Estatística II

Licenciatura em Gestão do Desporto

2.º Ano/2.º Semestre

2024/2025

Aulas Teórico-Práticas N.º 20 e 21 (Semana 11)

Docente: Elisabete Fernandes
E-mail: efernandes@iseg.ulisboa.pt

https://doity.com.br/estatistica-aplicada-a-nutricao

https://basiccode.com.br/produto/informatica-basica/

Conteúdos Programáticos

Aulas Teórico-Práticas (Semanas 1 a 3)	Aulas Teórico-Práticas (Semanas 4 a 7)	Aulas Teórico-Práticas (Semanas 7 a 9)	Aulas Teórico-Práticas (Semanas 10 a 13)
• **Capítulo 1**: Revisões e Distribuições de Amostragem	• **Capítulo 2**: Estimação	• **Capítulo 3**: Testes de Hipóteses	• **Capítulo 4**: Modelo de Regressão Linear Múltipla

Material didático: Exercícios do Livro Murteira et al (2015), Formulário e Tabelas Estatísticas

Bibliografia: B. Murteira, C. Silva Ribeiro, J. Andrade e Silva, C. Pimenta e F. Pimenta; *Introdução à Estatística*, 2ª ed., Escolar Editora, 2015.

1

Testes de Hipóteses para σ_1^2/σ_2^2

Hipóteses, Estatística de Teste e Decisão

Teste de Hipóteses para σ_1^2/σ_2^2: Formulário

- **POPULAÇÕES NORMAIS**

Variância corrigida

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

Média	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\frac{\bar{X}-\mu}{S'/\sqrt{n}} \sim t(n-1)$
Diferença de médias	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}} \sim N(0,1)$ $T = \frac{\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{1}{m}+\frac{1}{n}}}}{\sqrt{\frac{(m-1)S_1'^2+(n-1)S_2'^2}{m+n-2}}} \sim t(m+n-2)$	$Z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1'^2}{m}+\frac{S_2'^2}{n}}} \overset{a}{\sim} t(\nu)$ onde ν é o maior inteiro contido em r, $r = \frac{\left(\frac{s_1'^2}{m}+\frac{s_2'^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s_1'^2}{m}\right)^2+\frac{1}{n-1}\left(\frac{s_2'^2}{n}\right)^2}$
Variância	$\frac{nS^2}{\sigma^2} = \frac{(n-1)S'^2}{\sigma^2} \sim \chi^2(n-1)$	
Relação de variâncias	$\frac{S_1'^2}{S_2'^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F(m-1,n-1)$	

Teste de Hipóteses para σ_1^2/σ_2^2

Considere que duas populações Normais, com médias μ_1 e μ_2 e desvios padrão σ_1 e σ_2, das quais se extraíram aleatoriamente duas amostras independentes com dimensão n_1 e n_2, respectivamente. Pretende-se realizar um teste de hipóteses para comparar as variâncias populacionais σ_1^2 e σ_2^2, i. e., para o quociente de variâncias (σ_2^2/σ_1^2). A estatística de teste a utilizar é:

$$F = \frac{S_1^2}{S_2^2}\left(\frac{\sigma_2^2}{\sigma_1^2}\right)_0 \sim F_{n_1-1;\, n_2-1},$$

onde $(\sigma_2^2/\sigma_1^2)_0$ representa o valor que se assume para (σ_2^2/σ_1^2) em H_0.

Teste de Hipóteses para σ_1^2/σ_2^2

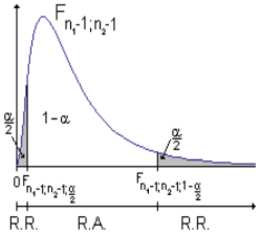
Nesta secção considera-se apenas a situação em que $(\sigma_2^2/\sigma_1^2)_0 = 1$, o que equivale a comparar a igualdade das variâncias populacionais.

T. bilateral

Hipóteses a testar:

$$H_0: \sigma_1^2 = \sigma_2^2 \ vs \ H_1: \sigma_1^2 \neq \sigma_2^2 \Leftrightarrow$$
$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \ vs \ H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Regiões críticas:

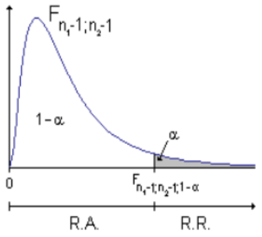


T. unilateral direito

Hipóteses a testar:

$$H_0: \sigma_1^2 \leq \sigma_0^2 \ vs: \sigma_1^2 > \sigma_2^2 \Leftrightarrow$$
$$H_0: \frac{\sigma_1^2}{\sigma_2^2} \leq 1 \ vs \ H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1$$

Regiões críticas:

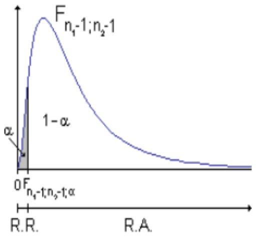


T. unilateral esquerdo

Hipóteses a testar:

$$H_0: \sigma_1^2 \geq \sigma_2^2 \ vs \ H_1: \sigma_1^2 < \sigma_2^2 \Leftrightarrow$$
$$H_0: \frac{\sigma_1^2}{\sigma_2^2} \geq 1 \ vs \ H_1: \frac{\sigma_1^2}{\sigma_2^2} < 1$$

Regiões críticas:



Nos primeiros 6 meses de vida dois grupos aleatórios de crianças seguiram esquemas de alimentação diferentes: o grupo 1 seguiu o esquema A e o grupo 2 seguiu o esquema B. No quadro seguinte apresentam-se os ganhos em peso, em kg, dessas crianças.

Grupo 1	2,7	3,2	3,6	4,1	2,7	3,2	4,5	3,6	2,7
Grupo 2	4,1	4,5	3,6	2,7	3,6	3,2	4,1		

Sabe-se que as crianças dos dois grupos tinham, ao nascer, aproximadamente pesos iguais.
Admita que as distribuições dos pesos seguem a distribuição Normal.

a) Com base num teste de hipóteses, ao nível de significância de 5%, pode concluir que a variabilidade é igual nos dois grupos?

b) Sem efectuar cálculos diga, justificando, qual a decisão que tomava no âmbito da alínea b), se considerasse um nível de significância de 1%?

c) Para o teste da alínea a) calcule o respectivo valor p e interprete.

ProbabilidadesEstatistica_2019 (uevora.pt)

Exercício (a)

Sejam:

- X_1 a v.a. que representa o ganho em peso, em kg, das crianças alimentadas segundo o esquema A,
- X_2 a v.a. que representa o ganho em peso, em kg, das crianças alimentadas segundo o esquema B,

Com $X_1 \sim N(\mu_1 = ?\,; \sigma_1 = ?)$ e $X_2 \sim N(\mu_2 = ?; \sigma_2 = ?)$.

$n_1 = 9, \quad \bar{x}_1 = 3{,}3667 \quad \text{e} \quad s_1^2 = 0{,}4150,$
$n_2 = 7, \quad \bar{x}_2 = 3{,}6857 \quad \text{e} \quad s_2^2 = 0{,}3714.$

a) $\alpha = 5\%, \sigma_1^2 = \sigma_2^2$?

$H_0: \sigma_1^2 = \sigma_2^2 \ vs\ H_1: \sigma_1^2 \neq \sigma_2^2$

$\Leftrightarrow H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \ vs\ \ H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$ (teste bilateral).

ProbabilidadesEstatistica_2019 (uevora.pt)

Exercício (a)

Estatística de teste:

$$F = \frac{S_1^2}{S_2^2}\left(\frac{\sigma_2^2}{\sigma_1^2}\right)_0 \sim F_{n_1-1;\, n_2-1=8;\, 6}.$$

$$f_{obs} = \frac{0{,}4150}{0{,}3714} \times 1 = 1{,}1173.$$

Pela tabela, $f_{n_1-1;\, n_2-1;\, \frac{\alpha}{2}} = f_{8;\, 6;\, 0{,}25} = \dfrac{1}{f_{6;\, 8;\, 0{,}975}} = \dfrac{1}{4{,}65} = 0{,}215$ e $f_{n_1-1;\, n_2-1;\, 1-\frac{\alpha}{2}} = f_{8;\, 6;\, 0{,}975} = 5{,}6.$

Logo, $R.A.$: $]0{,}215;\ 5{,}6[$ e $R.R.$: $[0;\ 0{,}215] \cup [5{,}6; +\infty[$

Como $f_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que a variabilidade nos pesos seja significativamente diferente nos dois grupos.

ProbabilidadesEstatistica_2019 (uevora.pt)

Exercícios (b) e (c)

b) Se $\alpha = 1\%$ a decisão tomada no teste anterior era a mesma, ou seja, não rejeitar H_0. Esta situação é originada pelo facto de quando se diminui o nível de significância também se diminui a $R.R.$ e consequentemente a $R.A.$ aumenta. Portanto, se quando $\alpha = 5\%$ f_{obs} está na $R.A.$ então $\alpha = 1\%$ a situação mantém-se.

c) valor $p = 2 \times min\{P(F \leq f_{obs}); P(F \geq f_{obs})\}$

$$= 2 \times min\{P(F \leq 1{,}1173); P(F \geq 1{,}1173)\}$$
$$= 2 \times min\{0{,}5409; 0{,}4591\}$$
$$= 2 \times 0{,}4591 = 0{,}9181.$$

A hipótese $H_0: \sigma_1^2 = \sigma_2^2$ é rejeitada para níveis de significância superiores ou iguais a 91,81%, indicando que não existe evidência de que as variâncias sejam diferentes.

ProbabilidadesEstatistica_2019 (uevora.pt)

Murteira et al (2015)
Capítulo 8

34. Dois programas de alimentação de gado bovino são comparados. A variável aleatória X representa o aumento de peso (em kg) de um animal alimentado segundo o programa 1, durante um mês, enquanto Y traduz o aumento de peso (em kg) de um animal alimentado segundo o programa 2, durante igual período de tempo. Sabe-se que $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$, e que as variáveis são independentes.

Um grupo de 8 animais foi submetido ao primeiro programa durante um mês, tendo-se obtido

$$\sum_{i=1}^{8} x_i = 416 \text{ e } \sum_{i=1}^{8} x_i^2 = 21807.$$

Outro grupo de 10 animais foi submetido ao segundo programa durante um mês, tendo-se obtido

$$\sum_{i=1}^{10} y_i = 468 \text{ e } \sum_{i=1}^{10} y_i^2 = 22172.$$

a) Teste, ao nível de 0.05, a igualdade entre as variâncias.
b) Teste, ao nível de 0.05, a igualdade entre os valores médios supondo variâncias iguais.
c) Teste, ao nível de 0.05, a igualdade entre os valores médios supondo variâncias diferentes.
d) Qual dos procedimentos adoptados nas alíneas anteriores, para testar a igualdade de médias, lhe parece mais adequado?

