhecida e independente de θ;
- é independente de qualquer outro parâmetro desconhecido.

Definição 2.4 (Obtenção de um intervalo de confiança)**.** Método:

1. Encontrar uma variável fulcral Z adequada ao problema em estudo.
2. Fixada a confiança desejada $(1 - \alpha)100\%$, determinar os dois valores no domínio de Z, $z_1(\alpha)$ e $z_2(\alpha)$, tal que

$$P\left(z_1(\alpha) < Z < z_2(\alpha)\right) = 1 - \alpha.$$

3. Passar de $z_1(\alpha) < Z < z_2(\alpha)$ para $T_1(X_1, X_2, \ldots, X_n) < \theta < T_2(X_1, X_2, \ldots, X_n)$, ou seja,

$$P\left(z_1(\alpha) < Z < z_2(\alpha)\right) = P\left(T_1(X_1, X_2, \ldots, X_n) < \theta < T_2(X_1, X_2, \ldots, X_n)\right) = 1 - \alpha.$$

4. Um intervalo aleatório com probabilidade $1-\alpha$ é:

$$(T_1(X_1, X_2, \ldots, X_n), T_2(X_1, X_2, \ldots, X_n)) = (T_1, T_2).$$

5. Finalmente, **um intervalo de confiança** a $(1-\alpha)100\%$ para θ é dado pela realização do intervalo aleatório para uma amostra observada:

$$IC_{(1-\alpha)100\%}(\theta) = (T_1(x_1, x_2, \ldots, x_n), T_2(x_1, x_2, \ldots, x_n)) = (t_1, t_2).$$

Nota . Em geral, apenas seguimos os pontos 1, 2 e 5. É importante observar que:

- as definições acima foram apresentadas para θ, mas a generalização para $h(\theta)$ é direta.
- Um intervalo de confiança é simplesmente a realização particular do intervalo aleatório, da mesma forma que uma estimativa é um valor particular de um estimador.
- Assim, atribuímos probabilidade apenas ao intervalo aleatório, não ao intervalo de confiança.
- O conceito de intervalo de confiança pode ser generalizado para dimensões superiores $(\theta_1, \theta_2, \ldots, \theta_k)$, $k > 1$, caso em que obtemos regiões de confiança.

Exercício 2.1. Considere a variável fulcral $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, e mostre que um intervalo de confiança a 95% para a média de uma população normal com σ conhecido é

$$\text{IC}_{95\%}(\mu) = \left(\overline{x} \mp z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}\right) = \left(\overline{x} - z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}\ ,\ \overline{x} + z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}\right).$$

```
#
```

```
#
```

2.3 Intervalos de confiança para populações normais

Definição 2.5. Intervalo de confiança para a **média** com **variância conhecida**:

- Variável fulcral:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

- Intervalo de confiança:

$$IC_{(1-\alpha)100\%}(\mu) = \left(\overline{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right), \qquad \Phi(z_{\alpha/2}) = 1 - \alpha/2.$$

Definição 2.6. Intervalo de confiança para a **média** com **variância desconhecida**:

- Variável fulcral:
$$T = \frac{\overline{X} - \mu}{S'/\sqrt{n}} \sim t(n-1).$$

- Intervalo de confiança:
$$IC_{(1-\alpha)100\%}(\mu) = \left(\overline{x} \mp t_{\alpha/2}\frac{s'}{\sqrt{n}}\right), \qquad P(T > t_{\alpha/2}) = \alpha/2.$$

Definição 2.7. Intervalo de confiança para a **variância**:

- Variável fulcral:
$$Q = \frac{(n-1)S'^2}{\sigma^2} \sim \chi^2(n-1).$$

- Intervalo de confiança:
$$IC_{(1-\alpha)100\%}(\sigma^2) = \left(\frac{(n-1)s'^2}{q_2}, \frac{(n-1)s'^2}{q_1}\right), \quad P(Q < q_1) = P(Q > q_2) = \alpha/2.$$

Considerem-se agora duas populações normais $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ e $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, e duas amostras casuais independentes $(X_{11}, X_{12}, \ldots, X_{1m})$ e $(X_{21}, X_{22}, \ldots, X_{2n})$.

Definição 2.8. Intervalo de confiança para a diferença de **médias** com **variâncias conhecidas**:

- Variável fulcral:
$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim \mathcal{N}(0,1).$$

- Intervalo de confiança:
$$IC_{(1-\alpha)100\%}(\mu_1 - \mu_2) = \left(\overline{x}_1 - \overline{x}_2 \mp z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right).$$

Definição 2.9. Intervalo de confiança para a diferença de **médias** com **variâncias desconhecidas (mas iguais)**:

- Variável fulcral:
$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\frac{(m-1)S_1'^2 + (n-1)S_2'^2}{m+n-2}}} \sim t(m+n-2).$$

- Intervalo de confiança:
$$IC_{(1-\alpha)100\%}(\mu_1 - \mu_2) = \left(\overline{x}_1 - \overline{x}_2 \mp t_{\alpha/2}\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\frac{(m-1)s_1'^2 + (n-1)s_2'^2}{m+n-2}}\right).$$

Definição 2.10. Intervalo de confiança para a diferença de **médias** com **variâncias desconhecidas (possivelmente diferentes)**:

- Variável fulcral:
$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1'^2}{m} + \frac{S_2'^2}{n}}} \overset{\text{a}}{\sim} t(r),$$

onde r é o maior número inteiro contido em

$$\frac{\left(\frac{s_1'^2}{m}+\frac{s_2'^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s_1'^2}{m}\right)^2+\frac{1}{n-1}\left(\frac{s_2'^2}{n}\right)^2}.$$

- Intervalo de confiança:

$$IC_{(1-\alpha)100\%}(\mu_1-\mu_2)=\left(\overline{x}_1-\overline{x}_2\mp t_{\alpha/2}\sqrt{\frac{s_1'^2}{m}+\frac{s_2'^2}{n}}\right).$$

Definição 2.11. Intervalo de confiança para a **razão de variâncias** σ_2^2/σ_1^2:

- Variável fulcral:

$$F=\frac{S_1'^2}{S_2'^2}\cdot\frac{\sigma_2^2}{\sigma_1^2}\sim F(m-1,n-1).$$

- Intervalo de confiança:

$$IC_{(1-\alpha)100\%}(\sigma_2^2/\sigma_1^2)=\left(f_1\frac{s_2'^2}{s_1'^2},f_2\frac{s_2'^2}{s_1'^2}\right),\quad P(F<f_1)=P(F>f_2)=\alpha/2.$$

Nota . Pela tabela 8 podemos obter f_2 diretamente, enquanto que f_1 é obtido da seguinte forma:

- ir à tabela 8 e nos valores de $F(n-1,m-1)$ encontrar o valor f que tem uma probabilidade à direita de $\alpha/2$;
- fazer $f_1=\frac{1}{f}$.

Exercício 2.2 (Exercício 7.31 do livro)**.** Considere uma população com distribuição normal de parâmetros desconhecidos, donde foi recolhida uma amostra casual de dimensão 25. Suponha que a amostra forneceu os seguintes resultados:

$$\sum_{i=1}^{25}x_i=75 \text{ e } \sum_{i=1}^{25}x_i^2=321.$$

(a) Construa um intervalo de confiança de 95% para a média.

```
#
```

```
#
```

(b) Construa um intervalo de confiança de 95% para o desvio padrão.

```
#
```

```
#
```

Exercício 2.3 (Exercício 7.30 do livro (adaptado))**.** Com base numa amostra casual de 16 observações, retirada de uma população normal, construiu-se, pelo processo habitual, o seguinte intervalo de confiança para o valor esperado: $(7.398,\ 12.602)$.

(a) Sabendo que, com a informação da amostra, se obteve $s = 3.872$, qual é o grau de confiança que pode atribuir ao intervalo de confiança atrás referido?

```
#
```

```
#
```

(b) Com base na mesma amostra, construa um intervalo de confiança a 95% para a variância da população.

```
#




#
```

2.4 Intervalos de confiança para grandes amostras

Quando n for grande ($n > 30$), recorremos ao TLC para obter intervalos assintóticos. Consideramos que todas as variáveis fulcrais acima mencionadas seguem assintoticamente uma distribuição normal padrão.

Exercício 2.4. Suponha que o gasto anual com bens de consumo dos habitantes das cidades A e B está associado a variáveis aleatórias com distribuição normal de parâmetros desconhecidos. Foram inquiridas 100 pessoas da cidade A e observou-se um gasto médio anual de 5100 unidades monetárias (u.m.) e um desvio padrão corrigido de 1020 u.m. Foram ainda inquiridas 125 pessoas da cidade B tendo-se verificado uma média de 6150 u.m. com um desvio padrão corrigido de 1300 u.m. Recorrendo a um intervalo de confiança a 95%, verifique se é possível afirmar que o gasto médio anual com bens de consumo é maior em alguma das duas cidades.

```
#




#
```

Capítulo 3

Testes de hipóteses paramétricos

3.1 Introdução

Ideia

- Estabelecer uma conjetura sobre aspetos desconhecidos da distribuição da população.
- Verificar se a informação existente na amostra observada $(x_1, \ldots, x_n)$ suporta ou não essa conjetura.

Definição 3.1 (Hipótese estatística)**.** Qualquer conjetura sobre aspetos desconhecidos da distribuição (forma funcional, parâmetros, etc) de X.

Exemplo 3.1. $X \sim exp(\lambda = 2)$?, $X \sim \mathcal{N}(\mu, \sigma^2)$?

Definição 3.2 (Hipótese não paramétrica)**.** A hipótese é feita sobre a distribuição (forma funcional) de X.

Exemplo 3.2. $X \sim F(\cdot)$?, $X \sim N(\cdot, \cdot)$?

Definição 3.3 (Hipótese paramétrica)**.** A hipótese é feita sobre parâmetros da distribuição de X. Neste caso, a distribuição de X é conhecida.

Exemplo 3.3. $X \sim N(\mu = 1, \sigma^2)$?, $X \sim exp(\lambda = 2)$?

Nota . Nesta UC apenas iremos estudar hipóteses paramétricas.

Exercício 3.1 (Exercício 8.1 do livro)**.** Para cada uma das seguintes proposições, indique se é ou não uma hipótese estatística:

(a) $\mu = 3$.

(b) $\overline{x} = 4$.

(c) $P(X < 2.5) = 0.4$.

(d) $2 < \sigma < 3$.

(e) $\overline{X} < 3$;

```
#
```

```
#
```

3.2 Testes de hipóteses

Definição 3.4 (Hipótese nula e hipótese alternativa)**.** Seja $X \sim f(x \mid \theta)$, $\theta \in \Theta$, θ conhecido.

- Hipótese nula (**geralmente corresponde ao que suspeitamos ser verdade**):

$$H_0 : \theta \in \Theta_0.$$

- Hipótese alternativa:

$$H_1 : \theta \in \Theta_1.$$

Qualquer hipótese paramétrica estabelece uma partição $\{\Theta_1, \Theta_2\}$ do espaço paramétrico Θ em Θ_0 e Θ_1:

$$\Theta = \Theta_0 \cup \Theta_1 \qquad \text{e} \qquad \Theta_0 \cap \Theta_1 = \varnothing.$$

Definição 3.5 (Hipótese estatística simples)**.** Quando o subespaço paramétrico contém apenas um elemento.

Definição 3.6 (Hipótese estatística composta)**.** Quando o subespaço paramétrico contém mais do que um elemento.

Exemplo 3.4. Pretende-se aferir se determinada moeda é equilibrada: $X \sim B(1, \theta)$ e $\theta = P(\text{“}sucesso\text{”})$

$$X = \begin{cases} 1, & \text{cara} \\ 0, & \text{coroa} \end{cases}$$

- Espaço parâmetro: $\theta \in \Theta = [0, 1]$.
- Hipótese nula: $H_0 : \ \theta = 0.5$ (é uma hipótese simples, neste caso).
- Hipótese alternativa: $H_1 : \ \theta \neq 0.5$ (é uma hipótese composta, neste caso).

Definição 3.7 (Teste de hipóteses)**.** É uma **regra** que permite especificar um subconjunto do espaço amostra (espaço de resultados da amostra) $W \subset \mathbb{R}^n$ tal que:

- se $(x_1, \ldots, x_n) \in W \Longrightarrow$ rejeita-se H_0;
- se $(x_1, \ldots, x_n) \notin W \Longrightarrow$ não se rejeita H_0.

A decisão é sempre referente a H_0 (rejeita-se H_0 ou não se rejeita H_0).

O teste estatístico introduz uma partição do espaço amostra em duas regiões, W e $\overline{W}$

$$W \cup \overline{W} = \mathbb{R}^n \qquad \text{e} \qquad W \cap \overline{W} = \varnothing,$$

onde W designa a **região de rejeição** (RR) ou **região crítica** (RC).

Definição 3.8 (Estatística de teste)**.** Em alternativa ao acima exposto e, em quase todos os casos práticos de interesse, é costume trabalhar-se um com uma **estatística de teste (ET)**:

$$T = T(X_1, \ldots, X_n) \Longrightarrow t_{obs} = T(x_1, \ldots, x_n).$$

Neste caso, a região de rejeição, W, é definida à custa da estatística de teste:

- se $t_{obs} \in W_T \Longrightarrow$ rejeita-se H_0;
- se $t_{obs} \notin W_T \Longrightarrow$ não se rejeita H_0.