Exercício 34

$X \sim N(\mu_1, \sigma_1^2)$ $\quad$ $Y \sim N(\mu_2, \sigma_2^2)$

$n_1 = 8$ $\quad$ $n_2 = 10$

$\sum_{i=1}^{8} x_i = 416$ $\quad$ $\sum_{i=1}^{10} y_i = 468$

$\sum_{i=1}^{8} x_i^2 = 21807$ $\quad$ $\sum_{i=1}^{10} y_i^2 = 22172$

Exercício 34 a)

a) $\alpha = 0.05$

$H_0: \sigma_1^2 = \sigma_2^2 \qquad H_1: \sigma_1^2 \neq \sigma_2^2$

sob H_0, $F = \dfrac{S_1'^2}{S_2'^2} \sim F(8-1, 10-1) = F(7,9)$

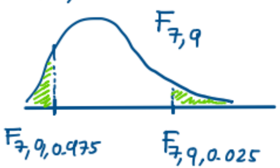
$s_1^2 = \sum_{i=1}^{8} x_i^2/8 - \bar{x}^2 = \dfrac{21807}{8} - \left(\dfrac{416}{8}\right)^2 = 21.875$

$s_1'^2 = \dfrac{n_1}{n_1 - 1} s_1^2 = \dfrac{8}{7} \times 21.875 = 25$

$s_2^2 = \sum_{i=1}^{10} y_i/10 - \bar{y}^2 = \dfrac{22172}{10} - \left(\dfrac{468}{10}\right)^2 = 26.96$

$s_2'^2 = \dfrac{n_2}{n_2 - 1} s_2^2 = \dfrac{10}{9} \times 26.96 = 29.9(5)$

$f_{obs} = \dfrac{s_1'^2}{s_2'^2} = \dfrac{25}{29.9(5)} \simeq 0.8346$



Exercício 34 a)

$$W_F = \{ f_{obs} : f_{obs} < F_{7,9,0.975} \vee f_{obs} > F_{7,9,0.025} \}$$

$$= \{ f_{obs} : f_{obs} < 0.21 \vee f_{obs} > 4.82 \}$$

$F_{7,9,0.025} = 4.82$ (tabela)

$$P(F < F_{7,9,0.975}) = 0.025 \Leftrightarrow$$

$$\Leftrightarrow P\left(\frac{1}{F} > \frac{1}{F_{7,9,0.975}} \right) = 0.025 \qquad \frac{1}{F} \sim F(9,7$$

$$\frac{1}{F_{7,9,0.975}} = F_{9,7,0.025} \Leftrightarrow$$

$$\Leftrightarrow F_{7,9,0.975} = \frac{1}{F_{9,7,0.025}} = \frac{1}{4.82} \simeq 0.21$$

$f_{obs} \notin W$ logo não se rejeita H_0.

2

Testes de Hipóteses para p

Hipóteses, Estatística de Teste e Decisão

Teste de Hipóteses para p: Formulário

População de Bernoulli

Proporção	$\dfrac{\overline{X}-\theta}{\sqrt{\dfrac{\theta(1-\theta)}{n}}} \overset{a}{\sim} N(0,1)$	$\dfrac{\overline{X}-\theta}{\sqrt{\dfrac{\overline{X}(1-\overline{X})}{n}}} \overset{a}{\sim} N(0,1)$
Diferença de proporções	$\dfrac{\overline{X}_1-\overline{X}_2-(\theta_1-\theta_2)}{\sqrt{\dfrac{\theta_1(1-\theta_1)}{m}+\dfrac{\theta_2(1-\theta_2)}{n}}} \overset{a}{\sim} N(0,1)$	$\dfrac{\overline{X}_1-\overline{X}_2-(\theta_1-\theta_2)}{\sqrt{\dfrac{\overline{X}_1(1-\overline{X}_1)}{m}+\dfrac{\overline{X}_2(1-\overline{X}_2)}{n}}} \overset{a}{\sim} N(0,1)$
Igualdade de proporções	$\dfrac{\overline{X}_1-\overline{X}_2-(\theta_1-\theta_2)}{\sqrt{\left(\dfrac{1}{m}+\dfrac{1}{n}\right)\hat{\theta}(1-\hat{\theta})}} \overset{a}{\sim} N(0,1)$ onde $\hat{\theta}=\dfrac{m\overline{X}_1+n\overline{X}_2}{m+n}$	

Estimador do Parâmetro p

Teste de Hipóteses

$$\frac{P^*-p_0}{\sqrt{\dfrac{p_0 q_0}{n}}} \overset{apr}{\frown} N(0,1)$$

Intervalo de Confiança

$$\left]\overline{P}-z_{0,95}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}};\ \overline{P}+z_{0,95}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}}\right[.$$

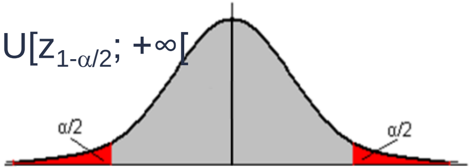
Exercício:
Considere-se uma amostra correspondente a 699 portugueses, sendo 205 fumadores.
Realize um Teste de Hipóteses para testar se a percentagem de fumadores portugueses é diferente de 40% para o nível de significância $\alpha = 1\%$.

Tipos de Testes de Hipóteses para uma Proporção Populacional

Um teste de hipóteses paramétrico para o parâmetro p (proporção populacional) pode ser:

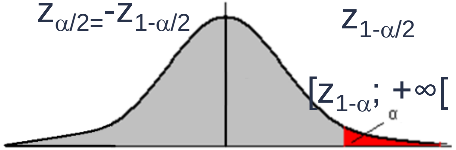
Teste

- Bilateral
  - $H_0: p = p_0$
  - $H_1: p \neq p_0$
  - $RR =]-\infty; -z_{1-\alpha/2}] \cup [z_{1-\alpha/2}; +\infty[$



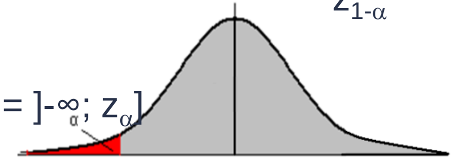
$z_{\alpha/2} = -z_{1-\alpha/2}$ $\qquad z_{1-\alpha/2}$

- Unilateral
  - à direita
    - $H_0: p \leq p_0$
    - $H_1: p > p_0$
    - $[z_{1-\alpha}; +\infty[$



$z_{1-\alpha}$

  - à esquerda
    - $H_0: p \geq p_0$
    - $H_1: p < p_0$
    - $RR =]-\infty; z_\alpha]$



$z_\alpha = -z_{1-\alpha}$

onde p_0 é o valor numérico específico considerado em H_0 e H_1.

Nota: Se $n \times p^* \geq 5$ e $n \times q^* \geq 5$ (corresponde a n grande), então a distribuição da estatística de teste é aproximadamente normal pelo Teorema de Limite Central (TLC).

Teste de Hipóteses para uma Proporção Populacional

$n \times p^* = 204{,}8 \geq 5$ e $n \times q^* = 494{,}2 \geq 5$

Hipóteses

Teste bilateral

H_0: $p = 0{,}4$ *versus* H_1: $p \neq 0{,}4$

Dados:
$n = 699 > 30$
$p^* = 205/699$
$p_0 = 0{,}40$
$q_0 = 0{,}60$
$\alpha = 0{,}01$
$z_{1-\alpha/2} = z_{0{,}995} = 2{,}576$

Tabela da Distribuição Normal

Estatística de teste

$$\frac{P^* - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \overset{apr}{\cap} N(0,1)$$

Valor observado da estatística de teste (VOE) $z_0 = z_{obs} = -5{,}775$

Região de rejeição ou crítica:
$RR =]-\infty; -z_{0,995}]U[z_{0,995}; +\infty[=]-\infty; -2{,}576]U[2{,}576; +\infty[$

Regra: $z_0 \in RR \Rightarrow$ Rejeita-se H_0

Tabela da Distribuição Normal

Decisão

Pela região de rejeição: $z_0 = -5{,}775 \in RR =]-\infty; -2{,}576]U[2{,}576; +\infty[$

Pelo valor-p: Valor-p $= 2\times P(Z \geq |-5{,}775|) = 2\times[1-P(Z \leq 5{,}775)] \sim 2\times(1-0{,}0005) = 0{,}001 < 0{,}01$

Rejeita-se H_0 para $\alpha = 1\%$. Existe evidência estatística para afirmar que, para $\alpha=1\%$, a proporção de indivíduos na pop. com a característica em estudo é diferente de 40%.

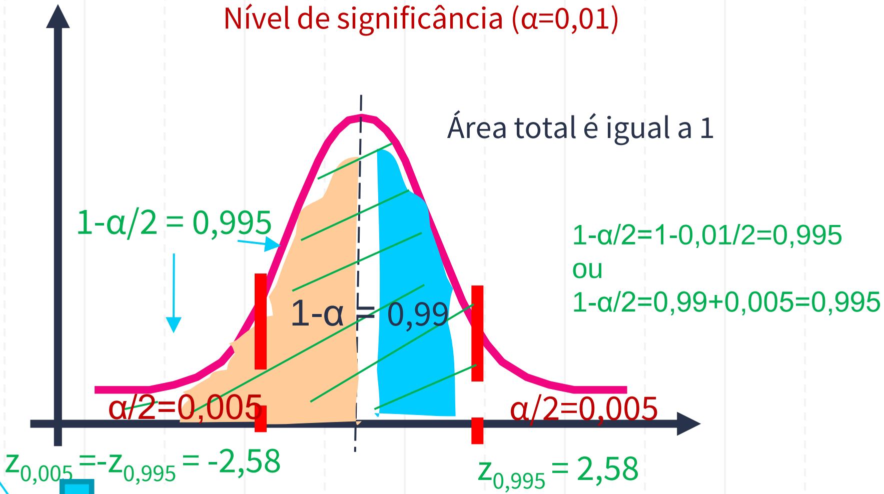
Regra: Valor-p $< \alpha \Rightarrow$ Rejeita-se H_0

Região de rejeição ou crítica:

RR =] -∞; $-z_{0,995}$]U[$z_{0,995}$; +∞[=] -∞; -2,5758]U[2,5758; +∞[

Cálculo do Quantil da Distribuição Normal Padrão de Probabilidade α

O nível de significância é igual a α = 0,01, então tem-se $z_{0,995}$ = 2,58 (ver Tabela)

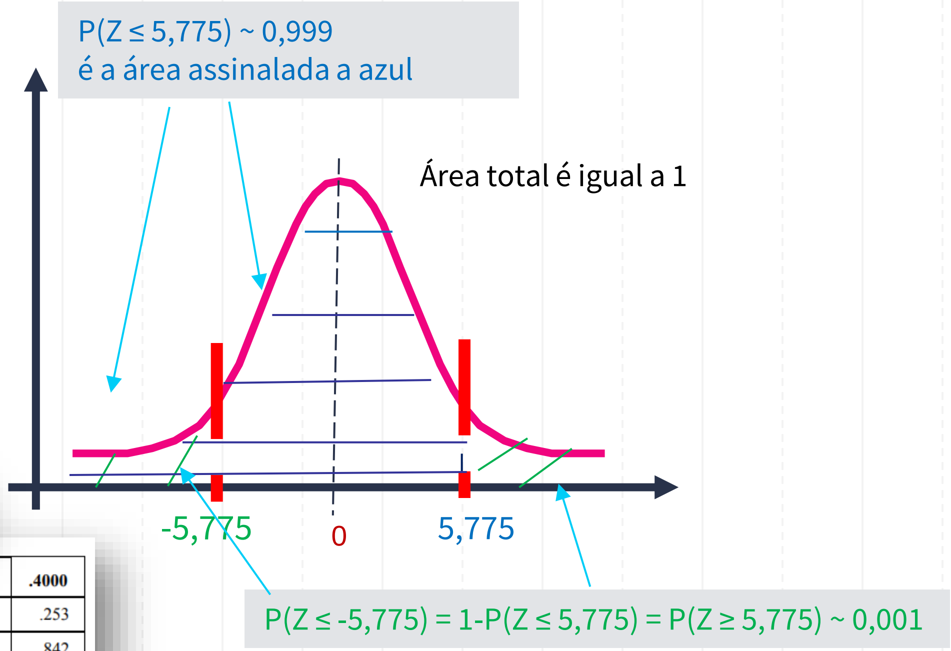


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986

Teste bilateral: valor-p = $P(Z \leq -z_0 \text{ ou } Z \geq z_0) = P(Z \leq -z_0) + P(Z \geq z_0) = 2 \times P(Z \geq |z_0|)$

Cálculo do Valor-p quando a Estatística de Teste tem Distribuição Normal Padrão

valor-p = $P(Z \leq -5{,}775 \text{ ou } Z \geq 5{,}775)$
$= 2 \times P(Z \geq 5{,}775) \sim 2 \times P(Z \geq 3{,}290) = 2 \times 0{,}0005 = 0{,}001$



$z_\varepsilon : P(Z > z_\varepsilon) = \varepsilon\,; \quad z_{\varepsilon/2} : P(|Z| > z_{\varepsilon/2}) = \varepsilon\,.$

ε	.0005	.0010	.0050	.0100	.0200	.0250	.0500	.1000	.2000	.3000	.4000
z_ε	3.290	3.090	2.576	2.326	2.054	1.960	1.645	1.282	.842	.524	.253
$z_{\varepsilon/2}$	3.481	3.290	2.807	2.576	2.326	2.241	1.960	1.645	1.282	1.036	.842

Murteira et al (2015)
Capítulo 8

45. A empresa “Trapos & Companhia” garante que pelo menos 95% das peças de fazenda que produz não têm defeito. De um grande fornecimento recebido por um armazém, foram seleccionadas 100 peças de fazenda, tendo-se apurado que sete delas apresentavam defeitos.

a) Pode pôr-se em causa a garantia da empresa? Justifique com base no teste de hipóteses adequado.

b) Qual a potência do teste efectuado na alínea anterior se a verdadeira proporção de peças defeituosas for de 0.1?

Exercício 45

A empresa "Trapos & Companhia" garante que pelo menos 95% das peças de fazenda que produz não têm defeito. De um grande fornecimento recebido por um armazém, foram seleccionadas 100 peças de fazenda, tendo-se apurado que sete delas apresentavam defeitos.

População de Bernoulli:

$$X \equiv \begin{cases} 1 & \text{(se a peça não é defeituosa)} \\ 0 & \text{(caso contrário)} \end{cases} \sim Ber(\theta)$$

Proporção de peças não defeituosas

Amostra: $n = 100$ peças $\quad \sum_{i=1}^{100} x_i = 100-7 = 93 \quad \bar{x} = \frac{93}{100} = 0.93$

Exercício 45 a)

a) Pode pôr-se em causa a garantia da empresa? Justifique com base no teste de hipóteses adequado.

$H_0: \theta \geq 0.95$ vs $H_1: \theta < 0.95$ $(\alpha = 0.05)$

$(\theta = 0.95)$

$$\frac{\bar{X}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \overset{a}{\sim} N(0,1)$$

Estatística do teste: $Z_0 = \dfrac{\bar{X} - \theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} \overset{a}{\sim} N(0,1)$

$W_Z = \{ z_{obs} : z_{obs} < -z_{0.05} = -1.645 \}$

$z_{obs} = \dfrac{\bar{x} - \theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} = \dfrac{0.93 - 0.95}{\sqrt{\frac{0.95(0.05)}{100}}} \simeq -0.9177 \notin W_Z$

$\downarrow$

Não se rejeita H_0

A evidência estatística corrobora a garantia da empresa.

Exercício 45 b)

$\hookrightarrow 1-\theta = 0.1 \Leftrightarrow \theta = 0.9$

$\hookrightarrow$ Proporção de peças defeituosas

$$1-\beta = P(\text{rejeitar } H_0 \mid H_0 \text{ é falsa}) = P(Z_0 < -1.645 \mid \theta = 0.9)$$

$$= P\left(\frac{\bar{X} - 0.95}{\sqrt{\frac{0.95(0.05)}{100}}} < -1.645 \,\middle|\, \theta = 0.9\right) =$$

$$= P\left(\bar{X} < 0.95 - 1.645\sqrt{\frac{0.95(0.05)}{100}} \,\middle|\, \theta = 0.9\right)$$

$$= P\left(\frac{\bar{X} - \theta}{\sqrt{\frac{\theta(1-\theta)}{m}}} < \frac{0.95 - 1.645\sqrt{\frac{0.95(0.05)}{100}} - 0.9}{\sqrt{\frac{0.9 \times 0.1}{100}}}\right) =$$

$\hookrightarrow \overset{a}{\sim} N(0,1)$

$$\simeq \underset{\text{Tabela}}{\Phi(0.47)} = 0.6808$$

A diferença para as soluções deve-se apenas ao arredondamento feitos antes de ir à tabela:

```
> pnorm((0.95-0.9-1.645*sqrt(0.95*0.05/100)) / sqrt(0.9*0.1/100))
[1] 0.6813945
```

Testes de Hipóteses para $p_1 - p_2$

Hipóteses, Estatística de Teste e Decisão

3

Teste de Hipóteses para p_1 - p_2: Formulário

População de Bernoulli

Proporção	$\dfrac{\overline{X}-\theta}{\sqrt{\dfrac{\theta(1-\theta)}{n}}} \overset{a}{\sim} N(0,1)$	$\dfrac{\overline{X}-\theta}{\sqrt{\dfrac{\overline{X}(1-\overline{X})}{n}}} \overset{a}{\sim} N(0,1)$
Diferença de proporções	$\dfrac{\overline{X}_1-\overline{X}_2-(\theta_1-\theta_2)}{\sqrt{\dfrac{\theta_1(1-\theta_1)}{m}+\dfrac{\theta_2(1-\theta_2)}{n}}} \overset{a}{\sim} N(0,1)$	$\dfrac{\overline{X}_1-\overline{X}_2-(\theta_1-\theta_2)}{\sqrt{\dfrac{\overline{X}_1(1-\overline{X}_1)}{m}+\dfrac{\overline{X}_2(1-\overline{X}_2)}{n}}} \overset{a}{\sim} N(0,1)$
Igualdade de proporções	$\dfrac{\overline{X}_1-\overline{X}_2-(\theta_1-\theta_2)}{\sqrt{\left(\dfrac{1}{m}+\dfrac{1}{n}\right)\hat{\theta}(1-\hat{\theta})}} \overset{a}{\sim} N(0,1)$ onde $\hat{\theta}=\dfrac{m\overline{X}_1+n\overline{X}_2}{m+n}$	

Teste de Hipóteses

$$Z = \frac{(\overline{P}_1-\overline{P}_2)-(p_1-p_2)_0}{\sqrt{\overline{P}^*\left(1-\overline{P}^*\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \overset{\circ}{\sim} N(0;1), \text{ onde } \overline{P}^* = \frac{n_1\overline{P}_1+n_2\overline{P}_2}{n_1+n_2}.$$

Intervalo de Confiança

Portanto, quando as amostras são grandes, o **I. C. para $p_1 - p_2$ com $100(1-\alpha)\%$ de confiança** é dado por:

$$\left]\overline{P}_1-\overline{P}_2-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{P}_1(1-\overline{P}_1)}{n_1}+\frac{\overline{P}_2(1-\overline{P}_2)}{n_2}};\ \overline{P}_1-\overline{P}_2+z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{P}_1(1-\overline{P}_1)}{n_1}+\frac{\overline{P}_2(1-\overline{P}_2)}{n_2}}\right[.$$

Teste de Hipóteses para p_1 - p_2

Considere duas populações Bernoulli, com parâmetros p_1 e p_2 das quais se extraíram aleatoriamente duas amostras independentes de grande dimensão n_1 e n_2, respectivamente. Pretende-se realizar um teste de hipóteses para comparar as duas proporções amostrais p_1 e p_2, i. e., para a diferença de proporções $p_1 - p_2$. Para este teste sabe-se que

$$Z = \frac{(\overline{P}_1 - \overline{P}_2) - (p_1 - p_2)_0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \overset{\circ}{\sim} N(0;1),$$

onde $(p_1 - p_2)_0$ representa o valor que se assume para $p_1 - p_2$ em H_0.

A expressão apresentada no denominador não é conhecida, apenas se conhecendo o valor da diferença $(p_1 - p_2)$ sob H_0. Como habitualmente este teste é realizado considerando $p_1 - p_2 = 0$, ou seja, $p_1 = p_2 = p$, e como esta proporção é desconhecida é substituída pelo seu estimador consistente. Desta forma, a estatística de teste a utilizar é:

$$Z = \frac{(\overline{P}_1 - \overline{P}_2) - (p_1 - p_2)_0}{\sqrt{\overline{P}^* \left(1 - \overline{P}^*\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{\circ}{\sim} N(0;1), \text{ onde } \overline{P}^* = \frac{n_1\overline{P}_1 + n_2\overline{P}_2}{n_1 + n_2}.$$

Realizou-se um estudo em duas cidades, A e B, sobre a percentagem de homens que viam o telejornal todos os dias. Na cidade A inquiriram-se aleatoriamente 150 homens tendo 54 afirmado que viam o telejornal todos os dias ao passo que na cidade B dos 200 inquiridos 80 fizeram tal afirmação.

a) Ao nível de significância de 5%, será de admitir que a proporção de homens que vê o telejornal todos os dias é:
   i. Diferente nas duas cidades?
   ii. Inferior na cidade B?
   iii. Superior na cidade B?
b) Para cada um dos testes anteriores calcule o nível de significância a partir do qual rejeita a hipótese nula.