Em resumo, os **componentes de um teste de hipótese estatístico** são:

- Hipótese nula, H_0: mantida, a menos que existam evidências que mostrem o contrário;
- Hipótese alternativa, H_1: adotada se H_0 for rejeitada;
- Estatística de teste, $T = T(X_1, \ldots, X_n)$: com base na qual será feita a regra de decisão;
- Região de rejeição (RR), W_T: a regra de decisão.

Tipos de erro

- O teste de hipóteses é realizado com base numa amostra casual (não temos acesso a toda a população).
- Assim, a decisão de rejeitar, ou não, a hipótese nula pode estar errada!
- Devemos considerar dois tipos de erro:
  - Erro do tipo I (ou erro de 1ª espécie): rejeitar H_0 quando H_0 é verdadeira.
  - Erro do tipo II (ou erro de 2ª espécie): não rejeitar H_0 quando H_0 é falsa.
- Vamos dividir o nosso estudo em 3 casos: hipóteses simples *vs* hipóteses simples, hipóteses simples *vs* hipóteses composta e testes bilaterais.

3.3 Hipóteses simples *vs* hipóteses simples

$$\boldsymbol{H_0 : \theta = \theta_0} \quad \text{versus} \quad \boldsymbol{H_1 : \theta = \theta_1}$$

Decisão/Realidade	H_0 verdadeira	H_0 falsa
Rejeitar H_0	Erro de 1ª espécie $\boldsymbol{\alpha = P(t \in W \mid \theta = \theta_0)}$	Decisão correta $\boldsymbol{\beta = P(t \in W \mid \theta = \theta_1)}$
Não rejeitar H_0	Decisão correta $\boldsymbol{1 - \alpha = P(t \notin W \mid \theta = \theta_0)}$	Erro de 2ª espécie $\boldsymbol{1 - \beta = P(t \notin W \mid \theta = \theta_1)}$

- **Dimensão do teste**: $\alpha = P(\text{rejeitar } H_0 \mid H_0 \text{ verdadeira}) = P(\text{rejeitar } H_0 \mid \theta = \theta_0)$.
- **Potência do teste**: $\beta = P(\text{rejeitar } H_0 \mid H_0 \text{ falsa}) = P(\text{rejeitar } H_0 \mid \theta = \theta_1)$.
- Idealmente: **menor valor de α e maior valor de β.**
- A redução das duas probabilidades de erro (ou de uma delas fixando a outra) só se consegue aumentando a dimensão da amostra (ver exemplo a seguir).
- Ao alterar a RR, obtêm-se outros valores para α e β (ver exemplo a seguir).
- Na impossibilidade de minimizar simultaneamente os 2 tipos de erro, recorremos ao Lema de Neyman-Pearson de forma a obtermos o teste mais potente (este semestre não falaremos da aplicação deste Lema).

Exemplo 3.5.

$$X \sim \mathcal{N}(\mu, \sigma^2 = 4), \qquad H_0 : \mu = 10 \quad vs \quad H_1 : \mu = 14 :$$

n	$RR: \ W = \{(x_1, \ldots, x_n) : \bar{x} > k\}$	$\alpha = P(T \in W_T \mid \mu = 10)$		$1 - \beta = Pr(T \notin W_T \mid \mu = 14)$	
1	$k = 12.5$ $W = \{x_1 : x_1 > 12.5\}$	0.1056		0.2266	
2	$k = 12.5$ $W = \{(x_1, x_2) : \bar{x} > 12.5\}$	0.0384	↓	0.1446	↓
2	$k = 13.5$ $W = \{(x_1, x_2) : \bar{x} > 13.5\}$	0.0068	↓	0.3632	↑

Exercício 3.2 (Exercício 8.5 do livro)**.** A duração em horas, X, de um determinado tipo de componente tem distribuição normal com desvio padrão igual a 50. Para testar:

$$H_0: \ \mu = 250 \ \text{ vs } \ H_1: \mu = 200,$$

a regra de decisão é: rejeitar H_0 se $\overline{x} < 230$.

(a) Se a decisão for tomada com base numa amostra casual de 16 componentes, calcule a dimensão e a potência associadas a este teste.

```
#
```

```
#
```

(b) Qual deverá ser a dimensão mínima da amostra para que a probabilidade de cometer o erro de tipo I seja menor a 0.025?

```
#
```

```
#
```

3.4 Hipótese simples *vs* hipóteses compostas

$$\boldsymbol{H_0: \theta = \theta_0} \quad \text{versus} \quad \boldsymbol{H_1: \theta > \theta_1} \quad \text{ou} \quad \boldsymbol{H_1: \theta < \theta_1}$$

- Nada se altera relativamente ao erro de primeira espécie (a dimensão do teste, α, apenas depende de H_0).

- A probabilidade do erro de 2^a espécie, $1-\beta$, e a potência do teste, β, passam a ser uma **função** de θ.

Definição 3.9 (Função potência)**.** Teste $H_0: \theta = \theta_0$ *vs* $H_1: \theta > \theta_0$, com RR W.

$$\beta(\theta) = P(\text{rej. } H_0 \mid H_0 \text{ falsa}) = P(\text{rej.} H_0 \mid \theta > \theta_0) = P((X_1, \ldots, X_n) \in W \mid \theta),$$

para $\theta \in \Theta_1 = \{\theta : \theta > \theta_0\}$.

Nota: o caso $H_1: \theta < \theta_0$ é semelhante (com as devidas adaptações).

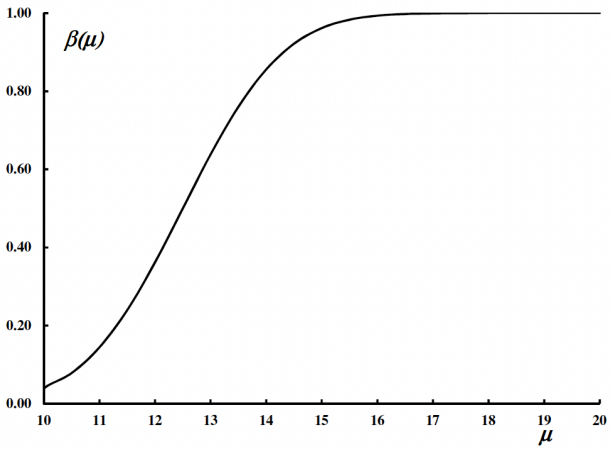


Figura 3.1: Exemplo de uma função potência.

Nota . É importante realçar que:

- Quando $H_0: \theta \leqslant \theta_0$ *vs* $H_1: \theta > \theta_0$ (hip. composta *vs* hip. composta unilateral) devemos proceder como se tivéssemos $H_0: \theta = \theta_0$ *vs* $H_1: \theta > \theta_0$.
- De modo semelhante, para $H_0: \theta \geqslant \theta_0$ *vs* $H_1: \theta < \theta_0$, devemos proceder como se tivéssemos $H_0: \theta = \theta_0$ *vs* $H_1: \theta < \theta_0$.
- Em ambos os casos estamos a escolher o pior cenário.

Exercício 3.3 (Exercício 8.16 do livro)**.** Seja X uma variável aleatória que representa a quantidade de vinho numa garrafa de 75 centilitros. Suponha que X tem distribuição normal com desvio padrão igual a 2. Para testar:

$$H_0: \mu = 75 \text{ vs } H_1: \mu < 75,$$

recolheu-se uma amostra casual de 10 garrafas, rejeitando-se a hipótese nula se $\overline{x} < 74.1$, onde $\overline{x}$ é a quantidade média de vinho por garrafa na amostra observada.

(a) Calcule a dimensão deste teste.

```
#
```

```
#
```

(b) Determine a função de potência e calcule o seu valor quando $\mu = 74$ e $\mu = 72.5$.

```
#



#
```

3.5 Testes bilaterais

$$\boldsymbol{H_0 : \theta = \theta_0} \quad \text{versus} \quad \boldsymbol{H_1 : \theta \neq \theta_0}$$

- Define-se a RR nas duas abas da distribuição da estatística teste, atribuindo-se igual probabilidade $\alpha/2$ a cada uma das 2 sub-regiões.

3.6 O valor-p

- Fixada a dimensão do teste, α, o resultado do teste consiste em rejeitar (ou não rejeitar) H_0.
- Não se tem em conta se o valor da estatística teste se situa longe ou perto do limiar de rejeição.
- O p-value é uma forma **alternativa** de reportar o resultado de um teste que permite ultrapassar esta limitação.

Definição 3.10. Seja $T(x_1, \ldots, x_n) = t_{obs}$ o valor observado de uma estatística de teste. O valor-p ou $p-$value:

- é uma ferramenta para verificar se a estatística de teste está na região de rejeição. É também uma medida da evidência para rejeitar H_0.
- quanto menor for o seu valor, menor será a consistência dos dados com a hipótese ("mais se rejeita" H_0);
- Regra de decisão: $p - value < \alpha \implies$ rejeita-se H_0.

3.7 Populações normais: teste para a média e a variância

Definição 3.11. Teste para a média com variância conhecida

- **Teste bilateral**

$$H_0 : \ \mu = \mu_0 \quad \text{vs} \quad H_1 : \ \mu \neq \mu_0$$

$$ET : \ Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$p - value = 2P(Z \geq |z_{obs}| \mid H_0)$$

$$RR = \{z : |z| > z_{\alpha/2}\} \text{ ou } RR = \left\{\overline{x} : \overline{x} > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} \cup \left\{\overline{x} : \overline{x} < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$

- **Teste unilateral à direita**

$$H_0: \mu = \mu_0 \quad \text{vs} \quad H_1: \mu > \mu_0$$

$$ET: Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$p - value = P(Z \geq z_{obs} \mid H_0)$$

$$RR = \{z : z > z_\alpha\} \text{ ou } RR = \left\{\overline{x} : \overline{x} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right\}$$

- **Teste unilateral à esquerda**

$$H_0: \mu = \mu_0 \quad \text{vs} \quad H_1: \mu < \mu_0$$

$$ET: Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$p - value = P(Z \leq z_{obs} \mid H_0)$$

$$RR = \{z : z < -z_\alpha\} \text{ ou } RR = \left\{\overline{x} : \overline{x} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}\right\}$$

Definição 3.12. Teste para a média com variância desconhecida

- **Teste bilateral**

$$H_0: \mu = \mu_0 \quad \text{vs} \quad H_1: \mu \neq \mu_0$$

$$ET: T = \frac{\overline{X} - \mu_0}{S'/\sqrt{n}} \sim t(n-1)$$

$$p - value = 2P(T \geq |t_{obs}| \mid H_0)$$

$$RR = \{t : |t| > t_{\alpha/2}\} \text{ ou } RR = \left\{\overline{x} : \overline{x} > \mu_0 + t_{\alpha/2} \frac{s'}{\sqrt{n}}\right\} \cup \left\{\overline{x} : \overline{x} < \mu_0 - t_{\alpha/2} \frac{s'}{\sqrt{n}}\right\}$$

- **Teste unilateral à direita**

$$H_0: \mu = \mu_0 \quad \text{vs} \quad H_1: \mu > \mu_0$$

$$ET: T = \frac{\overline{X} - \mu_0}{S'/\sqrt{n}} \sim t(n-1)$$

$$p - value = P(T \geq t_{obs} \mid H_0)$$

$$RR = \{t : t > t_\alpha\} \text{ ou } RR = \left\{\overline{x} : \overline{x} > \mu_0 + t_\alpha \frac{s'}{\sqrt{n}}\right\}$$

- **Teste unilateral à esquerda**

$$H_0: \mu = \mu_0 \quad \text{vs} \quad H_1: \mu < \mu_0$$

$$ET: T = \frac{\overline{X} - \mu_0}{S'/\sqrt{n}} \sim t(n-1)$$

$$p - value = P(T \leq t_{obs} \mid H_0)$$

$$RR = \{t : t < -t_\alpha\} \text{ ou } RR = \left\{\overline{x} : \overline{x} < \mu_0 - t_\alpha \frac{s'}{\sqrt{n}}\right\}$$

Definição 3.13. Teste para a variância

- **Teste bilateral**

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_1: \sigma^2 \neq \sigma_0^2$$

$$ET: Q = \frac{(n-1)S'^2}{\sigma_0^2} \sim \chi^2(n-1)$$

$$p - value = 2\min\{p_1, p_2\} \quad \text{(ver definição seguinte)}$$

$$RR = \{q : q < q_{1-\alpha/2}\} \cup \{q : q > q_{\alpha/2}\} \text{ ou } RR = \left\{s'^2 : s'^2 < \frac{q_{1-\alpha/2}\sigma_0^2}{n-1}\right\} \cup \left\{s'^2 : s'^2 > \frac{q_{\alpha/2}\sigma_0^2}{n-1}\right\}$$

- **Teste unilateral à direita**

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_1: \sigma^2 > \sigma_0^2$$

$$ET: Q = \frac{(n-1)S'^2}{\sigma_0^2} \sim \chi^2(n-1)$$

$$p - value = p_1 = P(Q \geq q_{obs} \mid H_0)$$

$$RR = \{q : q > q_\alpha\} \text{ ou } RR = \left\{s'^2 : s'^2 > \frac{q_\alpha \sigma_0^2}{n-1}\right\}$$

- **Teste unilateral à esquerda**

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_1: \sigma^2 < \sigma_0^2$$

$$ET: Q = \frac{(n-1)S'^2}{\sigma_0^2} \sim \chi^2(n-1)$$

$$p - value = p_2 = P(Q \leq q_{obs} \mid H_0)$$

$$RR = \{q : q < q_{1-\alpha}\} \text{ ou } RR = \left\{s'^2 : s'^2 < \frac{q_{1-\alpha}\sigma_0^2}{n-1}\right\}$$

Considerem-se agora duas populações normais $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ e $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, e duas amostras casuais independentes $(X_1, X_2, \ldots, X_m)$ e $(Y_1, Y_2, \ldots, Y_n)$.