ProbabilidadesEstatistica_2019 (uevora.pt)

Exercício (a) (i): Teste de Hipóteses para $p_1 - p_2$

Sejam:

- X_{1i} a v. a. que designa se o i-ésimo homem, da cidade A, afirmou ver o telejornal, $i = 1, \ldots, n_1$,
- X_{2i} a v. a. que designa se o i-ésimo homem, da cidade B, afirmou ver o telejornal, $i = 1, \ldots, n_2$,
- $\overline{P}_1$ a v. a. que representa a proporção de homens, da cidade A, que afirmaram ver o telejornal, em n_1 homens,
- $\overline{P}_2$ a v. a. que representa a proporção de homens, da cidade B, que afirmaram ver o telejornal, em n_2 homens.

$$n_1 = 150; \ \overline{p}_1 = \frac{54}{150} = 0{,}36; \ n_2 = 200 \ \text{e} \ \overline{p}_2 = \frac{80}{200} = 0{,}4.$$

a) $\alpha = 5\%$.

ProbabilidadesEstatistica_2019 (uevora.pt)

Exercício (a) (i): Teste de Hipóteses para $p_1 - p_2$

i) $p_1 \neq p_2$?

Hipóteses:

$H_0: p_1 = p_2 \; vs. \; H_1: p_1 \neq p_2$
$\Leftrightarrow H_0: \; p_1 - p_2 = 0 \; vs \; H_1: \; p_1 - p_2 \neq 0$ (teste bilateral).

Estatística de Teste:

Estatística de teste:

$$Z = \frac{(\overline{P}_1 - \overline{P}_2) - (p_1 - p_2)_0}{\sqrt{\overline{P}^* \left(1 - \overline{P}^*\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{\circ}{\sim} N(0;1), \text{ onde } \overline{P}^* = \frac{n_1 \overline{P}_1 + n_2 \overline{P}_2}{n_1 + n_2}.$$

$$\overline{p}^* = \frac{n_1 \overline{p}_1 + n_2 \overline{p}_2}{n_1 + n_2} = \frac{150 \times 0{,}36 + 200 \times 0{,}4}{150 + 200} = 0{,}3829,$$

$$z_{obs} = \frac{(0{,}36 - 0{,}4) - 0}{\sqrt{0{,}3829(1 - 0{,}3829)\left(\frac{1}{150} + \frac{1}{200}\right)}} = -0{,}7619.$$

ProbabilidadesEstatistica_2019 (uevora.pt)

Pela tabela $z_{1-\frac{\alpha}{2}} = z_{0{,}975} = 1{,}96$.

Logo, $R.A.:]-1{,}96;\ 1{,}96[$ e $R.R.:]-\infty;\ -1{,}96] \cup [1{,}96; +\infty[$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que a proporção de homens que afirmam ver o telejornal todos os dias é significativamente diferente nas duas cidades.

Exercício (a) (ii): Teste de Hipóteses para p_1 - p_2

Hipóteses:

Estatística de Teste:

ii) $p_1 > p_2$?

$H_0: p_1 \leq p_2$ $vs.$ $H_1: p_1 > p_2$
$\Leftrightarrow H_0: p_1 - p_2 \leq 0$ vs $H_1: p_1 - p_2 > 0$ (teste unilateral direito).

Estatística de teste:

$$Z = \frac{(\overline{P}_1 - \overline{P}_2) - (p_1 - p_2)_0}{\sqrt{\overline{P}^* \left(1 - \overline{P}^*\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{\circ}{\sim} N(0;1), \text{ onde } \overline{P}^* = \frac{n_1\overline{P}_1 + n_2\overline{P}_2}{n_1 + n_2}.$$

Como a estatística de teste é a mesma da alínea anterior, $z_{obs} = -0{,}7619$.

Pela tabela $z_{1-\alpha} = z_{0,95} = 1{,}645$. Logo, $R.A.:]-\infty;\ 1{,}645[$ e $R.R.: [1{,}645; +\infty[$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que a proporção de homens que afirmam ver o telejornal todos os dias seja significativamente inferior na cidade B.

ProbabilidadesEstatistica_2019 (uevora.pt)

Exercício (a) (iii): Teste de Hipóteses para p_1 - p_2

iii) $p_1 < p_2$?

Hipóteses:

$H_0: p_1 \geq p_2 \; vs. H_1: p_1 < p_2$
$\Leftrightarrow H_0: p_1 - p_2 \geq 0 \; vs \; H_1: p_1 - p_2 < 0$ (teste unilateral esquerdo).

Estatística de Teste:

Estatística de teste:

$$Z = \frac{(\overline{P}_1 - \overline{P}_2) - (p_1 - p_2)_0}{\sqrt{\overline{P}^* \left(1 - \overline{P}^*\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{\circ}{\sim} N(0;1), \text{ onde } \overline{P}^* = \frac{n_1\overline{P}_1 + n_2\overline{P}_2}{n_1 + n_2}.$$

Como a estatística de teste é a mesma da alínea anterior, $z_{obs} = -0{,}7619$.

Pela tabela $z_{1-\alpha} = z_{0,95} = 1{,}645$. Logo, $R.A.:]-1{,}645; +\infty[$ e $R.R.:]-\infty; -1{,}645]$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que a proporção de homens que afirmam ver o telejornal todos os dias seja significativamente superior na cidade B.

ProbabilidadesEstatistica_2019 (uevora.pt)

Exercício (b): Teste de Hipóteses para $p_1 - p_2$

Decisão (pela valor-p):

b) i) valor $p = 2 \times P(Z \geq |z_{obs}|) = 2 \times P(Z \geq 0{,}7619) = 2 \times \big(1 - \Phi(0{,}7619)\big)$
$= 2 \times (1 - 0{,}7769) = 0{,}4461.$

A hipótese $H_0: p_1 - p_2 = 0$ é rejeitada para níveis de significância superiores ou iguais a 44,61%, logo não existe evidência de que as proporções sejam diferentes, para qualquer nível se significância usual (1%, 5%, 10%).

ii) valor $p = P(Z \geq z_{obs}) = P(Z \geq -0{,}7619) = 1 - \Phi(-0{,}7619) = \Phi(0{,}7619) = 0{,}7769.$
Alternativa, com base no $valor\text{-}p$ bilateral calculado na alínea i): substituindo em H_1 p_1 e p_2 por $\overline{p}_1$ e $\overline{p}_2$, respectivamente, $\overline{p}_1 - \overline{p}_2 = 0{,}36 - 0{,}4 < 0$ dá uma proposição falsa. Logo,

$$\text{valor } p_{uni} = 1 - \frac{0{,}4461}{2} = 0{,}7769.$$

A hipótese $H_0: p_1 - p_2 \leq 0$ é rejeitada para níveis de significância superiores ou iguais a 77,69%. logo não existe evidência de que a proporção na cidade A seja superior à verificada na cidade B.

iii) valor $p = P(Z \leq z_{obs}) = P(Z \leq -0{,}7619) = \Phi(-0{,}7619) = 1 - \Phi(0{,}7619)$
$= 1 - 0{,}7769 = 0{,}2231.$
Alternativa, com base no $valor\text{-}p$ bilateral calculado na alínea i): substituindo em H_1 p_1 e p_2 por $\overline{p}_1$ e $\overline{p}_2$, respectivamente, $\overline{p}_1 - \overline{p}_2 = 0{,}36 - 0{,}4 < 0$ dá uma proposição verdadeira. Logo,

$$\text{valor } p_{uni} = \frac{0{,}4461}{2} = 0{,}2231.$$

A hipótese $H_0: p_1 - p_2 \geq 0$ é rejeitada para níveis de significância superiores ou iguais a 22,31%, logo não existe evidência de que a proporção na cidade A seja inferior à verificada na cidade B.

ProbabilidadesEstatistica_2019 (uevora.pt)

Murteira et al (2015)
Capítulo 8

48. O Ministério das Finanças está interessado em averiguar se a proporção de declarações de IRS enviadas pela *Internet* com erros de preenchimento é menor do que aquela que se verifica nas declarações entregues nos balcões das repartições de Finanças. Uma amostra de 500 declarações enviadas pela *Internet* evidenciou que 125 continham erros de preenchimento, enquanto uma amostra de 450 declarações entregues nos balcões revelou 128 mal preenchidas. Efectue o teste adequado e interprete o resultado obtido.

Exercício 48

Duas populações de Bernoulli

Declarações enviadas pela internet:

$$X = \begin{cases} 1 & \text{(se a declaração tem erros)} \\ 0 & \text{(caso contrario)} \end{cases} \qquad X \sim Ber(\theta_X)$$

Declarações entregues nos balcões:

$$Y = \begin{cases} 1 & \text{(se a declaração tem erros)} \\ 0 & \text{(caso contrário)} \end{cases} \qquad Y \sim Ber(\theta_Y)$$

Exercício 48

Amostras:

$m_X = 500$ $\quad \sum_{i=1}^{500} x_i = 125$ $\quad \bar{x} = \frac{125}{500} = 0.25$

$m_Y = 450$ $\quad \sum_{i=1}^{450} y_i = 128$ $\quad \bar{y} = \frac{128}{450} = 0.28(4)$

$H_0: \theta_X - \theta_Y \geqslant 0 \quad \text{vs} \quad H_1: \theta_X - \theta_Y < 0$

$(\theta_X - \theta_Y = 0)$

Duas populações de Bernoulli, igualdade de proporções

$$\frac{\bar{X} - \bar{Y} - (\theta_X - \theta_Y)}{\sqrt{\left(\frac{1}{m_X} + \frac{1}{m_Y}\right)\hat{\theta}(1-\hat{\theta})}} \overset{a}{\sim} N(0,1)$$

$$\hat{\theta} = \frac{m_X \bar{X} + m_Y \bar{Y}}{m_X + m_Y}$$

Exercício 48

Sob H_0, $Z_0 = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right) \hat{\theta}(1-\hat{\theta})}} \overset{a}{\sim} N(0,1)$

$W_Z = \{ z_{obs} : z_{obs} < -z_{0.05} = -1.645 \}$

$\hat{\theta} = \dfrac{n_X \bar{x} + n_Y \bar{y}}{n_X + n_Y} = \dfrac{125 + 128}{500 + 450} \simeq 0.2663$

$z_{obs} = \dfrac{(0.25 - 0.284) - 0}{\sqrt{\left(\frac{1}{500} + \frac{1}{450}\right) 0.2663(1-0.2663)}} = -1.18376 \notin W_Z$

Não se rejeita H_0.