Definição 3.14. Teste à igualdade de médias com variâncias conhecidas

- **Teste bilateral**

$$H_0: \mu_X = \mu_Y \iff \mu_X - \mu_Y = 0 \quad \text{vs} \quad H_1: \mu_X \neq \mu_Y \iff \mu_X - \mu_Y \neq 0$$

$$ET: Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim \mathcal{N}(0,1)$$

$$p - value = 2P(Z \geq |z_{obs}| \mid H_0)$$

$$RR = \{z : |z| > z_{\alpha/2}\} \text{ ou } RR = \left\{ |\overline{x} - \overline{y}| : |\overline{x} - \overline{y}| > z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} \right\}$$

- **Teste unilateral à direita**

$$H_0: \mu_X = \mu_Y \quad \text{vs} \quad H_1: \mu_X > \mu_Y$$

$$ET: Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim \mathcal{N}(0,1)$$

$$p - value = P(Z \geq z_{obs} \mid H_0)$$

$$RR = \{z : z > z_{\alpha}\} \text{ ou } RR = \left\{ \overline{x} - \overline{y} : \overline{x} - \overline{y} > z_{\alpha} \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} \right\}$$

- **Teste unilateral à esquerda**

$$H_0: \mu_X = \mu_Y \quad \text{vs} \quad H_1: \mu_X < \mu_Y$$

$$ET: Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim \mathcal{N}(0,1)$$

$$p - value = P(Z \leq z_{obs} \mid H_0)$$

$$RR = \{z : z < -z_{\alpha}\} \text{ ou } RR = \left\{ \overline{x} - \overline{y} : \overline{x} - \overline{y} < -z_{\alpha} \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} \right\}$$

Definição 3.15. Teste à igualdade de médias com variâncias desconhecidas mais iguais

- **Teste bilateral**

$$H_0:\ \mu_X = \mu_Y \iff \mu_X - \mu_Y = 0 \quad \text{vs} \quad H_1:\ \mu_X \neq \mu_Y \iff \mu_X - \mu_Y \neq 0$$

$$ET:\ T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\frac{(m-1)S_X'^2 + (n-1)S_Y'^2}{m+n-2}}} \sim t(m+n-2)$$

$$p - value = 2P(T \geq |t_{obs}| \mid H_0)$$

$$RR = \{t : |t| > t_{\alpha/2}\} \text{ ou } RR = \left\{ |\overline{x} - \overline{y}| : |\overline{x} - \overline{y}| > t_{\alpha/2}\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\frac{(m-1)s_X'^2 + (n-1)s_Y'^2}{m+n-2}} \right\}$$

- **Teste unilateral à direita**

$$H_0:\ \mu_X = \mu_Y \quad \text{vs} \quad H_1:\ \mu_X > \mu_Y$$

$$ET:\ T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\frac{(m-1)S_X'^2 + (n-1)S_Y'^2}{m+n-2}}} \sim t(m+n-2)$$

$$p - value = P(T \geq t_{obs} \mid H_0)$$

$$RR = \{t : t > t_{\alpha}\} \text{ ou } RR = \left\{ \overline{x} - \overline{y} : \overline{x} - \overline{y} > t_{\alpha}\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\frac{(m-1)s_X'^2 + (n-1)s_Y'^2}{m+n-2}} \right\}$$

- **Teste unilateral à esquerda**

$$H_0:\ \mu_X = \mu_Y \quad \text{vs} \quad H_1:\ \mu_X < \mu_Y$$

$$ET:\ T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\frac{(m-1)S_X'^2 + (n-1)S_Y'^2}{m+n-2}}} \sim t(m+n-2)$$

$$p - value = P(T \leq t_{obs} \mid H_0)$$

$$RR = \{t : t < -t_{\alpha}\} \text{ ou } RR = \left\{ \overline{x} - \overline{y} : \overline{x} - \overline{y} < -t_{\alpha}\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\frac{(m-1)s_X'^2 + (n-1)s_Y'^2}{m+n-2}} \right\}$$

Definição 3.16. Teste à igualdade de médias com variâncias desconhecidas e possivelmente diferentes

- **Teste bilateral**

$$H_0:\ \mu_X = \mu_Y \iff \mu_X - \mu_Y = 0 \quad \text{vs} \quad H_1:\ \mu_X \neq \mu_Y \iff \mu_X - \mu_Y \neq 0$$

$$ET:\ T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X'^2}{m} + \frac{S_Y'^2}{n}}} \sim t(r),$$

onde r é o maior número inteiro contido em

$$\frac{\left(\frac{s'^2_X}{m}+\frac{s'^2_Y}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s'^2_X}{m}\right)^2+\frac{1}{n-1}\left(\frac{s'^2_Y}{n}\right)^2}.$$

$$p-value = 2P(T \geq |t_{obs}| \mid H_0)$$

$$RR = \{t : |t| > t_{\alpha/2}\} \text{ ou } RR = \left\{|\overline{x}-\overline{y}| : |\overline{x}-\overline{y}| > t_{\alpha/2}\sqrt{\frac{s'^2_X}{m}+\frac{s'^2_Y}{n}}\right\}$$

- **Teste unilateral à direita**

$$H_0 : \mu_X = \mu_Y \quad \text{vs} \quad H_1 : \mu_X > \mu_Y$$

$$ET : T = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{S'^2_X}{m}+\frac{S'^2_Y}{n}}} \sim t(r),$$

onde r é o maior número inteiro contido em

$$\frac{\left(\frac{s'^2_X}{m}+\frac{s'^2_Y}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s'^2_X}{m}\right)^2+\frac{1}{n-1}\left(\frac{s'^2_Y}{n}\right)^2}.$$

$$p-value = P(T \geq t_{obs} \mid H_0)$$

$$RR = \{t : t > t_{\alpha}\} \text{ ou } RR = \left\{\overline{x}-\overline{y} : \overline{x}-\overline{y} > t_{\alpha}\sqrt{\frac{s'^2_X}{m}+\frac{s'^2_Y}{n}}\right\}$$

- **Teste unilateral à esquerda**

$$H_0 : \mu_X = \mu_Y \quad \text{vs} \quad H_1 : \mu_X < \mu_Y$$

$$ET : T = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{S'^2_X}{m}+\frac{S'^2_Y}{n}}} \sim t(r),$$

onde r é o maior número inteiro contido em

$$\frac{\left(\frac{s'^2_X}{m}+\frac{s'^2_Y}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s'^2_X}{m}\right)^2+\frac{1}{n-1}\left(\frac{s'^2_Y}{n}\right)^2}.$$

$$p-value = P(T \leq t_{obs} \mid H_0)$$

$$RR = \{t : t < -t_{\alpha}\} \text{ ou } RR = \left\{\overline{x}-\overline{y} : \overline{x}-\overline{y} < -t_{\alpha}\sqrt{\frac{s'^2_X}{m}+\frac{s'^2_Y}{n}}\right\}$$

Definição 3.17. Teste para o quociente de variâncias

- **Teste bilateral**

$$H_0: \sigma_X^2 = \sigma_Y^2 \iff \sigma_X^2/\sigma_Y^2 = 1 \text{ vs } H_1: \sigma_X^2 \neq \sigma_Y^2 \iff \sigma_X^2/\sigma_Y^2 \neq 1$$

$$ET: F = \frac{S_X'^2}{S_Y'^2} \sim F(m-1, n-1)$$

$$p - value = 2\min\{p_1, p_2\} \quad \text{(ver definição seguinte)}$$

$$RR = \left\{F : F < 1/F^*_{\alpha/2}\right\} \cup \left\{F : F > F_{\alpha/2}\right\} \text{ ou}$$
$$RR = \left\{s_X'^2/s_Y'^2 : s_X'^2/s_Y'^2 < 1/F^*_{\alpha/2}\right\} \cup \left\{s_X'^2/s_Y'^2 : s_X'^2/s_Y'^2 > F_{\alpha/2}\right\} \text{ onde}$$

$$F_{\alpha/2} : P(F > F_{\alpha/2}) = \alpha/2,\ F^*_{\alpha/2} : P(1/F > F^*_{\alpha/2}) = \alpha/2$$

- **Teste unilateral à direita**

$$H_0: \sigma_X^2 = \sigma_Y^2 \iff \sigma_X^2/\sigma_Y^2 = 1 \text{ vs } H_1: \sigma_X^2 > \sigma_Y^2 \iff \sigma_X^2/\sigma_Y^2 > 1$$

$$ET: F = \frac{S_X'^2}{S_Y'^2} \sim F(m-1, n-1)$$

$$p - value = p_1 = P(F \geq F_{obs} \mid H_0)$$

$$RR = \{F : F > F_\alpha\} \text{ ou } RR = \left\{s_X'^2/s_Y'^2 : s_X'^2/s_Y'^2 > F_\alpha\right\}$$

- **Teste unilateral à esquerda**

$$H_0: \sigma_X^2 = \sigma_Y^2 \iff \sigma_X^2/\sigma_Y^2 = 1 \text{ vs } H_1: \sigma_X^2 < \sigma_Y^2 \iff \sigma_X^2/\sigma_Y^2 < 1$$

$$ET: F = \frac{S_X'^2}{S_Y'^2} \sim F(m-1, n-1)$$

$$p - value = p_2 = P(F \leq F_{obs} \mid H_0)$$

$$RR = \{F : F < 1/F^*_\alpha\} \text{ ou } RR = \left\{s_X'^2/s_Y'^2 : s_X'^2/s_Y'^2 < 1/F^*_\alpha\right\}$$

Exercício 3.4 (Exercício 8.24 do livro)**.** Um certo produtor de vinho garante às autoridades de fiscalização que o seu vinho tem uma acidez média não superior a 0.5 g/l. Assume-se que o teor de acidez é uma variável aleatória com distribuição normal de parâmetros desconhecidos.

(a) Com base numa amostra de dimensão n, formalize um teste de hipóteses que lhe permita analisar a veracidade do afirmação do produtor.

```
#
```

```
#
```

(b) Observada uma amostra de 20 garrafas, obteve-se média de 0.7 g/l e um desvio-padrão corrigido de 0.08. As autoridades de fiscalização deverão agir contra o produtor? Justifique a sua resposta através de um teste de hipóteses adequado.

```
#
```

```
#
```

Exercício 3.5 (Exercício 8.31 do livro)**.** Uma repartição de finanças tem dois funcionários que recebem declarações de impostos. Suponha que o tempo que cada funcionário leva para atender uma pessoa tem distribuição normal, com desvio padrão igual a 2 minutos. O Sr. Diogo Costa, ao chegar para entregar a sua declaração, verifica que a fila junto ao balcão A tem 20 pessoas, enquanto a fila junto ao balcão B tem 15 pessoas e, naturalmente, opta por esta. Quando começa a ser atendido (uma hora e quinze minutos depois), o Sr. Diogo percebe que a vigésima primeira pessoa da fila ao lado acaba de ser atendida. Pode-se dizer que o tempo médio gasto pelos dois funcionários para atender uma pessoa é idêntico?

```
#
```

```
#
```

Exercício 3.6 (Exercício 8.32 do livro)**.** Para avaliar a qualidade do ambiente nas duas maiores cidades portuguesas são consideradas duas variáveis aleatórias, X e Y, que representam o número de partículas suspensas no ar (microgramas$/m^3$) em Lisboa e no Porto, respetivamente (quanto mais partículas em suspensão, pior a qualidade do ar). Suponha que as duas variáveis aleatórias seguem uma distribuição normal. O Ministério do Ambiente recolheu duas amostras casuais: uma de tamanho 16 em Lisboa e outra de tamanho 13 no Porto. Os resultados observados são os seguintes: $\overline{x} = 92.9$, $s'_X = 25.4$, $\overline{y} = 86.1$, $s'_Y = 28.1$.

(a) Com base num teste de hipóteses adequado, mostre que a igualdade de variâncias das duas variáveis aleatórias não é rejeitada. Considere $\alpha = 0.05$.

```
#
```

```
#
```

(b) Assumindo variâncias iguais em Lisboa e no Porto, determine o p-value do teste adequado para avaliar se a qualidade do ar é pior no centro de Lisboa do que no centro do Porto?

```
#
```

```
#
```

3.8 Grandes amostras

Quando n for grande ($n > 30$), recorremos ao TLC para obter distribuições assintóticas para as estatísticas de teste. Consideramos que todas as ET acima mencionadas seguem assintoticamente uma distribuição normal padrão.

Exercício 3.7 (Exercício 8.39 do livro)**.** Numa empresa de aluguer de viaturas, o principal parâmetro para a definição da tarifa diária de aluguer de viaturas ligeiras de passageiros na categoria de quilometragem ilimitada é o número médio de quilómetros percorridos diariamente, que, à tarifa em vigor, se assume não ser superior a 275. Para avaliar se é necessário rever este tarifário, foi recolhida uma amostra casual de 500 alugueres nesta categoria, tendo-se verificado uma média de 278.5 e uma variância corrigida de 6430.5. Através de um teste de hipóteses adequado e, considerando $\alpha = 0.05$, avalie se é necessário rever a tarifa diária para este tipo de aluguer.

```
#
```

```
#
```

Capítulo 4

Regressão linear

4.1 Introdução

- Greene (Econometric Analysis, 1997) define **Econometria** como o "domínio da Economia que se preocupa com a aplicação da **Estatística Matemática** e dos instrumentos da **Inferência Estatística** à medição empírica das relações postuladas pela teoria económica."
- Ponto de partida: estudar, com base em **dados**, um determinado fenómeno de natureza económica. Exemplos: evolução do consumo das famílias, PIB, dívida, ...

Definição 4.1. Modelo teórico:

- A teoria (bom senso e/ou intuição) leva a construir um modelo teórico que é sempre uma representação abstracta (uma aproximação) da realidade.
- Este modelo estabelece uma relação entre variáveis.
- Os modelos que se vão estudar consistem na análise do comportamento de uma variável dependente, z, como função de outras variáveis $w_1, w_2, \ldots, w_p$ (denominadas variáveis independentes ou explicativas) $z = h(w_1, \ldots, w_p)$, onde a relação geralmente envolve um conjunto de parâmetros $(\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k$.
- É uma função no sentido matemático do termo: a cada valor do argumento corresponde um único valor da função.

Exemplo 4.1. Função consumo, $consumo = f(rendimento)$:

$$consumo = \alpha_1 + \alpha_2 rendimento,$$

onde $0 < \alpha_2 < 1$ é a propensão marginal ao consumo (incremento no consumo de uma pessoa quando há um acréscimo no seu rendimento).

Exemplo 4.2. Função de produção de um bem (Cobb-Douglas), $Q = f(K, L)$, onde K, L são fatores produtivos, por exemplo $K = capital\ investido$, $L = trabalho\ (labour)$:

$$Q = \alpha_1 K^{\alpha_2} L^{\alpha_3},$$

α_2 e α_3 representam elasticidades da quantidade produzida em relação ao capital e ao trabalho, respetivamente.

Nota. Relembrar da Matemática I: a elasticidade de uma variável X face a uma variável Y corresponde à variação percentual que ocorre em X por cada variação percentual unitária (variação de 1pp) que ocorre em Y.

Exemplo 4.3. Função de produção de substituição de elasticidade constante (modelo muito mais complicado):

$$Q = \beta \left[(1-\delta) L^{-\rho} + \delta K^{-\rho} \right]^{-\gamma/\rho},$$

com $\beta, \gamma > 0$, $0 < \delta < 1$.

- Apenas vamos estudar os modelos que envolvem uma relação linear (ou linearizável) em relação aos parâmetros porque abrangem uma variedade significativa de situações e são de tratamento matemático mais fácil.
- Uma relação linear relativamente aos parâmetros $\beta_1, \ldots, \beta_k$ pode ser definida como $y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$, independentemente de a forma funcional ser linear ou linearizável.
- Linearidade relativa aos parâmetros:

$$z = \alpha_1 + \alpha_2 w + \alpha_3 w^2$$

é linear relativamente aos parâmetros α_i, mas não relativamente à variável w.

- Linearidade relativa às variáveis:

$$z = \alpha_1 + \alpha_2 w_2 + \alpha_3^2 w_3$$

é linear relativamente às variáveis w_2, w_3, mas não é linear (nem linearizável) relativamente aos parâmetros α_i.

Exemplo 4.4. Relações lineares ou linearizáveis:

- Linearidade relativa aos parâmetros

$$consumo = \beta_1 + \beta_2 rendimento.$$

- A relação

$$Q = \alpha_1 K^{\alpha_2} L^{\alpha_3}$$

admite linearização através da aplicação da função logarítmo

$$\ln Q = \ln \alpha_1 + \alpha_2 \ln K + \alpha_3 \ln L \iff \ln Q = A + \alpha_2 \ln K + \alpha_3 \ln L,$$

onde $A = \ln \alpha_1$ é uma constante.

- A relação seguinte não é linear nem é linearizável (não estudamos esses casos aqui)

$$Q = \beta \left[(1-\delta) L^{-\rho} + \delta K^{-\rho} \right]^{-\gamma/\rho}.$$

- Iremos estudar modelos definidos como

$$z = h(w_1, w_2, \ldots, w_p)$$

ou, depois de uma possível linearização (assumimos sempre $x_1 = 1$)

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k.$$

Este modelo teórico não é um modelo estatístico. Porquê?

Nota . Antes de converter o modelo acima num modelo estatístico, é importante abordar duas questões:

1. Qual a natureza (aleatória ou determinística) das variáveis envolvidas no modelo? Vamos postular que as variáveis do modelo, assim como as suas observações, são de natureza aleatória.
2. Flexibilidade relacional do modelo teórico: ao considerar

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k,$$

está implicito que os únicos fatores explicativos de y são $x_1, x_2, \ldots, x_k$, o que é geralmente uma hipótese absurda! A flexibilidade obtém-se introduzindo uma variável adicional u que abrange todos os fatores que não foram considerados e que podem afetar o comportamento de y.

- Incorporando u no modelo $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k$ permite-nos obter

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u.$$

Note que u não é observável.

Definição 4.2 (Variável residual)**.** A variável u introduzida acima chama-se **variável residual** e representa tudo o que precisa ser adicionado a $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k$ para se obter y.

Exemplo 4.5. Pretende-se estudar o efeito da escolaridade sobre o salário. É consensual admitir que, para além da escolaridade, outras variáveis influenciam o salário, como por exemplo a experiência profissional, o setor de atividade, etc:

- *salar*: salário mensal médio num determinado ano do trabalhador;
- *educ*: número de anos de escolaridade do trabalhador;
- *exper*: número de anos de experiência profissional após atingir o nível de escolaridade;
- *empc*: número de anos de trabalho no emprego corrente;
- *mulher*: variável binária, assume valor 1 se mulher e 0 se homem.

Para modelo original pode propor-se uma função exponencial:

$$salar = exp(\alpha_1 + \alpha_2 educ + \alpha_3 exper + \alpha_4 empc + \alpha_5 mulher).$$

Linearizando obtém-se

$$\overbrace{\underbrace{\ln(salar)}_{y}}^{lsalar} = \underbrace{\alpha_1}_{\beta_1} + \underbrace{\alpha_2}_{\beta_2}\underbrace{educ}_{x_2} + \underbrace{\alpha_3}_{\beta_3}\underbrace{exper}_{x_3} + \underbrace{\alpha_4}_{\beta_4}\underbrace{empc}_{x_4} + \underbrace{\alpha_5}_{\beta_5}\underbrace{mulher}_{x5}.$$

Introduzindo a variável residual:

$$lsalar = \alpha_1 + \alpha_2 educ + \alpha_3 exper + \alpha_4 empc + \alpha_5 mulher + u.$$

4.2 Tipos de dados económicos

- Para realizar uma análise econométrica, precisamos de dados.
- A econometria, como mencionado anteriormente, desenvolveu-se como uma ferramenta estatística independente, principalmente devido à especificidade dos dados em estudo.
- Em geral, os dados disponíveis são observacionais, ou seja, não experimentais (ex. taxas de juro).
- O tipo de dados disponíveis é uma questão crucial, pois determina tanto os tipos de perguntas que podem realmente ser respondidas quanto os tipos de técnicas que devem ser usadas.
- Existem basicamente três tipos de dados: dados seccionais, séries temporais e dados em painel.

Definição 4.3 (Dados seccionais)**.** Uma amostra de indivíduos (ou empresas, países, etc.) observada num instante específico. Exemplos: dados sobre rendimentos anuais de famílias, dados sobre o desemprego, ...

Definição 4.4 (Séries temporais)**.** Observações sobre um conjunto de variáveis ao longo do tempo para uma unidade estatística (indivíduo, indústria, setor, país, etc.). Exemplo: taxas de juro.

- A frequência de observação dos dados é importante (diária, semanal, mensal, anual).
- Os dados são naturalmente ordenados cronologicamente.
- Naturalmente supõe-se que existe dependência das observações ao longo do tempo, pois o passado em geral é relevante.

Definição 4.5 (Dados em painel)**.** Dados em painel são dados no qual o comportamento de várias entidades/indivíduos é observado ao longo do tempo. Essas entidades podem ser países, empresas, indivíduos, etc.

Nation	Government debt as a percentage of GNP	Unemployment rate
Finland	6.6	2.6
Denmark	5.7	1.6
United States	27.5	5.6
Spain	13.9	3.2
Sweden	15.9	2.7
Belgium	45.0	2.4
Japan	11.2	1.4
New Zealand	44.6	0.5
Ireland	63.8	5.9
Italy	42.5	4.7
Portugal	6.6	2.1
Norway	28.1	1.7
Netherlands	23.6	2.1
Germany	6.7	0.9
Canada	26.9	6.3

Figura 4.1: Exemplo de dados seccionais.

Month	Presidential approval
2002.01	83.7
2002.02	82.0
2002.03	79.8
2002.04	76.2
2002.05	76.3
2002.06	73.4
2002.07	71.6

Figura 4.2: Exemplo de uma série temporal.

country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4
3	2002	9.1	5.2	6.9	2.1

Figura 4.3: Exemplo de dados em painel.

4.3 O modelo de regressão linear

Definição 4.6 (Modelo de regressão linear)**.** O modelo de regressão linear é definido por

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u,$$

onde:

- y: variável dependente ou regressando ou variável explicada;
- x_j, $j = 2, \ldots, k$: variável independente ou regressor ou variável explicativa;
- β_j, $j = 1, \ldots, k$: coeficientes da regressão (constantes) de cada regressor (β_1 chama-se termo independente ou constante);
- u: variável residual (variável não observada).

Atenção:

- y pode ser um regressando mas não ser uma variável explicada: y = variável explicada e $\ln(y)$ = regressando, por exemplo.
- x_j pode ser um regressor mas não ser uma variável explicativa: x_2 = variável explicativa e $\ln(x_2)$ = regressor, por exemplo.

Para estimar os parâmetros (coeficientes de regressão) do modelo teórico, é necessário partir de uma amostra, de dimensão n:

$$\{(y_t, x_{t2}, \ldots, x_{tk}) : t = 1, 2, \ldots, n\}$$

que origina n relações amostrais

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \cdots + \beta_k x_{tk} + u_t,$$

onde u_t é a variável residual associada à observação t das outras variáveis.

As n desigualdades podem apresentar-se utilizando notação matricial:

$$\begin{cases} y_1 &=& \beta_1 + \beta_2 x_{12} + \beta_3 x_{13} + \cdots + \beta_k x_{1k} + u_1 \\ y_2 &=& \beta_1 + \beta_2 x_{22} + \beta_3 x_{23} + \cdots + \beta_k x_{2k} + u_2 \\ \vdots & \vdots & \vdots \\ y_n &=& \beta_1 + \beta_2 x_{n2} + \beta_3 x_{n3} + \cdots + \beta_k x_{nk} + u_n \end{cases} \iff \boldsymbol{Y = X\beta + U.}$$

4.3.1 Hipóteses do modelo de regressão linear

H1 - Linearidade:

$$\boldsymbol{Y = X\beta + U}.$$

- Trataremos apenas de modelos lineares (modelos mais fáceis de tratar).

H2 - Exogeneidade (os regressores são exógenos):

$$E(u_t \mid X) = 0,\ t = 1, 2, \ldots, n.$$

- Nenhuma da informação contida em X pode ser utilizada para calcular $E(u_t)$. Consequências importantes deste pressuposto:
  - O valor esperado não condicionado da variável residual é nulo (ver prop. do valor esperado iterado):

$$E(u_t) = E(E(u_t \mid X)) = E(0) = 0, \quad t = 1, \ldots, n.$$

- $E(y_t \mid X) = E(\beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t \mid X) = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + E(u_t \mid X) =$
  $= \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + 0 = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} \iff E(y_t \mid X) = y_t - u_t \iff$

$$u_t = y_t - E(y_t \mid X)$$

u_t continua a ser desconhecido, pois os parâmetros são desconhecidos.