Exercício 48

De acordo com o teste efetuado, a evidência estatística parece não corroborar a hipótese de que as declarações enviadas pela internet contêm menos erros.

Nota: Para obter o resultado das soluções basta arredondar as contas apenas no ultimo passo.

```
> theta <- (125 + 128) / (500 + 450)
> (0.25 - 128 / 450) / sqrt((1 / 500 + 1 / 450)* theta*(1 - theta))
[1] -1.199211
```

Teste de Hipóteses de Ajustamento do Q-Q

4

Hipóteses, Estatística de Teste e Decisão

4. O tabagismo é um fator de risco para neoplasia gástrica. Pretende-se saber se se pode considerar que a ocorrência de neoplasia gástrica, em fumadores, é igualmente provável na região pilórica, no corpo gástrico e na região do cárdia. Observada uma amostra aleatória constituída por 161 indivíduos com neoplasia gástrica e fumadores de 20 cigarros/dia pelo menos durante 20 anos, obteve-se a seguinte tabela de frequências:

Região pilórica	**Corpo gástrico**	**Região do cárdia**
45	54	62

Realizado o teste estatístico adequado, obteve-se o *output*:

Test Statistics

	Neoplasia gástrica
Chi-Square[a]	2,696
df	2
Asymp. Sig.	,260

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is ....

4.1. Identifique, justificando, o teste estatístico utilizado.

4.2. Formule as hipóteses estatísticas associadas ao teste.

4.3. Calcule as frequências esperadas sob a hipótese nula, e complete o output.

4.4. Indique o valor observado da estatística de teste e a forma como foi obtido.

4.5. O que pode afirmar ao nível de significância de 5%?

Exercício 4: Variável

Localização_neoplasia	Freq
• Localização de neoplasia gástrica (1-Região pilórica, 2-Corpo gástrico, 3-Região do cárdia) em fumadores • Qualitativa nominal	• Frequência absoluta

Exercícios 4.1. e 4.2: Teste de Ajustamento do Qui-Quadrado

Hipóteses

H_0: $p_1 = p_2 = p_3 = 1/3 = 33{,}3\%$ ("Ocorrência de neoplasia gástrica é igualmente provável nas 3 regiões")

Dados
n = 161
p_i = probabilidade da ocorrência de neoplasia gástrica na i-ésima região, i = 1, 2, 3

Versus

H_1: Pelo menos uma destas probabilidades regista outro valor diferente na população (ou "Ocorrência de neoplasia gástrica não é igualmente provável nas 3 regiões")

Objetivo do Teste de Ajustamento do Qui-Quadrado:

- ✓ Pretende-se saber se a ocorrência de neoplasia gástrica, em fumadores, é igualmente provável em três regiões do corpo (região pilórica, corpo gástrico e região do cárdia). De outra forma, pretende-se testar se a frequência dessa doença é igual nas 3 regiões do corpo.

Exercícios 4.3. e 4.4: Teste de Ajustamento do Qui-Quadrado

Formulário

Teste de Ajustamento: $Q=\sum_{j=1}^{m}\frac{(N_j-fe_j)^2}{fe_j}\overset{a}{\sim}\chi^2(m-1)$

Com estimação de k parâmetros para obter as estimativas $\hat{p}_{\circ j}$: $\chi^2_{(m-k-1)}$

Estatística de teste

$$\sum_i\frac{(O_i-E_i)^2}{E_i}\overset{apr}{\cap}\chi^2_{(k-1)}$$

Região	Freq. obser. (O_i)	Freq. esper. ($E_i = n \times p_i$)	Resíduos (O_i-E_i)
1 - Região pilórica	45	161×1/3=53,67	-8,67
2 - Corpo gástrico	54	53,67	0,33
3 - Região do cárdia	62	53,67	8,33
Total	161		

Pela fórmula:
VOE=(-8,67)²/53,67+0,33²/53,67
+8,33²/53,67 = 2,695

Resposta: E_i mínimo é 53,67

Condições de Aplicabilidade dos Testes do Qui-Quadrado:
- As frequências esperadas devem ser ≥ 5.
- No caso de tal não se verificar, então pelo menos 80% das frequências esperadas ≥ 5 e todas > 1

A Condição de Aplicabilidade dos Testes do Qui-Quadrado é satisfeita:
- Todas as frequências esperadas $E_i \geq 5$.

Exercício 4.5: Teste de Ajustamento do Qui-Quadrado

Região de rejeição ou crítica:
$2{,}695 \notin RR = [\chi^2_{95;2}; +\infty[= [5{,}99; +\infty[$

Decisão (para $\alpha = 0{,}05$)

Pelo valor crítico: $VOE = 2{,}695 < \chi^2_{0,95;2} = 5{,}99$

Tabela da Distribuição do Qui-Quadrado

Quantil de probabilidade $1\text{-}\alpha$ da distribuição Qui-Quadrado

Regra de decisão pelo valor crítico ou região de rejeição (RR):

$$\begin{cases} VOE \geq \chi^2_{1-\alpha} \\ VOE \in RR = [\chi^2_{1-\alpha}; +\infty[\end{cases} \Rightarrow \text{Rejeita-se } H_0 \text{ para } \alpha$$

Regra de decisão pelo valor-p:
Valor-p $= P(X^2 \geq VOE) < \alpha \Rightarrow$ Rejeita-se H_0 para α

Pelo valor-p: valor-p $= P(X^2 \geq 2{,}695) = 0{,}260 > 0{,}05$

Cálculo do Quantil da Distribuição Qui-Quadrado de Probabilidade 1-α e com n-1 g.l.´s

Nível de confiança (1-α=0,95)

Nível de significância (α=0,05)

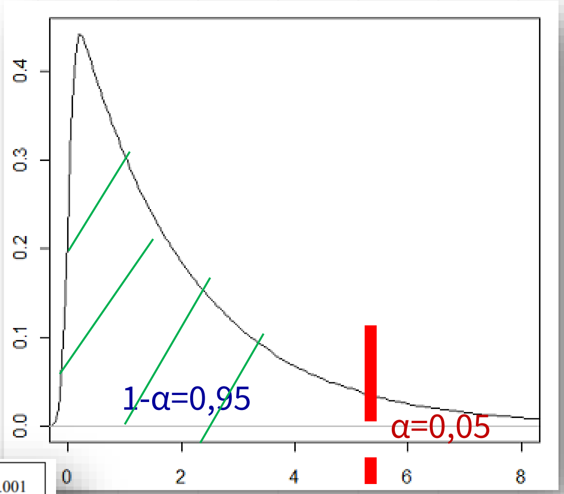
Área total é igual a 1

O nível de significância é igual a α = 0,05, então tem-se 1-α = 0,95

$\chi^2_{0,95;2}$ = 5,991 (ver tabela)

$$\chi^2_{n.\varepsilon} : P(X > \chi^2_{n.\varepsilon}) = \varepsilon$$

n \ ε	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005	.001
1	.000	.000	.001	.004	.016	.102	.455	1.323	2.706	3.841	5.024	6.635	7.879	10.827
2	.010	.020	.051	.103	.211	.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.815
3	.072	.115	.216	.352	.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	.207	.297	.484	.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.466
5	.412	.554	.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.832	15.086	16.750	20.515
6	.676	.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.457
7	.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588

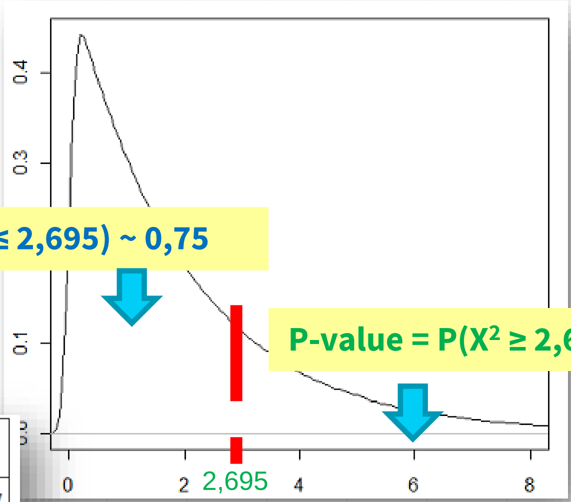


$\chi^2_{0,95;2}$ = 5,991

Cálculo do Valor-p quando a Estatística de Teste tem Distribuição Qui-Quadrado

valor-p = $P(X^2 \geq 2{,}695) \sim P(X^2 \geq 2{,}773) = 0{,}25$

Área total é igual a 1



$P(X^2 \leq 2{,}695) \sim 0{,}75$

P-value = $P(X^2 \geq 2{,}695) \sim 0{,}25$

$\chi^2_{n.\varepsilon} : P(X > \chi^2_{n.\varepsilon}) = \varepsilon$

n \ ε	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005	.001
1	.000	.000	.001	.004	.016	.102	.455	1.323	2.706	3.841	5.024	6.635	7.879	10.827
2	.010	.020	.051	.103	.211	.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.815
3	.072	.115	.216	.352	.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	.207	.297	.484	.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.466
5	.412	.554	.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.832	15.086	16.750	20.515
6	.676	.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.457
7	.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588

Regra de decisão pelo valor-p:
Valor-p = $P(X^2 \geq VOE) < \alpha \Rightarrow$ Rejeita-se H_0 para α

Murteira et al (2015)
Capítulo 9

1. Com o objectivo de remodelar determinado centro comercial, realizou-se uma pesquisa sobre o movimento de entradas e saídas pelas suas três portas. Inquiriu-se qual a porta de entrada para uma amostra aleatória de 201 pessoas:

Entrada	1	2	3
N.º de pessoas	83	62	56

Foi afirmado que não havia preferência por qualquer uma das três entradas. Comente para uma dimensão de 0.05.