- Não existe associação linear entre os regressores e a variável residual (**ponto chave do modelo**)

$$Cov(x_{sj}, u_t) = E(x_{sj}u_t) - E(x_{sj})E(u_t) = E(E(x_{sj}u_t \mid X)) = E(x_{sj}E(u_t \mid X)) = E(0) = 0,$$

$t, s = 1, \ldots, n;\ j = 2, 3, \ldots, k.$

H3 - Homocedasticidade condicionada:

$$Var(u_t \mid X) = \sigma^2 > 0,\ t = 1, 2, \ldots, n.$$

- A variância (condicionada) da variável residual é constante $\forall t$, o que implica 2 consequências importantes:
  - A variância **não condicionada** das variáveis residuais é constante:

$$Var(u_t) = Var[E(u_t \mid X)] + E[Var(u_t \mid X)] = Var[0] + E[\sigma^2] = \sigma^2.$$

  - A variância **condicionada** do regressando y_t é constante:

$$Var(y_t \mid X) = Var(\beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t \mid X) = Var(u_t \mid X) = \sigma^2.$$

H4 - Ausência de autocorrelação:

$$Cov(u_t, u_s \mid X) = 0,\ t = 1, 2, \ldots, n,\ t \neq s.$$

- Hipótese importante para modelos com séries temporais:
  - existe uma "ordem" nas observações, o que não acontece nos modelos seccionais;
  - com dados temporais é frequente especificar modelos em que H4 não se verifica, i.e., $Cov(u_t, u_s \mid X) \neq 0$ para $t \neq s$.
- Consequências:
  - H2 + H4 $\Longrightarrow E(u_t u_s \mid X) = Cov(u_t, u_s \mid X) = 0, t \neq s,\ t, s = 1, \ldots, n.$
  - Ausência de correlação (não condicionada)

$$Cov(u_t, u_s) = E[E(u_t u_s \mid X)] - E(u_t)E(u_s) = 0 - 0 = 0.$$

  - As covariâncias condicionadas entre as observações do regressando não dependem das observações dos regressores:

$$Cov(y_t, y_s \mid X) = Cov(u_t, u_s \mid X) = 0, \quad t \neq s.$$

H5 - Não existência de multicolinearidade exata. A caraterística da matriz X é igual a k (número de coeficientes de regressão) e $k < n$:

- Hipótese mais técnica que se destina a garantir que a matriz $X^T X$ admite inversa $(X^T X)^{-1}$, isto é, as colunas da matriz X são linearmente independentes.
- Não existência de multicolinearidade: x_j não é combinação linear dos restantes regressoress $j = 1, \ldots, k$:

$$x_j \neq \gamma_1 + \gamma_2 x_2 + \cdots \gamma_{j-1} x_{j-1} + \gamma_{j+1} x_{j+1} + \cdots + \gamma_k x_k.$$

- Exemplo de multicolinearidade: x_2 : rendimento euros e x_3 : rendimento em milhares de euros $\Longrightarrow x_2 = 1000 x_3$.

H6 - Distribuição normal da variável residual condicionada por X (útil para a inferência estatística no MRL):

$$u_t \mid X \sim \mathcal{N}(0, \sigma^2).$$

Nota . Matriz de variâncias-covariâncias, U:

- As hipóteses H3 e H4 permitem determinar a matriz de variâncias-covariâncias U condicionada por X:

$$Cov(U \mid X) = \begin{bmatrix} Var(u_1 \mid X) & Cov(u_1, u_2 \mid X) & \cdots & Cov(u_1, u_n \mid X) \\ Cov(u_2, u_1 \mid X) & Var(u_2 \mid X) & \cdots & Cov(u_2, u_n \mid X) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(u_n, u_1 \mid X) & Cov(u_n, u_2 \mid X) & \cdots & Var(u_n \mid X) \end{bmatrix}$$

- Sob H3: elementos da diagonal são todos σ^2.
- Sob H4: restantes elementos são nulos.

$$Cov(U \mid X) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

O único parâmetro desconhecido desta matriz é σ^2.

4.4 Método dos mínimos quadrados

- Verificadas as hipóteses H1 a H5, os parâmetros a estimar são :

$$\beta_1, \beta_2, \ldots, \beta_k \text{ e } \sigma^2.$$

- Os parâmetros são estimados através do método dos mínimos quadrados.
- Iremos obter as estimativas (e outras informações) através de *outputs* de software (MS Excel e STATA).

Definição 4.7. Estimador para $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k).$$

Para estimar $\boldsymbol{\beta}$ parte-se de:

- uma amostra com n observações das k variáveis e de y;
- $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_k)$: "aproximação" para $\boldsymbol{\beta}$.

Método dos mínimos quadrados:

Obter o vetor $\boldsymbol{b} = (b_1, b_2, \ldots, b_k)$ que minimiza a soma do quadrado dos resíduos:

$$\begin{aligned} \varphi(\tilde{\beta}) &= \sum_{t=1}^{n} \tilde{u}_t^2 = \sum_{t=1}^{n} (y_t - \boldsymbol{x}_t \cdot \tilde{\boldsymbol{\beta}})^2 = \sum_{t=1}^{n} (\text{valor.obs}_t - \text{valor.aprox}_t)^2 = \sum_{t=1}^{n} (\text{residuo}_t)^2 \\ &= \sum_{t=1}^{n} (y_t - (\tilde{\beta}_1 + \tilde{\beta}_2 x_{t2} + \cdots + \tilde{\beta}_k x_{tk}))^2. \end{aligned}$$

Exemplo 4.6. Modelo com um regressor e uma constante: $y = \beta_1 + \beta_2 x + u$.

$$\tilde{u}_t = y_t - (\tilde{\beta}_1 + \tilde{\beta}_2 x_{t2}) = y_t - x_t \cdot \tilde{\boldsymbol{\beta}},$$

com $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2)$ e $x_t = (1, x_{t2})$.

Como obter b?

- O vetor $\boldsymbol{b}$ minimiza a soma dos resíduos quadrados e dá-nos o modelo ("reta") ajustado. É possível mostrar que:

$$(X^T X)\boldsymbol{b} = X^T Y \quad \underbrace{\Leftrightarrow}_{H5:\ X^T X \text{ é invertível}} \quad \boldsymbol{b} = (X^T X)^{-1} X^T Y.$$

- A escolha de minimizar o quadrado dos resíduos tem por principal consequência a de dar maior peso aos grandes resíduos em detrimento dos pequenos.
- Exemplo (modelo com um regressor e uma constante):

$$b_2 = \frac{n\sum_{t=1}^n x_t y_t \ - \ (\sum_{t=1}^n x_t)\,(\sum_{t=1}^n y_t)}{n\sum_{t=1}^n x_t^2 - (\sum_{t=1}^n x_t)^2}, \qquad b_1 = \bar{y} - b_2 \bar{x}.$$

Definição 4.8 (Função de regressão linear ajustada)**.** Uma vez determinado o estimador (e a estimativa) de mínimos quadrados (MQ) dos coeficientes de regressão

$$\boldsymbol{b} = (X^T X)^{-1} X^T Y = (b_1, \ldots, b_k)$$

obtém-se a **função de regressão linear ajustada** (aos dados):

$$\hat{y}_t = b_1 + b_2 x_{t2} + \cdots + b_k x_{tk}.$$

$\boldsymbol{b} - \boldsymbol{\beta}$ representa o desvio entre o estimador $\boldsymbol{b}$ e o verdadeiro valor do vetor dos coeficientes de regressão, $\boldsymbol{\beta}$:

$$\begin{aligned} b &= (X^T X)^{-1} X^T \underbrace{Y}_{X\beta + U} = (X^T X)^{-1} X^T (X\beta + U) = \beta + (X^T X)^{-1} X^T U \iff \\ \iff b - \beta &= (X^T X)^{-1} X^T U \iff (X^T X)(b - \beta) = X^T U. \end{aligned}$$

Este desvio nunca pode ser determinado de forma exata, porque U não é observável.

Definição 4.9 (Resíduos dos mínimos quadrados)**.** O resíduo dos mínimos quadrados relativo à observação t é dado por:

$$\hat{u}_t = y_t - \hat{y}_t, \qquad t = 1, 2, \ldots, n$$

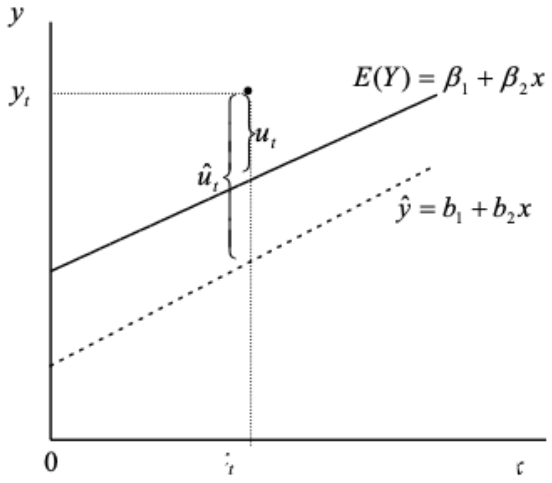
Em notação matricial,

$$\hat{U} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}, \qquad \hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = Xb \qquad \Longrightarrow \hat{U} = Y - Xb = Y - \hat{Y}.$$

Nota . É importante não confundir o seguinte:

- resíduos MQ $\neq$ variáveis resíduais, $\hat{u}_t \neq u_t$.
- função de regressão linear teórica $\neq$ função de regressão linear ajustada:
  - regressão linear teórica: $E(y_t \mid X = x) = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk}$.
  - regressão linear ajustada: $\hat{y}_t = \widehat{E}(y_t \mid X = x) = b_1 + b_2 x_{t2} + \cdots + b_k x_{tk}$.

– o valor b_j representa a estimativa MQ de β_j, $j = 2, \ldots, k$.



4.5 Interpretaçaõs dos parâmetros estimados

As interpretações são feitas em termos do **valor esperado condicionado** de Y

$$E(Y \mid X) = E(y_t \mid x_{t2}, \ldots, x_{tk}) = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk},$$

onde, para cada conjunto $(x_{t2}, \ldots, x_{tk})$ é estimado por

$$\hat{y}_t = b_1 + b_2 x_{t2} + \cdots + b_k x_{tk}.$$

Para exemplificar as situações mais correntes, consideram-se dois exemplos::

1. Variável y é a variável de interesse

$$E(y_t \mid x_{t2}, \ldots, x_{tk}) = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} \Longrightarrow \hat{y}_t = b_1 + b_2 x_{t2} + \cdots + b_k x_{tk}.$$

· β_j: representa a variação marginal, ou seja, a variação em $E(y \mid X)$ quando x_j aumenta uma unidade, tudo o resto mantendo-se constante (*ceteris paribus*), $j = 2, \cdots, k$;

· β_1 : termo independente (em geral, não possui interpretação própria).

2. Variável de interesse z com $y = \ln(z)$, $x_2 = \ln w_2$ e $x_3, \ldots, x_k$ não resultam de transformação.

$$E(\ln z_t \mid w_{t2}, x_{t_3} \ldots, x_{tk}) = \beta_1 + \beta_2 \ln w_{t2} + \beta_3 x_{t3} + \cdots + \beta_k x_{tk}$$

$$\Longleftrightarrow$$

$$E(y_t \mid x_{t2}, \cdots, x_{tk}) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \cdots + \beta_k x_{tk} \Longrightarrow \hat{y}_t = b_1 + b_2 x_{t2} + b_3 x_{t3} + \cdots + b_k x_{tk}.$$

A interpretação dos parâmetros deve ser feita agora de forma mais cuidadosa:

· β_2: representa a variação marginal, ou seja, a variação em $E(\ln z \mid X)$ quando $\ln w_2$ aumenta uma unidade, tudo o resto mantendo-se constante (*ceteris paribus*); esta interpretação não tem interesse (ver ponto seguinte);

· β_2: representa uma elasticidade pontual de z (ou melhor, do valor esperado condicional de z) em relação a w_2, ou seja, quando w_2 varia em 1%, o valor esperado condicional de z irá variar aproximadamente β_2%, *ceteris paribus*;

· β_j, $j = 3, \cdots, k$: representa a semi-elasticidade pontual de z (ou melhor, o valor esperado condicional de z) em relação a x_j. Em termos concretos, quando x_j aumentar uma unidade, o valor esperado condicional de z irá alterar aproximadamente $100\beta_j$%, $j = 3, \cdots, k$, c.p.;

· A variação percentual exata em z é dada por

$$100(e^{b_j \Delta x_j} - 1).$$

Para uma melhor compreensão recomenda-se a leitura atenta do Capítulo 10 do livro de referência, onde estes aspetos são apresentados de forma pormenorizada.

Resumo

escala y	escala x	Interpretação
Y	X_j	$\Delta X_j = 1 \Longrightarrow \Delta E(Y \mid X) = \beta_j$
$\ln(Y)$	X_j	$\Delta X_j = 1 \Longrightarrow \Delta E(Y \mid X) = 100\beta_j\%$
Y	$\ln(X_j)$	$\Delta X_j = 1\% \Longrightarrow \Delta E(Y \mid X) = \beta_j/100$
$\ln(Y)$	$\ln(X_j)$	$\Delta X_j = 1\% \Longrightarrow \Delta E(Y \mid X) = \beta_j\%$

Exemplo 4.7.

$$\widehat{lsalar}_t = 5.81505 + 0.055383 educ_t + 0.022988 exper_t + 0.003953 empc_t$$

Interpretação das estimativas:

- A estimativa MQ da semi-elasticidade do salário (esperado) em relação ao número de anos de escolaridade (retorno da educação) é de 0.0554, isto é, se a escolaridade aumentar um ano, o salário aumenta aproximadamente $100 \times 0.0554\% = 5.54\%$, c.p. (o valor exato é $e^{0.0554} - 1 \approx 5.69\%$).
- Interpretação semelhante é dada aos outros coeficientes;
- Os sinais das três estimativas coincidem com os sinais esperados para os respetivos parâmetros.

4.6 Propriedades dos resíduos de MQ

1. A soma dos resíduos é zero:

$$\sum_{t=1}^{n} \hat{u}_t = 0.$$

2. A soma dos produtos das observações de cada regressor pelos resíduos é zero:

$$\sum_{t=1}^{n} x_{tj} \hat{u}_t = 0, \quad j = 1, \ldots, k.$$

3. A soma dos produtos dos valores ajustados pelos resíduos é igual a zero:

$$\sum_{t=1}^{n} \hat{y}_t \hat{u}_t = 0.$$

4. A soma dos quadrados das observações do regressando é igual à soma dos quadrados dos respetivos valores ajustados mais a soma dos quadrados dos resíduos:

$$\sum_{t=1}^{n} y_t^2 = \sum_{t=1}^{n} \hat{y}_t^2 + \sum_{t=1}^{n} \hat{u}_t^2.$$

5. O estimador MQ de β, designado por b, condicionado ou não por X, é centrado (não enviesado):

$$E(b \mid X) = E(b) = \beta.$$

6. O estimador dos mínimos quadrados b, condicionado ou não por X, é linear em Y:

$$b = (X^T X)^{-1} X^T Y = \phi(Y), \quad \phi \text{ função linear de } Y.$$

7. A matriz de variâncias-covariâncias do estimador dos mínimos quadrados b, condicionado por X, é

$$Cov(b \mid X) = \sigma^2 (X^T X)^{-1},$$

logo,

$$Var(b_j \mid X) = \sigma^2_{b_j} = \sigma^2 m^{jj}, \qquad j = 1, \ldots, k,$$

onde m^{jj} = elemento da diagonal de ordem j da matriz $(X^T X)^{-1}$.

8. O estimador MQ de β é consistente.
9. (Teorema de Gauss-Markov)

Teorema 4.1 (Gauss-Markov)**.** *Qualquer que seja o estimador $\hat{\beta}$ de β, linear e não enviesado, o estimador b condicionado por X é mais eficiente do que $\hat{\beta}$ (tem menor variância). Diz-se que b é **BLUE** (**B**est **L**inear **U**nbiased **E**stimator) para β.*

Esta propriedade pode ser estendida a uma combinação linear dos coeficientes de regressão

$$\delta = c_1\beta_1 + c_2\beta_2 + \cdots + c_k\beta_k,$$

mostrando-se que $\hat{\delta} = c_1 b_1 + \cdots + c_k b_k$ é BLUE para δ.

Definição 4.10 (Estimador não enviesado da variância das variáveis residuais)**.** Agora queremos estimar

$$\sigma^2 = Var(u_t) = E(u_t^2) - E^2(u_t) \underbrace{=}_{E(u_t)=0} E(u_t^2).$$

Não se pode usar $\hat{\sigma}^2 = \dfrac{1}{n}\sum_{t=1}^{n} u_t^2$, já que u_t não é observável. Utiliza-se

$$s^2 = \frac{1}{n-k}\sum_{t=1}^{n} \hat{u}_t^2$$

que é um estimador centrado para σ^2.

Podemos definir o **erro padrão** da regressão por

$$s = \sqrt{s^2}.$$

Estimado σ^2, a matriz de variâncias-covariâncias de b, condicionada por X, pode ser estimada por:

$$\widehat{Cov}(b \mid X) = s^2 (X^T X)^{-1} \quad \text{e, em particular,} \quad \widehat{Var}(b_j \mid X) = s^2 m^{jj} = s^2_{b_j}.$$

O **erro padrão** da estimativa b_j é dado por

$$s_{b_j} = \sqrt{\widehat{Var}(b_j \mid X)} = s\sqrt{m^{jj}}.$$

Estimado o modelo, como avaliar a qualidade do ajustamento?

Definição 4.11 (Coeficiente de determinação)**.** Para avaliar a "qualidade do ajustamento" (da regressão aos dados) pode-se recorrer ao **Coeficiente de determinação** (corresponde ao coeficiente de correlação empírico):

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}, \qquad 0 \leqslant R^2 \leqslant 1,$$

onde

Total **S**um of **S**quares (SST): variação total na variável dependente (VT)

$$SST = VT = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Explained **S**um of **S**quares (SSE): variação (na variável dependente) explicada pela regressão (VE)

$$SSE = VE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

Residual **S**um of **S**quares (SSR): variação (na variável dependente) não explicada pela regressão (VR)

$$SSR = VR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{u}_i{}^2.$$

Nota: $SST = SSE + SSR \iff VT = VE + VR$.

Quanto mais próximo de 1 estiver o coeficiente de determinação, melhor é o "grau de ajustamento''.

Nota . Apenas se deve usar R^2 para comparar modelos que tenham a **mesma variável dependente**.

O coeficiente de determinação R^2 tem dois grandes inconvenientes:

- Ser uma medida sumária de **interpretação** nem sempre fácil: o que é um valor elevado/baixo?
- Quando se acrescenta ao modelo mais um regressor, qualquer que ele seja, o R^2 nunca decresce (para a mesma amostra), pois $\sum_{t=1}^{n} \hat{u}_t^2$ não cresce.

Definição 4.12 (Coeficiente de determinação ajustado)**.** Para contornar as insuficiências da utilização de R^2, utiliza-se o **coeficiente de determinação ajustado** que penaliza a introdução de mais variáveis (regressores):

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

Este coeficiente tem o inconveniente de poder ser negativo e de nunca atingir o valor 1.

Apenas se deve usar para comparar modelos que tenham a mesma variável dependente.

4.7 Inferência estatística no modelo de regressão linear

A partir de uma amostra, podemos estar interessados em:

- Estimar os coeficientes e as suas variâncias;
- Construir intervalos de confiança para os coeficientes;
- Testar hipóteses sobre os parâmetros.

O objetivo agora é fazer inferência estatística sobre os parâmetros β_j do modelo.

Exemplo: $\ln(sal) = \beta_0 + \beta_1 educ + u$.

Será que $\beta_1 = 0$? Ou $\beta_1 < 0$? Ou ainda $\beta_1 = 1$?

É necessário encontrar uma estatística de teste com uma distribuição totalmente conhecida.

Nota: relembre o capítulo relativo aos testes de hipóteses!

Inferência estatística sobre a variância das variáveis residuais

Interesse direto reduzido, mas importante em Econometria...

Hipóteses:

$$H_0 : \sigma^2 = \sigma_0^2 \quad vs \quad H_1 : \sigma^2 \neq \sigma_0^2 \quad (ou \;\; H_1 : \sigma^2 < \sigma_0^2, \; ou \;\; H_1 : \sigma^2 > \sigma_0^2).$$

ET:

$$Q = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2(n-k), \quad \text{com} \quad s^2 = \frac{1}{n-k}\sum_{i=1}^{n} \hat{u}_i^2.$$

Inferência estatística sobre um coeficiente de regressão

Hipóteses:

$$H_0 : \beta_j = \beta_{0j} \quad vs \quad H_1 : \beta_j \neq \beta_{0j} \quad (ou \;\; H_1 : \beta_j < \beta_{0j}, \; ou \;\; H_1 : \beta_j > \beta_{0j}).$$

ET:

$$T_j = \frac{b_j - \beta_j}{s_{b_j}} = \frac{b_j - \beta_j}{s\sqrt{m^{jj}}} \sim t(n-k).$$

Caso particular (e muito importante): $\beta_{0j} = 0$ (significância estatística do regressor).

Nota: no livro principal aparece t_j de modo a aliviar a notação. Na regressão podemos usar t_j ou T_j.

Inferência estatística sobre um coeficiente de regressão (sinal do regressor)

Hipótese:

$$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j > 0 \; (ou \;\; < 0).$$

ET:

$$T_j = \frac{b_j}{s_{b_j}} \sim t(n-k).$$

Inferência estatística sobre uma combinação linear dos coeficientes de regressão

O objetivo agora é testar

$$\delta = c_1\beta_1 + c_2\beta_2 + \cdots + c_k\beta_k = c\beta^T.$$

O estimador de MQ de δ é

$$\hat{\delta} = c_1 b_1 + c_2 b_2 + \cdots + c_k b_k = c b^T.$$

Mostra-se que

$$t_{\hat{\delta}} = \frac{\hat{\delta} - \delta}{s_{\hat{\delta}}} \sim t(n-k).$$

$s_{\hat{\delta}} = s\sqrt{c(X^TX)^{-1}c^T}$ é o erro padrão de $\hat{\delta}$ e pode ser obtido da matriz de covariâncias de b.

Adiante se verá uma solução prática quando não se tem a matriz de covariâncias de b (no MSExcel não é fácil obter essa matriz).

Teste de nulidade conjunta de coeficientes de regressão (nulidade de um subconjunto)

Averiguar se alguns dos coeficientes de regressão são conjuntamente iguais a zero.

Suponha-se que queremos testar se os últimos $m = k - p$ coeficientes são iguais a zero:

Hipóteses:

$$H_0 : \beta_{p+1} = 0, \beta_{p+2} = 0, \ldots, \beta_k = 0$$

$$vs$$

$$H_1 : \exists\, \beta_j \neq 0, \quad j = p+1, \ldots, k.$$

Concebe-se um teste em 3 passos:

Passo 1 - Estimar o modelo **sem restrições**, i.e., com todos os regressores, e obter

$$SSR_1 = VR_1 = \sum_{t=1}^{n} \hat{u}_t^2.$$

Passo 2 - Estimar o modelo **com restrições**, i.e., eliminando os regressores que se admite terem coeficiente nulo:

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_p x_{tp} + \underbrace{\beta_{p+1}}_{=0} x_{tp+1} + \cdots + \underbrace{\beta_k}_{=0} x_{tk} \Longrightarrow y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_p x_{tp},$$

e obter

$$SSR_0 = VR_0 = \sum_{t=1}^{n} \hat{u}_t^2.$$

Passo 3 - Comparar os modelos através da estatística

$$F = \frac{(SSR_0 - SSR_1)/m}{SSR_1/(n-k)} = \frac{SSR_0 - SSR_1}{ms^2} \sim F(m, n-k),$$

onde $m = k - p$ representa o número de restrições (número de coeficientes nulos).

Adicionalmente, $s^2 = \dfrac{SSR_1}{n-k}$ representa uma estimativa de σ^2 com base no modelo sem restrições.

Nota:

Em alternativa à utilização de SSR_0 e SSR_1, podemos recorrer aos respetivos coeficientes de determinação:

$$F = \frac{(R^2 - R_0^2)/m}{(1 - R^2)/(n-k)} \sim F(m, n-k).$$

Se o teste individual de cada um dos coeficientes incluídos em H_0 não rejeita a nulidade e o teste conjunto rejeita, desconfiar de possível multicolinearidade.

Teste de significância global da regressão

Testa-se a nulidade de todos os coeficientes com exceção do termo independente:

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0 \quad \text{vs} \quad H_1 : \exists \beta_j \neq 0, \quad j = 2, \ldots, k.$$

Não rejeitar a hipótese nula corresponde a afirmar que o modelo proposto não é globalmente adequado para descrever o comportamento da regressão.

A estatística do teste obtém-se pelo sistema anterior, sendo agora o número de restrições $m = k - 1$:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} = \frac{SSE/(k-1)}{SSR/(n-k)} \sim F(k-1, n-k).$$

Nota final

Quando a variável residual não tem distribuição normal (violação da hipótese H6), mas a amostra é grande, o TLC pode ser aplicado:

$$T_j = \frac{b_j - \beta_j}{s_{b_j}} \overset{a}{\sim} \mathcal{N}(0,1).$$

Exercício 4.1 (Exercício 10.4 do livro (adaptado)). Num estudo sobre colesterol, recolheu-se uma amostra aleatória de 30 pacientes, tendo-se observado as seguintes variáveis:

- $x =$ número de gramas de gordura consumida por dia;
- $y =$ quantidade de colesterol no sangue, em miligramas por decilitro.

Os resultados da estimação, obtidos com o MSExcel, são os seguintes:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,586050325
R Square	0,343454983 A
Adjusted R Square	0,320006947 B
Standard Error	39,39128311 C
Observations	30 D

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1 E	22728,12448 H	22728,12448 K	14,64749452 M	0,00066655 N
Residual	28 F	43446,84918 I	1551,673185 L		
Total	29 G	66174,97367 J			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	91,5885625 O	30,5126691 Q	3,00165686 S	0,005594418 U	29,08619321 W	154,0909318 Y
x (fat)	1,167693221 P	0,305103428 R	3,827204531 T	0,00066655 V	0,542717181 X	1,792669262 Z

(a) Descreva o significado de todas as letras de cor laranja.

```
#
```

```
#
```

(b) Escreva os modelos teórico e estimado.

```
#
#
```

(c) Interprete o valor da letra P.

```
#
#
```

(d) Teste a hipótese do declive da recta ser igual a 1 contra a alternativa de ser superior a 1.

```
#
#
```

(e) Construa um intervalo de confiança a 90% para o declive da recta de regressão.

```
#
#
```

(f) A partir do intervalo construído na alínea anterior, o que pode concluir quanto à relação entre colesterol e gordura ingerida?

```
#
```

```
#
```

Exercício 4.2 (Exercício 10.19 do livro)**.** Considere o seguinte modelo de regressão (a verificar as hipóteses básicas):

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + u_t, \quad t = 1, 2, \cdots, 500,$$

onde y é o preço (em unidades monetárias) por m^2 de um apartamento em determinada cidade, x_2 é a área do apartamento, e x_3 é a distância ao centro da cidade em quilómetros. Estimado o modelo com o software SPSS, obteve-se:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.479 (a)	.229	.226	478.86068

(a) Predictors: (Constant), DCC, Area

ANOVA (b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33858983.423	2	16929491.711	73.829	.000 (a)
	Residual	113965850.759	497	229307.547		
	Total	147824834.182	499			

(a) Predictors: (Constant), DCC, Area
(b) Dependent Variable: Preçom2

Coefficients (a)

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	2241.934	72.425		30.955	.000
	Area	-2.503	.211	-.467	-11.847	.000
	DCC	-18.947	7.281	-.102	-2.602	.010

(a) Dependent Variable: Preçom2

Nota: a coluna "Coeficientes padronizados" não é importante nesta fase.