Exercício 1

X = Porta escolhida para entrar no centro comercial

$D_X = \{1, 2, 3\}$ $\qquad n = 201$ pessoas

$p_j = P(X = j) \quad (j = 1, 2, 3)$

$H_0: \; p_1 = p_2 = p_3 = \frac{1}{3} \qquad H_1: \exists j : p_j \neq \frac{1}{3} \quad (j = 1, 2, 3)$

$p_{0j} = P(X = j \mid H_0) = \frac{1}{3} \quad (j = 1, 2, 3)$

Exercício 1

$$Q = \sum_{j=1}^{m} \frac{(N_j - fe_j)^2}{fe_j} \stackrel{a}{\sim} \chi^2(m-1) = \chi^2(2)$$

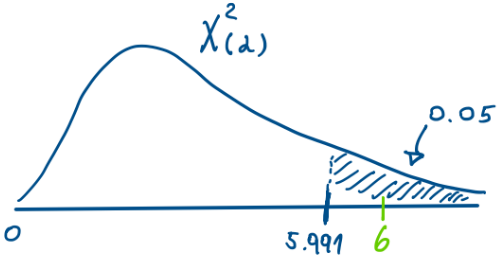
m = nº de categorias

Porta	N_j	$fe_j = n\,P_{0j}$	
1	83	$201 \times \frac{1}{3} = 67$	>5
2	62	$201 \times \frac{1}{3} = 67$	>5
3	56	$201 \times \frac{1}{3} = 67$	>5

$m = 3$

$$Q_{obs} = \frac{(83-67)^2}{67} + \frac{(62-67)^2}{67} + \frac{(56-67)^2}{67} = 6$$

Exercício 1



$\alpha = 0.05$

$\chi^2_{2,0.05} = 5.991$

$W_Q = \{ Q_{obs} : Q_{obs} > \chi^2_{2,0.05} = 5.991 \}$

$Q_{obs} = 6 \in W_Q$ logo rejeita-se H_0.

Conclusão: Tendo como referência um teste de dimensão 0.05 a hipótese nula é rejeitada, o que significa que a evidência estatística presente na amostra recolhida não é favorável à afirmação enunciada.

Nota:

```
> pchisq(6, 2, lower.tail = FALSE)
[1] 0.04978707
```

$p_{obs} = 0.0498 < \alpha = 0.05$ logo rejeita-se H_0.

Com $\alpha = 0.01$ já não se rejeitava H_0.

Murteira et al (2015)
Capítulo 9

4. O número de erros de impressão por página de um livro é frequentemente considerado uma variável aleatória de Poisson. A contagem dos erros de impressão em 100 páginas de um livro deu o seguinte resultado:

N.º de erros	0	1	2	3
N.º de páginas	65	25	8	2

a) Teste a hipótese, ao nível de 0.01, de que o número de erros por página é uma variável aleatória de Poisson de parâmetro $\lambda = 0.4$.

Exercício 4 a)

a) $X \equiv$ nº de erros por página
amostra: $n = 100$ páginas

$H_0: X \sim Poi(0.4)$ $\quad H_1: X \nsim Poi(0.4)$ $\quad (\alpha = 0.01)$

sob H_0, $f_X(x) = \dfrac{e^{-0.4}\, 0.4^x}{x!}$ $\quad (x = 0, 1, 2, \dots)$

erros	N_j	$n\,p_{0j}$
0	65	67.03
1	25	26.81
2	8	5.36
≥ 3	2	0.8 < 5

→ Temos que agrupar estas duas categorias:

$m = 3 \quad N_3 = 10$

$n\,p_{03} = 5.36 + 0.8 = 6.16$

Exercício 4 a)

$$P_{0j} = P(X = j-1 \mid H_0) = \frac{e^{-0.4}\, 0.4^{j-1}}{(j-1)!} \quad (j = 1, 2, 3)$$

$$P_{04} = P(X \geq 3 \mid H_0) = 1 - P_{01} - P_{02} - P_{03}$$

Assim sendo:

Tabela 2

$$n P_{01} = 100 P(X = 0 \mid H_0) = 100 \times \underline{0.6703} = 67.03$$

$$n P_{02} = 100 P(X = 1 \mid H_0) = 100 \times \underline{0.2681} = 26.81$$

$$n P_{03} = 100 P(X = 2 \mid H_0) = 100 \times \underline{0.0536} = 5.36$$

$$n P_{04} = 100 P(X \geq 3 \mid H_0) = 100 \times (1 - P_{01} - P_{02} - P_{03}) =$$

$$= 100(1 - 0.6703 - 0.2681 - 0.0536) = 100 \times 0.008 = 0.8$$

Exercício 4 a)

A Hipótese que realmente se testa é a hipótese "aparentada":

$$H_0': \begin{cases} P(X = x) = \dfrac{e^{-0.4}\, 0.4^{x}}{x!} \quad (x = 0,1) \\ P(X \geq 2) = 1 - P(X=0) - P(X=2) \end{cases}$$

$$Q = \sum_{j=1}^{3} \frac{(N_j - n P_{0j})^2}{n P_{0j}} \overset{a}{\sim} \chi^2_{(m-1)} = \chi^2_{(2)}$$

$$\chi^2_{2,\,0.01} = 9.21$$

$$W_Q = \{Q_{obs} : Q_{obs} > 9.21\}$$

Exercício 4

$$Q_{obs} = \frac{(65-67.03)^2}{67.03} + \frac{(25-26.81)^2}{26.81} + \frac{(10-6.16)^2}{6.16} =$$

$$= 2.58 \notin W_Q \text{ logo não se rejeita } H_0'.$$

Teste de Hipóteses de Independência do Q-Q

Hipóteses, Estatística de Teste e Decisão

5

5. Um inspetor de qualidade recolheu uma amostra de 176 produtos alimentares num centro de distribuição. Sabendo que cada produto pode ser proveniente de uma de três fábricas e pode ou não estar contaminado; o inspetor avaliou todos os produtos e obteve os seguintes resultados:

	Fábrica A	**Fábrica B**	**Fábrica C**	**Total**
Contaminado	8	15	11	34
Não contaminado	55	67	20	142
Total	63	82	31	176

Pode-se afirmar que o facto de um produto estar contaminado é independente da sua fábrica de origem, considerando $\alpha = 0,01$?

[*Adaptado da fonte: https://www.ime.unicamp.br/~veronica/Coordenadas1s/aula8pr.pdf*]

Formulário

Teste de Independência: $Q = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(N_{ij} - fe_{ij})^2}{fe_{ij}} \overset{a}{\sim} \chi^2_{((r-1)(s-1))}$

Exercício: Teste de Independência do Qui-Quadrado

Dados
N = 176
VOE = 7,024
Valor-p = 0,03
α = 1% $\Rightarrow$ 1-α = 99%
df = g.l.´s = 2

Hipóteses

H_0: As variáveis são independentes

Versus

H_1: As variáveis não são independentes

Frequências esperadas

$$E_{ij} = \frac{L_i \times C_j}{N}$$

L = total linhas e C = total colunas
N = nº total de elementos

Estatística de teste

$$\sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{apr}{\cap} \chi^2_{(l-1)(c-1)}$$

Tabela da distribuição do Qui-Quadrado

Decisão

Pelo valor crítico: Valor da Estatística de Teste = 7,024 não pertence a RR = [$\chi^2_{0,99;2}$; +∞[= [9,21; +∞[

Pelo valor-p: valor-p = 0,030 > 0,01

Não se rejeita H_0 para α = 1%. Assim, não existe evidência estatística para afirmar que as variáveis não são independentes para α = 1%.

Exercício: Teste de Independência do Qui-Quadrado

Estatística de Teste:

$$\sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{apr}{\cap} \chi^2_{(l-1)(c-1)}$$

Frequências esperadas

$$E_{ij} = \frac{L_i \times C_j}{N}$$

L = total linhas e C = total colunas
N = nº total de elementos

Contaminado * Fábrica Tabulação cruzada

			Fábrica			
			Fábrica A	Fábrica B	Fábrica C	Total
Contaminado	Sim	Contagem	8	15	11	34
		Expected Count	12,2	15,8	6,0	34,0
	Não	Contagem	55	67	20	142
		Expected Count	50,8	66,2	25,0	142,0
Total		Contagem	63	82	31	176
		Expected Count	63,0	82,0	31,0	176,0

Valor da Estatística de Teste

Testes de chi-quadrado

	Valor	df	Sig. Assint. (2 lados)
Chi-quadrado de Pearson	7,024[a]	2	,030
Razão de probabilidade	6,450	2	,040
Associação Linear por Linear	6,099	1	,014
N de Casos Válidos	176		

a. 0 células (0,0%) esperam contagem menor do que 5. A contagem mínima esperada é 5,00.

A Condição de Aplicabilidade dos Testes do Qui-Quadrado é satisfeita:

- Todas as frequências esperadas $E_i \geq 5$.

Teste de Independência do Qui-Quadrado

Região de rejeição ou crítica:
7,024 não pertence a RR = $[\chi^2_{0,99;2}; +\infty[$ = $[9,21; +\infty[$

Decisão (para $\alpha = 0,01$)
Pelo valor crítico: VOE = 7,024 > $\chi^2_{0,99;2} = 9,21$

Tabela da Distribuição do Qui-Quadrado

Regra de decisão pelo valor crítico ou região de rejeição (RR):

$$\begin{cases} VOE \geq \chi^2_{1-\alpha} \\ VOE \in RR = [\chi^2_{1-\alpha}; +\infty[\end{cases} \Rightarrow \text{Rejeita-se } H_0 \text{ para } \alpha$$

Quantil de probabilidade 1-α da distribuição do Qui-Quadrado

Pelo valor-p: valor-p = 0,03 > 0,01

Regra de decisão pelo valor-p:
Valor-p = $P(X^2 \geq VOE) < \alpha \Rightarrow$ Rejeita-se H_0 para α

Não se rejeita-se H_0 para $\alpha = 0,01$. Assim, não existe evidência estatística para afirmar que as variáveis não são independentes para $\alpha = 1\%$.