(a) Analise sumariamente os resultados obtidos em termos estatísticos (considere $\alpha = 0.05$) e práticos ("económicos"), interpretando as estimativas obtidas para os coeficientes da regressão.

```
#
```

```
#
```

(b) Teste

$$H_0 : \beta_2 = -2 \text{ vs } H_1 : \beta_2 < -2.$$

No caso de não rejeitar H_0, seria razoável excluir o regressor x_2 do modelo?

```
#
```

```
#
```

Exercício 4.3 (Exercício 10.10 do livro)**.** A empresa ELECTRIK pretende construir um modelo explicativo do consumo familiar (em unidades monetárias) de energia eléctrica, y, em função do rendimento familiar, x_2, do número de indivíduos em cada família, x_3, e da área do fogo respectivo em metros quadrados, x_4. Os resultados da estimação com o EXCEL encontram-se a seguir:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.932985993
R Square	0.870462864
Adjusted R Square	0.805694295
Standard Error	4.571741129
Observations	10

ANOVA

	df	*SS*	*MS*	*F*
Regression	3	842.6950983	280.8983661	13.43959
Residual	6	125.4049017	20.90081695	
Total	9	968,1		

	Coefficients	*Standard Error*	*t Stat*	*p-value*
Intercept	12.47998124	9.282257233	1.344498534	0.227386
x_2	0.060527128	0.024073045	2.514311292	0.045637
x_3	-2.81451648	2.679535456	-1.050374786	0.334
x_4	0.020535759	0.057861717	0.354910983	0.734798

(a) No seu conjunto, considera que os regressores incluídos neste modelo são úteis para a explicação do consumo? Teste a 0.05.

```
#
```

```
#
```

(b) Observando apenas o valor-p (p-value) no quadro, diga qual ou quais dos regressores parecem ser significativos.

```
#
#
```

(c) Critique o modelo adoptado tendo em conta o número de observações.

```
#
#
```

(d) Teste a 0.05 a hipótese de o aumento de rendimento implicar, em média, um aumento de consumo.

```
#
#
```

(e) A ELECTRIK resolveu estimar um outro modelo em que apenas incluiu o rendimento familiar como regressor. Os resultados da estimação encontram-se a seguir:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.90358399
R Square	0.816464028
Adjusted R Square	0.793522031
Standard Error	4.712764247
Observations	10

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	790.4188252	790.4188252	35.58819
Residual	8	177.6811748	22.21014685	
Total	9	968.1		

	Coefficients	*Standard Error*	*t Stat*	*p-value*
Intercept	11.02028658	2.963330537	3.718885369	0.005881
Rendimento	0.050015429	0.008383996	5.965583414	0.000336

Teste a 0.05 a nulidade simultânea dos coeficientes respeitantes às variáveis "número de indivíduos" e "área do fogo familiar". Qual dos dois modelos parece preferível?

```
#




#
```

Exercício 4.4 (Exercício 10.27 do livro)**.** Um economista de uma associação de produtores de vinho decidiu construir um modelo explicativo do preço (em euros por garrafa) dos vinhos de determinada região demarcada, PR. Para tal seleccionou como variáveis explicativas a classificação, de 1 a 10 pontos, dada pela revista "Espírito do Vinho" (CL), a antiguidade da colheita em anos (ID), a quantidade produzida em milhares de garrafas (QT) e uma variável (TN), que assume o valor 1 se o vinho tem maioritariamente uvas da casta "Touriga Nacional" e 0 caso contrário, e propôs a seguinte especificação:

$$LPR = \beta_1 + \beta_2 CL + \beta_3 ID + \beta_4 LQT + \beta_5 TN + u,$$

onde LPR representa logaritmo natural do preço da garrafa, LQT é o logaritmo natural da quantidade produzida. Supondo satisfeitas as hipóteses do modelo de regressão linear múltipla, e observada uma amostra aleatória de 65 produtores (uma garrafa por produtor), estimou-se o modelo tendo-se obtido os seguintes resultados:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.83421390
R Square	0.69591283
Adjusted R Square	0.67564035
Standard Error	0.18470493
Observations	65

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	4.68451870	1.17112967	34.32796005	6.75987E-15
Residual	60	2.04695474	0.03411591		
Total	64	6.73147343			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2.1218282	0.44818010	4.7343204	1.3843E-05
CL	0.1041714	0.02042785	5.0994816	3.6707E-06
ID	0.0503412	0.01992331	2.5267481	0.01417673
LQT	-0.1875930	0.09870456	-1.9005501	0.06216823
TN	0.4510302	0.04717505	9.5607791	1.1381E-13

$$\hat{Cov}(b \mid X) = \begin{bmatrix} 0.200870 & & & & \\ -0.001640 & 0.000417 & & & \\ -0.003265 & 0.000064 & 0.000397 & & \\ -0.040315 & -0.000214 & 0.000064 & 0.009743 & \\ -0.005002 & -0.000085 & 0.000119 & 0.000855 & 0.002225 \end{bmatrix}$$

Utilize uma dimensão 5% para todos os testes que tiver de realizar:

(a) Tanto no seu conjunto como individualmente, considera que os regressores incluídos neste modelo são úteis na explicação do logaritmo do preço de venda do vinho dessa região demarcada?

```
#
```

```
#
```

(b) Interprete as estimativas obtidas para os coeficientes β_2 e β_4.

```
#
```

```
#
```

(c) Construa um intervalo de confiança a 95% para β_3. Interprete.

```
#
```

```
#
```

(d) Comente, justificando, a seguinte frase: "Para uma dimensão de 5% existe evidência estatística de que em média, e em iguais circunstâncias relativamente às restantes variáveis explicativas, quanto maior for a quantidade produzida mais baixo o preço do vinho."

```
#
```

```
#
```

(e) Teste

$$H_0: \ \beta_2 = 2\beta_3 \ \text{ vs } \ H_1: \ \beta_2 \neq 2\beta_3.$$

```
#
```

```
#
```

Capítulo 5

Complementos ao modelo de regressão linear

5.1 Variáveis artificiais

- Quando uma (ou mais) das variáveis explicativas é de natureza qualitativa (nominal ou ordinal), a sua representação direta por uma variável quantitativa não é possível. Recorre-se então às **variáveis artificiais (*dummy*)**.

Exemplo 5.1. · grau de escolaridade: básico/secundário/superior (ordinal).

· género: masculino/feminino (nominal).

· distrito de residência: Aveiro/Braga/Coimbra/Porto/... (nominal)

Modelo de partida:

Por questões de simplicidade, consideramos o modelo com apenas um regressor quantitativo

$$y_t = \beta_1 + \beta_2 x_t + u_t.$$

Contudo, o que se vai dizer é válido para situações mais gerais, em que o modelo tem mais do que um regressor.

Um fator quantitativo com 2 modalidades:

- d_t tem duas modalidades: A e B (exemplo: Feminino e Masculino);
- introduz-se uma **variável binária** d tal que

$$d = \begin{cases} 1, & \text{se } A \text{ se verifica} \\ 0, & \text{se } A \text{ não se verifica} \end{cases}$$

- Não se devem introduzir 2 variáveis artificiais, para não originar multicolinearidade perfeita (*dummy trap*).
- Se fosse introduzida

$$d^B = \begin{cases} 1, & \text{se } B \text{ se verifica} \\ 0, & \text{se } B \text{ não se verifica} \end{cases}$$

então $d^B = 1 - d$ (multicolinearidade, i.e, a matriz $X^T X$ não é invertível).

- Nota: a variável artificial pode ter mais do que dois fatores. Ex: Trimestre 1, Trimestre 2, Trimestre 3 e Trimestre 4.
- A variável d designa-se **variável artificial (*dummy*)**.
- A escolha dos valores 0 e 1 é arbitrária, mas tem consequências na interpretação dos parâmetros.

- Usualmente o nome da variável binária corresponde à situação a que se atribui o valor 1. Exemplo:

$$Feminino = \begin{cases} 1, & \text{se Feminino} \\ 0, & \text{se Homem} \end{cases}$$

- A variável qualitativa (artificial) pode ter efeito no termo independente ou no declive ou em ambos. vamos estudar esses 3 casos.

5.1.1 Efeito apenas no termo independente

- Introduzir a variável d como um regressor

$$y_t = \beta_1 + \delta d_t + \beta_2 x_t + u_t$$

onde

$$d = \begin{cases} 1, & \text{se } A \text{ se verifica} \\ 0, & \text{se } A \text{ não se verifica} \end{cases}$$

logo

$$y_t = \begin{cases} (\beta_1 + \delta) + \beta_2 x_t + u_t, & \text{se } d_t = 1 \\ \beta_1 + \beta_2 x_t + u_t, & \text{se } d_t = 0 \end{cases}$$

- Em termos de estimação, nada se altera: estima-se β_1, δ e β_2 a partir de $y_t = \beta_1 + \delta d_t + \beta_2 x_t + u_t$.
- Cuidado com a interpretação a dar agora aos parâmetros β_1 e δ. Em termos teóricos:

$$\delta = E(y_t \mid d_t = 1, x_1, \ldots, x_n) - E(y_t \mid d_t = 0, x_1, \ldots, x_n),$$

isto é, δ corresponde ao incremento no termo independente quando se passa de uma observação com $d_t = 0$ para a situação com $d_t = 1$.

- Acontecimento padrão (ou acontecimento referência): aquele que corresponde ao valor 0 da variável artificial e que se inclui em β_1.

5.1.2 Efeito no coeficiente de um regressor quantitativo

- Introduzir uma interação entre a variável d e o regressor quantitativo

$$y_t = \beta_1 + \beta_2 x_t + \delta d_t x_t + u_t = \beta_1 + (\beta_2 + \delta d_t) x_t + u_t,$$

onde $d_t x_t$ representa a **interação** entre o regressor quantitativo e a variável artificial:

$$d_t x_t = \begin{cases} x_t, & \text{se } d_t = 1 \\ 0, & \text{se } d_t = 0 \end{cases}$$

- Para todos os efeitos, $d_t x_t$ comporta-se como uma nova variável (regressor). Logo

$$y_t = \begin{cases} \beta_1 + (\beta_2 + \delta) x_t + u_t, & \text{se } d_t = 1 \\ \beta_1 + \beta_2 x_t + u_t, & \text{se } d_t = 0 \end{cases}$$

- Estimam-se β_1, δ e β_2 a partir de $y_t = \beta_1 + \beta_2 x_t + \delta d_t x_t + u_t$.
- O modelo pode agora escrever-se como

$$y_t = \beta_1 + \beta_2 x_t + \delta d_t x_t + u_t = \beta_1 + (\beta_2 + \delta d_t) x_t + u_t.$$

- Interpretação dos coeficientes:
  - β_2: efeito marginal de x_t sobre y_t, quando $d_t = 0$.
  - $\beta_2 + \delta$: efeito marginal de x_t sobre y_t, quando $d_t = 1$.
  - δ: traduz a diferença dos 2 efeitos $d_t = 0$ e $d_t = 1$:

$$\delta x_t = E(y_t \mid d_t = 1, x_1, \ldots, x_n) - E(y_t \mid d_t = 0, x_1, \ldots, x_n).$$

  A variação do valor esperado de y_t não é medida por δ (constante), mas por δx_t (depende de x_t).

5.1.3 Efeito no termo independente e no coeficiente de um regressor quantitativo

- Juntam-se os 2 efeitos:

$$y_t = \beta_1 + \delta_1 d_t + \beta_2 x_t + \delta_2 d_t x_t + u_t$$

$$y_t = \begin{cases} (\beta_1 + \delta_1) + (\beta_2 + \delta_2) x_t + u_t, & \text{se } d_t = 1 \\ \beta_1 + \beta_2 x_t + u_t, & \text{se } d_t = 0 \end{cases}$$

Notas finais:

· Quando o fator qualitativo tem mais do que 2 modalidades (ex: 4 trimestres T1, T2, T3 e T4), consideram-se tantas variáveis binárias quantas as modalidades do fator menos uma.

· Se uma variável qualitativa tem demasiadas modalidades (ex: posição da empresa no ranking das 1000 maiores empresas portuguesas) não se podem definir as variáveis artificiais necessárias.

· Nestas situações devem agrupar-se modalidades por classes. Por exemplo, definir 5 classes de acordo com as classificações no ranking: 1 a 10, 11 a 50, 51 a 200, 201 a 500, 501 a 1000.

· Quando existem vários fatores qualitativos generaliza-se a abordagem que se acabou de ver: definir as variáveis artificiais necessárias a cada fator e inclui-las no modelo (efeitos no termo independente e/ou coeficientes).

Exercício 5.1 (Exercício 11.1 do livro)**.** Considere o seguinte modelo de regressão linear para explicar o comportamento do consumo de leite das crianças de determinado país que frequentam o ensino básico:

$$y_t = \beta_1 + \beta_2 x_{t2} + u_t,$$

onde y_t representa o consumo de leite pela criança t, e x_{t2} o rendimento *per capita* da família a que pertence a criança t.

(a) Reespecifique o modelo de forma a incluir os efeitos do sexo e do local de residência (campo ou cidade) sobre o termo independente.

```
#



#
```

(b) Qual é o consumo autónomo de uma criança do sexo masculino que resida na cidade?

```
#



#
```

5.2 Testes de alteração de estrutura

- Suponha-se que num modelo de regressão linear é possível dividir as observações das variáveis em grupos de forma a que seja possível estimar **separadamente** os coeficientes do modelo para **cada um dos grupos**:

$$\left.\begin{array}{ll} \text{grupo 1:} & n_1 \text{ observações} \\ \text{grupo 2:} & n_2 \text{ observações} \end{array}\right\} n = n_1 + n_2$$

- Para cada grupo, considera-se um modelo diferente

$$\left\{\begin{array}{rcll} y_t & = & \beta_{11} + \beta_{21}x_{t2} + \cdots + \beta_{k1}x_{tk} + u_t, & t = 1, \ldots, n_1 \\ y_t & = & \beta_{12} + \beta_{22}x_{t2} + \cdots + \beta_{k2}x_{tk} + u_t, & t = n_1 + 1, \ldots, n \end{array}\right.$$

com β_{ji} tal que j : regressor ($j = 1, \ldots, k$) e i : grupo ($i = 1, 2$).

- Podem-se juntar os 2 modelos introduzindo a **dummy**

$$d_t = \left\{\begin{array}{ll} 0, & t = 1, 2, \ldots, n_1 \\ 1, & t = n_1 + 1, \ldots, n \end{array}\right.$$

e utiliza-se o modelo

$$y_t = \beta_1 + \delta_1 d_t + \beta_2 x_{t2} + \delta_2 d_t x_{t2} + \cdots + \beta_k x_{tk} + \delta_k d_t x_{tk} + u_t$$

- Tem-se que $\delta_j = \beta_{j2} - \beta_{j1}$, para $j = 1, \ldots, k$. Ex: $\delta_2 = \beta_{22} - \beta_{21}$.

Se não recorrermos à introdução de variáveis artificiais podemos aplicar o teste de Chow:

Definição 5.1 (Teste de Chow (sem recurso às variáveis artificiais))**.** Supondo $n_1 > k$ e $n_2 > k$, o teste consiste em:

- Hipóteses:

$$H_0 : \beta_{1j} = \beta_{2j}, \quad \text{permanência de estrutura (não existe alteração na estrutura)}$$

$$\text{vs}$$

$$H_1 : \exists j \; \beta_{1j} \neq \beta_{2j}, \quad j = 1, 2, \ldots, k, \;\; \text{existe alteração de estrutura}$$

- Estatística de teste:
  - Estima-se o modelo com as n observações (modelo restrito) e calcula-se $VR_0 = \sum_{t=1}^{n} \hat{u}_t^2$.
  - Estima-se o modelo com n_1 observações do grupo $d_t = 0$ e calcula-se $\sum_{t=1}^{n_1} \hat{u}_t^2$.
  - Estima-se o modelo com n_2 observações do grupo $d_t = 1$ e calcula-se $\sum_{t=n_1+1}^{n} \hat{u}_t^2$.
  - É possível mostrar que $VR_1 = \sum_{t=1}^{n_1} \hat{u}_t^2 + \sum_{t=n_1+1}^{n} \hat{u}_t^2$.

A estatística de teste é

$$F_{Chow} = \frac{(VR_0 - VR_1)/k}{VR_1/(n-2k)} \sim F(k, n-2k).$$

- Região de rejeição: aba direita.
- Não rejeitar significa que não há alteração de estrutura, ou seja, não é necessário "dividir" o modelo!

Recorrendo à introdução de variáveis artificiais podemos aplicar um teste mais geral:

Definição 5.2 (Testes de alteração de estrutura (com recurso às variáveis artificiais))**.** O teste consiste em:

- Hipóteses:

$$H_0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0, \quad \text{permanência de estrutura (não existe alteração na estrutura)}$$

vs

$$H_1 : \exists j \; \delta_j \neq 0, \quad \text{existe alteração de estrutura}$$

É um teste habitual de nulidade de um subconjunto de parâmetros, comparando o modelo sem restrições com o modelo com as restrições dadas por H_0.

- Estatística de teste:

$$F = \frac{(VR_0 - VR_1)/k}{VR_1/(n-2k)} \sim F(k, n-2k).$$

- Região de rejeição: aba direita.
- Não rejeitar significa que não há alteração de estrutura!

Exercício 5.2 (Exercício 11.22 do livro (adaptado)).

Para tentar explicar a classificação obtida pelos seus alunos, um docente de uma disciplina de Matemática do primeiro ano de um curso universitário de Lisboa resolveu estimar um modelo em que utiliza como regressando a classificação obtida (y), e como regressores a nota média (numa escala de 0 a 20 valores) em Matemática nos 3 anos do ensino secundário (x_2), o número de horas dedicado ao estudo da disciplina ao longo do semestre (x_3) e a percentagem de aulas frequentadas (x_4). Os resultados relativos a 54 alunos escolhidos ao acaso foram

$$\begin{aligned} \hat{y}_t &= 2.339 + 0.1212\, x_{t2} + 0.05329\, x_{t3} + 3.9664\, x_{t4}, \\ &\qquad (0.03776) \quad (0.00622) \quad (0.8226) \\ R^2 &= 0.90935\,, \; \sum \hat{u}_t^2 = 35.0649\,. \end{aligned}$$

Suponha que as hipóteses básicas do modelo de regressão linear se encontram satisfeitas.

(a) Teste a adequação estatística global do modelo.

```
#
```

```
#
```

(b) O docente estimou ainda duas regressões repartindo a amostra em dois grupos: 41 alunos não repetentes e 13 alunos repetentes. As respectivas somas dos quadrados dos resíduos foram 27.5074 e 1.5138. Explique o objectivo destas regressões e comente o resultado obtido.

```
#
```

#

(c)

O docente introduziu então uma variável, d_t, que assume o valor 1 para os alunos residentes fora da área urbana de Lisboa, e 0 no caso contrário. Estimada nova regressão, obteve-se

$$\hat{y}_t = 2.818 + 0.1158\, x_{t2} + 0.04970\, x_{t3} + 3.838\, x_{t4} - 0.45\, d_t,$$
$$\qquad\qquad (0.03757) \quad (0.00665) \quad (0.8193) \quad (0.3162)$$

$$R^2 = 0.9129\,,\ \sum \hat{u}_t^2 = 33.67\,.$$

Que efeito pretende o docente captar com a introdução da variável d_t? Que conclusões pode tirar? Obtenha a soma dos resíduos para os alunos residentes fora da área urbana de Lisboa.

#

#

5.3 Previsão

- O objetivo da previsão é estimar os valores previstos para a variável y_t, para valores fixos das variáveis explicativas. A previsão pode ser em **média**, isto é, prever $E(y_t \mid \cdots)$ ou pontual, isto é, prever y_0 dados valores particulares das variáveis explicativas.
- Considere-se o modelo

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \cdots + \beta_k x_{tk} + u_t,$$

e designe-se os valores para os quais se quer efetuar a previsão por (c_k constantes):

$$x_{t1} = 1, x_{t2} = c_2, x_{t3} = c_3, \ldots, x_{tk} = c_k.$$

Definição 5.3 (Previsão em média)**.** O objetivo é prever

$$\theta = E(y_t \mid x_{t2} = c2, x_{t3} = c_3, \ldots, x_{tk} = c_k) = \beta_1 + \beta_2 c_2 + \cdots + \beta_k c_k.$$

- Estimador BLUE para θ:

$$\hat{\theta} = b_1 + b_2 c_2 + \cdots + b_k c_k.$$

- Sabemos que

$$Var(\hat{\theta} \mid X, c) = Var(\mathbf{cb} \mid X, c) = \mathbf{c} Cov(\mathbf{b} \mid X, c)\mathbf{c}^T = \sigma^2 \mathbf{c}(X^T X)^{-1}\mathbf{c}^T.$$

- Podemos estimar σ^2 por s^2. Logo,
- Erro padrão da previsão em média:

$$s_{\hat{\theta}} = s\sqrt{\mathbf{c}(X^T X)^{-1}\mathbf{c}^T}.$$

- Nota: nesta UC, o valor $\mathbf{c}(X^T X)^{-1}\mathbf{c}^T$ é fornecido nos enunciados.

Definição 5.4 (Inferência sobre a previsão em média)**.** Admitindo a hipótese de normalidade das variáveis residuais

$$\frac{\hat{\theta} - \theta}{s_{\hat{\theta}}} \sim t(n-k).$$

- Com este resultado podem determinar-se intervalos de confiança **intervalos de previsão** para

$$\theta = (y_t \mid x_{t2} = c2, x_{t3} = c_3, \ldots, x_{tk} = c_k)$$

e conduzir testes de hipóteses.

- Quando se abandona a hipótese de normalidade, o resultado é assintótico, sendo válida apenas para grandes amostras.
- Intervalo de previsão a $100(1-\alpha)\%$ para θ:

$$IP_{100(1-\alpha)\%}(\theta) = \left(\widehat{\theta} \mp t_{\alpha/2} s_{\widehat{\theta}}\right).$$

- Testes de hipóteses: proceder como habitualmente.

Definição 5.5 (Previsão pontual)**.** O objetivo é prever y_0, mantendo $x_{t1} = 1, x_{t2} = c_2, x_{t3} = c_3, \ldots, x_{tk} = c_k$ e supondo que

$$y_0 = \beta_1 + \beta_2 c_2 + \beta_3 c_3 + \cdots + \beta_k x_k + u_0,$$

onde

$$E(u_0 \mid X, c) = 0 \quad \text{e} \quad Var(u_0 \mid X, c) = \sigma^2.$$

O estimador dos mínimos quadrados de y_0 é

$$\hat{y}_0 = b_1 + b_2 c_2 + b_3 c_3 + \cdots + b_k c_k.$$

- Enquanto na previsão em média se pretendia estimar $E(y_0 \mid X, c)$, na previsão pontual procura-se prever os valores assumidos por y_0.
- Erro padrão da previsão:

$$s_d = s\sqrt{1 + \mathbf{c}(X^T X)^{-1}\mathbf{c}^T},$$

onde $d = y_0 - \hat{y}_0$.

- Nota: nesta UC, o valor $\mathbf{c}(X^T X)^{-1}\mathbf{c}^T$ é fornecido nos enunciados.

Exercício 5.3 (Exercício 11.14 do livro (adaptado))**.** Retome-se o exercício 19 do capítulo 10 (corresponde ao exercício 4.2 desta sebenta). Considere um apartamento com 250 m^2, que dista 8 km do centro da cidade e admita $s^2 c(X^T X)^{-1} c^T = 558.9769$.

(a) Determine a previsão pontual e por intervalo (nível de confiança de 95%) para o preço médio por m^2 para um apartamento nas condições indicadas.

```
#
```

(b) Determine a previsão pontual e por intervalo (nível de confiança de 95%) para o preço por m^2 de um apartamento nas condições indicadas.

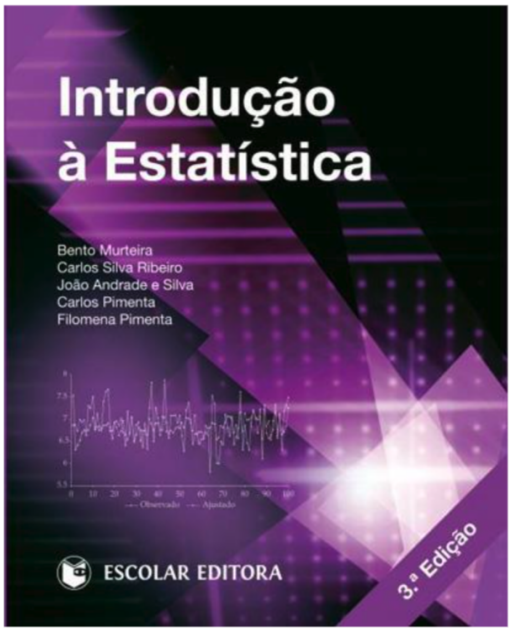
```
#
```

```
#
```

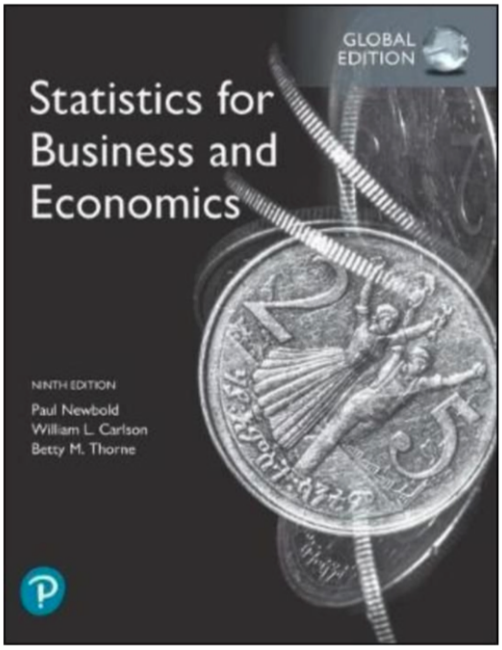
Bibliografia

Alguns exercícios e exemplos foram inspirados na seguinte bibliografia:

- B. Murteira, C. Ribeiro, J. Andrade e Silva, C. Pimenta, F. Pimenta, (2015) Introdução à Estatística (3ª Edição). Escolar Editora.



- P. Newbold, W. Carlson, B. Thorne, (2020) Statistics for Business and Economics (9th Edition), Pearson Education.



- G. Casella, R. Berger, (2002) Statistical Inference (2nd Edition), Thomson Learning.