Cálculo do Quantil da Distribuição Qui-Quadrado de Probabilidade 1-α e com (l-1)×(c-1) g.l.´s

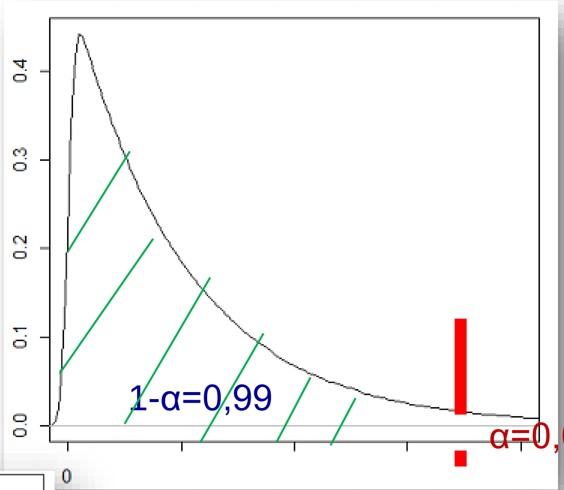
Nível de confiança (1-α=0,99)

Nível de significância (α=0,01)

Área total é igual a 1

O nível de significância é igual a α = 0,01, então tem-se 1-α = 0,99

$\chi^2_{0,99;2}$ = 9,21 (ver tabela)



α=0,01

$\chi^2_{0,99;2} = 9,21$

$\chi^2_{n.\varepsilon} : P(X > \chi^2_{n.\varepsilon}) = \varepsilon$

n \ ε	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005	.001
1	.000	.000	.001	.004	.016	.102	.455	1.323	2.706	3.841	5.024	6.635	7.879	10.827
2	.010	.020	.051	.103	.211	.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.815
3	.072	.115	.216	.352	.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	.207	.297	.484	.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.466
5	.412	.554	.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.832	15.086	16.750	20.515
6	.676	.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.457
7	.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588

Regra de decisão pelo valor-p:
Valor-p = P($X^2 \geq$ VOE) < $\alpha \Rightarrow$ Rejeita-se H_0 para α

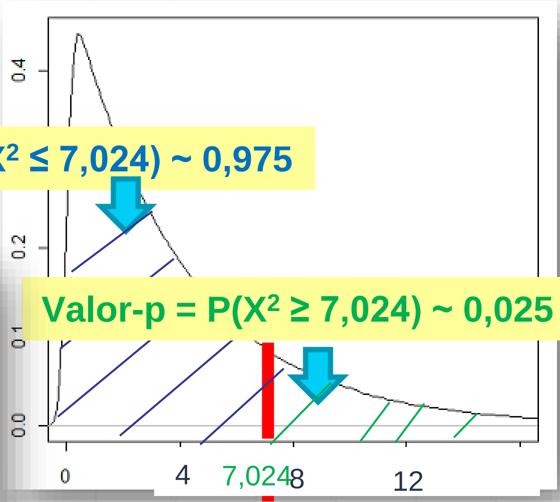
Cálculo do Valor-p quando a Estatística de Teste tem Distribuição Qui-Quadrado

valor-p = P($X^2 \geq$ 7,024) ~ P($X^2 \geq$ 7,378) = 0,025

$$\chi^2_{n.\varepsilon} : P(X > \chi^2_{n.\varepsilon}) = \varepsilon$$

n \ ε	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005	.001
1	.000	.000	.001	.004	.016	.102	.455	1.323	2.706	3.841	5.024	6.635	7.879	10.827
2	.010	.020	.051	.103	.211	.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.815
3	.072	.115	.216	.352	.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	.207	.297	.484	.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.466
5	.412	.554	.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.832	15.086	16.750	20.515
6	.676	.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.457
7	.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588

Área total é igual a 1



P($X^2 \leq$ 7,024) ~ 0,975

Valor-p = P($X^2 \geq$ 7,024) ~ 0,025

Exercício: Teste de Independência do Qui-Quadrado

Condições de Aplicabilidade dos Testes do Qui-Quadrado:
- As frequências esperadas devem ser ≥ 5.
- No caso de tal não se verificar, então pelo menos 80% das frequências esperadas ≥ 5 e todas > 1 (não é válido para tabelas 2x2).

Frequências esperadas

$$E_{ij} = \frac{L_i \times C_j}{N}$$

L = total linhas e C = total colunas
N = nº total de elementos

Verificação das condições de aplicabilidade

Neste caso, todas as células têm frequências esperadas superiores a 5.

O teste do Qui-Quadrado apenas informa sobre a independência entre variáveis, mas nada diz sobre o grau de associação existente. Para esse efeito calculam-se **medidas de associação** tais como o coeficiente Phi, o coeficiente V de Cramer e o coeficiente de contingência.

Contaminado * Fábrica Tabulação cruzada

			Fábrica			
			Fábrica A	Fábrica B	Fábrica C	Total
Contaminado	Sim	Contagem	8	15	11	34
		Expected Count	12,2	15,8	6,0	34,0
	Não	Contagem	55	67	20	142
		Expected Count	50,8	66,2	25,0	142,0
Total		Contagem	63	82	31	176
		Expected Count	63,0	82,0	31,0	176,0

Murteira et al (2015)
Capítulo 9

16. Uma companhia de seguros está interessada em saber se existe independência entre a realização de seguros de dois tipos: " Seguro de Vida" e "Seguro de Saúde". Para tal analisou as apólices detidas por uma amostra casual de 200 clientes tendo-se verificado que: 90 tinham simultaneamente um seguro de vida e um seguro de saúde; 35 tinham apenas seguro de vida; 42 tinham apenas seguro de saúde; os restantes não tinham qualquer destes dois seguros. Com base num teste a 0.05, que pode concluir?

Exercício 16

$$H_0: \quad P_{ij} = P_{i\bullet} P_{\bullet j} \quad (i, j = 1, 2)$$

$$H_1: \exists (i,j): P_{ij} \neq P_{i\bullet} P_{\bullet j}$$

$$Q = \frac{\sum_{i=1}^{2} \sum_{j=1}^{2} \left(N_{ij} - fe_{ij} \right)^2}{fe_{ij}} \overset{a}{\sim} \chi^2_{\{(\overset{\#\text{ linhas}}{2}-1)(\overset{\#\text{ colunas}}{2}-1)\}} = \chi^2_{(1)}$$

Total | 132 68 | 200

$$\hat{P}_{\bullet 1} = \frac{132}{200} = 0.66 \qquad \frac{68}{200} = 0.34 = P_{\bullet 2}$$

Exercício 16

$H_0: \quad P_{ij} = P_{i\bullet} P_{\bullet j} \quad (i, j = 1, 2)$

$H_1: \exists (i, j): P_{ij} \neq P_{i\bullet} P_{\bullet j}$

$$Q = \frac{\sum_{i=1}^{2} \sum_{j=1}^{2} (N_{ij} - fe_{ij})^2}{fe_{ij}} \overset{a}{\sim} \chi^2 \{ (\overset{\#\text{ linhas}}{2} - 1)(\overset{\#\text{ colunas}}{2} - 1) \} = \chi^2(1)$$

Exercício 16

Frequencias esperadas sob H_0 (fe_{ij}):

$$fe_{ij} = n\,\hat{P}_{i\bullet}\,\hat{P}_{\bullet j} \quad (i, j = 1, 2)$$

		seguro saúde: sim	seguro saúde: não
seguro vida	sim	$200 \times 0.625 \times 0.66 = 82.5$	$200 \times 0.625 \times 0.34 = 42.5$
	não	$200 \times 0.375 \times 0.66 = 49.5$	$200 \times 0.375 \times 0.34 = 25.5$

Exercício 16

$$\chi^2_{1,\,0.05} = 3.841 \qquad W_Q = \{Q_{obs} : Q_{obs} > 3.841\}$$

$$Q_{obs} = \frac{(90-82.5)^2}{82.5} + \frac{(35-42.5)^2}{42.5} + \frac{(42-49.5)^2}{49.5} + \frac{(33-25.5)^2}{25.5} =$$

$$= 5.3476 \in W_Q \rightarrow \text{Rejeitamos } H_0.$$

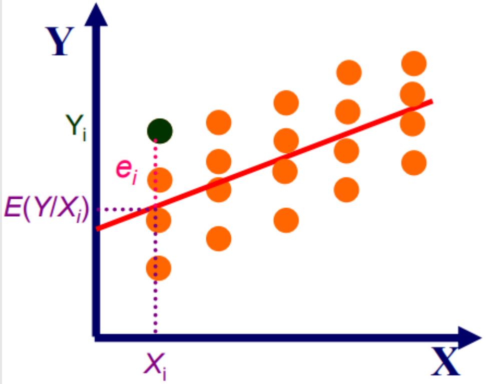
6

Modelo de Regressão Linear Múltipla

Estimação dos Coeficientes da Reta de Regressão

Modelo de Regressão Linear Simples: Revisão

Seja a relação entre *Y* e *X* na população:



$$Y_i = \alpha + \beta X_i + e_i$$

ou

$$E(Y/X_i) = \alpha + \beta X_i$$

Modelo de Regressão Linear Simples para *Y* na população

Onde:

Y é a variável dependente ou regressando

X é a variável independente ou regressor

α é o intercepto ou constante do modelo

β é o coeficiente angular do modelo

Beta é o declive

Alfa é a ordenada na origem

Erro de previsão:

Seja X_i a *i*-ésima observação de *X*, teremos:

Y_i é o valor observado em *Y* para o *i*-ésimo valor de *X*

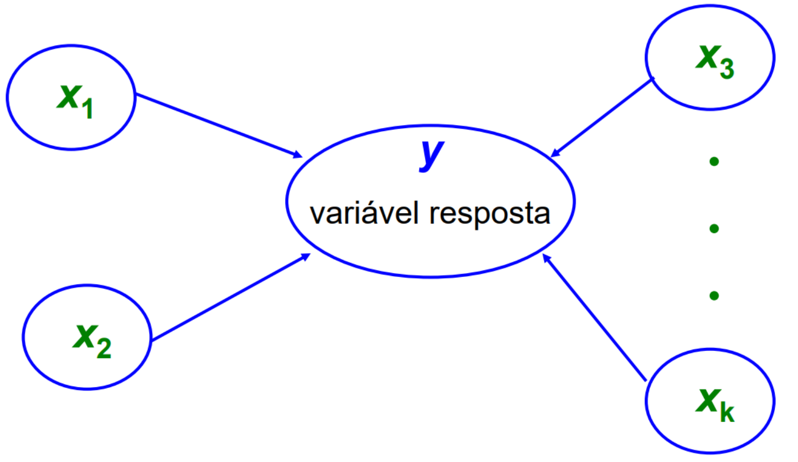
$E(Y/X_i)$ é a esperança condicional de *Y* e representa o valor esperado de *Y* para o *i*-ésimo valor de *X*

e_i é o erro, ou variação de Y_i não explicada pelo modelo

Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

Modelo de Regressão Linear Múltipla (MRLM)

Chamamos Modelo de Regressão Linear Múltipla a qualquer modelo de regressão linear com duas ou mais variáveis explicativas.



x_1, x_1, ..., x_k: variáveis explicativas (regressores)

[Index of /wp-content/uploads/2014/02 (hedibert.org)](#)

MRLM

Vamos admitir que X_1, X_2, ..., X_k sejam as variáveis independentes e Y a variável dependente.

Dada uma amostra de *n* observações,

$$(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i), \quad i = 1, 2, \ldots, n,$$

Index of /wp-content/uploads/2014/02 (hedibert.org)

MRLM

o modelo de regressão linear múltipla será dado por:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \,,$$

ou

$$E[y_i \mid x_{1i}, x_{2i}, \ldots, x_{ki}] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} \,,$$

$i = 1, 2, \ldots, n$ em que $n > (k+1)$.

Neste modelo, k é o nº de variáveis independentes e k+1 é o nº de coeficientes

Index of /wp-content/uploads/2014/02 (hedibert.org)

Estimação dos Coeficientes do MRLM: Método dos Mínimos Quadrados (MMQ)

Para determinarmos os estimadores de mínimos quadrados de β_0, β_1, ..., β_k, devemos minimizar o erro quadrático total ($\Sigma\varepsilon_i^2$):

$$\sum_{i=1}^{n}(\varepsilon_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_k x_{ki})^2$$

Index of /wp-content/uploads/2014/02 (hedibert.org)

Estimação dos Coeficientes do MRLM: MMQ

O mínimo da função

$$S(\beta_0,\ \beta_1,\ \ldots,\beta_k) = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_k x_{ki})^2$$

é obtido derivando-a em relação a β_0, β_1, ..., β_k, e igualando o resultado a zero. Ou seja,

$$\frac{\partial}{\partial \beta_0} S(\beta_0,\ \beta_1,\ \ldots,\beta_k) = 0 \qquad \cdots \qquad \frac{\partial}{\partial \beta_k} S(\beta_0,\ \beta_1,\ \ldots,\beta_k) = 0$$

Index of /wp-content/uploads/2014/02 (hedibert.org)

Estimação dos Coeficientes do MRLM: MMQ

$$\frac{\partial}{\partial \beta_0} S(\beta_0,\ \beta_{1,}\ \ldots,\beta_k) = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right) = 0$$

$$\frac{\partial}{\partial \beta_1} S(\beta_0,\ \beta_{1,}\ \ldots,\beta_k) = -2\sum_{i=1}^{n}\left[\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right) x_{1i}\right] = 0$$

$$\vdots$$

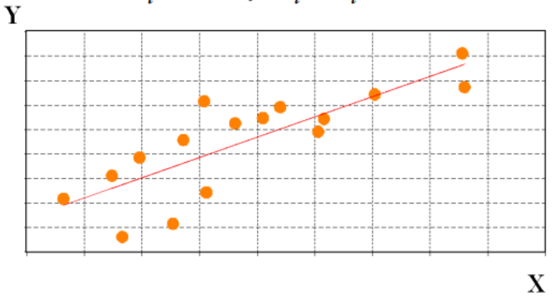
$$\frac{\partial}{\partial \beta_k} S(\beta_0,\ \beta_{1,}\ \ldots,\beta_k) = -2\sum_{i=1}^{n}\left[\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right) x_{ki}\right] = 0$$

Index of /wp-content/uploads/2014/02 (hedibert.org)

Estimação dos Coeficientes do MRLM: MMQ

Regressão Linear Simples:

$$Y_i = \alpha + \beta X_i + e_i$$



Onde:

$$EQT(\hat{\alpha}, \hat{\beta}) = \sum \hat{e}_i^2 = \sum [Y_i - (\hat{\alpha} + \hat{\beta} X_i)]^2$$

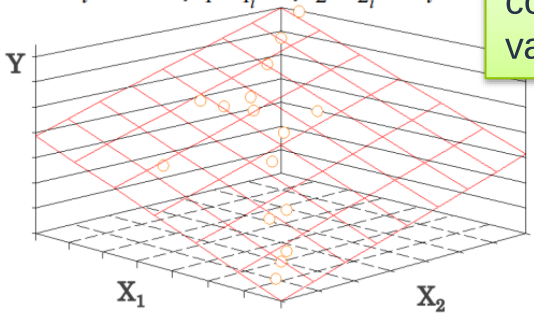
Minimizando EQT:

$$\frac{\partial EQT}{\partial \hat{\alpha}} = 0 \;\Rightarrow\; \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\frac{\partial EQT}{\partial \hat{\beta}} = 0 \;\Rightarrow\; \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Regressão Linear Múltipla:

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i$$



Onde:

$$EQT(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2) = \sum \hat{e}_i^2 = \sum [Y_i - (\hat{\alpha} + \hat{\beta}_1 X_{1_i} + \hat{\beta}_2 X_{2_i})]^2$$

Minimizando EQT:

$$\frac{\partial EQT}{\partial \hat{\alpha}} = 0 \;\Rightarrow\; \hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

$$\frac{\partial EQT}{\partial \hat{\beta}_1} = 0 \;\Rightarrow\; \hat{\beta}_1 = \frac{(\sum y_i x_{1i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}$$

$$\frac{\partial EQT}{\partial \hat{\beta}_2} = 0 \;\Rightarrow\; \hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}$$

Caso particular: MRLM com apenas duas variáveis regressoras

Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

MRLM: Abordagem Matricial

Devido à complexidade das fórmulas envolvidas, utilizaremos a abordagem matricial, que nos permitirá, entre outras coisas:

i. encontrar o vetor de estimadores;

ii. verificar as propriedades estatísticas de (i);

iii. obter a distribuição de probabilidades de (i);

qualquer que seja o número de regressores presentes no modelo.

Index of /wp-content/uploads/2014/02 (hedibert.org)

MRLM: Abordagem Matricial

Assim, a equação

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i\ , \ i = 1, 2, \ldots, n.$$

também pode ser escrita como

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \ldots + \beta_k x_{k1} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \ldots + \beta_k x_{k2} + \varepsilon_2$$

$$y_3 = \beta_0 + \beta_1 x_{13} + \beta_2 x_{23} + \ldots + \beta_k x_{k3} + \varepsilon_3$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \ldots + \beta_k x_{kn} + \varepsilon_n$$

Index of /wp-content/uploads/2014/02 (hedibert.org)

MRLM: Abordagem Matricial

As igualdades anteriores podem ser alocadas facilmente em dois vetores colunas $(n\times1)$, descritos a seguir:

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{(n\times1)} = \underbrace{\begin{pmatrix} \beta_0 + \beta_1 x_{11} + \ldots + \beta_k x_{k1} + \varepsilon_1 \\ \beta_0 + \beta_1 x_{12} + \ldots + \beta_k x_{k2} + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_{1n} + \ldots + \beta_k x_{kn} + \varepsilon_n \end{pmatrix}}_{(n\times1)}$$

Index of /wp-content/uploads/2014/02 (hedibert.org)

MRLM: Abordagem Matricial

Ainda,

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{(n\times 1)} = \underbrace{\begin{pmatrix} \beta_0 + \beta_1 x_{11} + \ldots + \beta_k x_{k1} \\ \beta_0 + \beta_1 x_{12} + \ldots + \beta_k x_{k2} \\ \vdots \\ \beta_0 + \beta_1 x_{1n} + \ldots + \beta_k x_{kn} \end{pmatrix}}_{(n\times 1)} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{(n\times 1)}$$

Index of /wp-content/uploads/2014/02 (hedibert.org)

MRLM: Abordagem Matricial

Finalmente,

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{(n\times 1)} = \underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix}}_{(n\times(k+1))} \cdot \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{((k+1)\times 1)} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{(n\times 1)}$$

Index of /wp-content/uploads/2014/02 (hedibert.org)

MRLM: Abordagem Matricial

Vamos definir:

$$\underset{\sim}{\mathbf{y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \underset{\sim}{\mathbf{X}} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix} \qquad \underset{\sim}{\omega_{\mathbf{i}}} = \begin{pmatrix} 1 & x_{1i} & \cdots & x_{ki} \end{pmatrix}$$

$$\underset{\sim}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \underset{\sim}{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

15

Index of /wp-content/uploads/2014/02 (hedibert.org)

Modelo de Regressão Linear Múltipla (MRLM)

Assim, utilizando os resultados do *slide* anterior, podemos escrever o modelo de regressão linear múltipla como:

$$\underset{\sim}{\mathbf{y}} = \underset{\sim}{\mathbf{X}}\,\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\boldsymbol{\varepsilon}},$$

que é chamado Modelo Linear Geral.

Index of /wp-content/uploads/2014/02 (hedibert.org)

Estimação do MRLM: Métodos dos Mínimos Quadrados (MMQ)

Para determinarmos os estimadores de MQO de β_0, β_1, ..., β_k, devemos minimizar

$$S=\sum_{i=1}^{n}(\varepsilon_i)^2=\varepsilon_1^2+\ldots+\varepsilon_n^2=\underset{\sim}{\boldsymbol{\varepsilon}}'\underset{\sim}{\boldsymbol{\varepsilon}}$$

ou, ainda,

$$S=\underset{\sim}{\boldsymbol{\varepsilon}}'\underset{\sim}{\boldsymbol{\varepsilon}}=\left(\underset{\sim}{\mathbf{y}}-\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}\right)'\left(\underset{\sim}{\mathbf{y}}-\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}\right)$$

Index of /wp-content/uploads/2014/02 (hedibert.org)

Estimação do MRLM: MMQ

Curiosidade

Abrindo a expressão anterior, vem que

$$S = \left(\underset{\sim}{\mathbf{y}} - \underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}\right)'\left(\underset{\sim}{\mathbf{y}} - \underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}\right) = \left(\underset{\sim}{\mathbf{y}}' - \underset{\sim}{\boldsymbol{\beta}}'\underset{\sim}{\mathbf{X}}'\right)\left(\underset{\sim}{\mathbf{y}} - \underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}\right) =$$

$$= \underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} - \underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}} - \underset{\sim}{\boldsymbol{\beta}}'\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{y}} + \underset{\sim}{\boldsymbol{\beta}}'\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}$$

[Index of /wp-content/uploads/2014/02 (hedibert.org)]

Estimação do MRLM: MMQ

Curiosidade

Como

$$\underset{\sim}{y}'\underset{\sim}{X}\underset{\sim}{\beta} \qquad \text{e} \qquad \underset{\sim}{\beta}'\underset{\sim}{X}'\underset{\sim}{y}$$

são escalares e

$$\underset{\sim}{y}'\underset{\sim}{X}\underset{\sim}{\beta} = \left(\underset{\sim}{\beta}'\underset{\sim}{X}'\underset{\sim}{y}\right)'$$

então

$$\underset{\sim}{y}'\underset{\sim}{X}\underset{\sim}{\beta} = \underset{\sim}{\beta}'\underset{\sim}{X}'\underset{\sim}{y}$$

Index of /wp-content/uploads/2014/02 (hedibert.org)

Estimação do MRLM: MMQ

Curiosidade

Assim

$$S = \underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} - 2\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\boldsymbol{\beta}}'\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}$$

Logo, nosso interesse, agora, é encontrar o resultado para

$$\frac{\partial S}{\partial \underset{\sim}{\boldsymbol{\beta}}}$$

21

[Index of /wp-content/uploads/2014/02 (hedibert.org)]

Estimação do MRLM: MMQ

Curiosidade

Lembrando que objetivamos minimizar

$$S\left(\underset{\sim}{\boldsymbol{\beta}}\right) = \underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}} - 2\,\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{X}}\,\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\boldsymbol{\beta}}'\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\,\underset{\sim}{\boldsymbol{\beta}}$$

e, utilizando os resultados vistos anteriormente, temos que

$$\frac{\partial S\left(\underset{\sim}{\boldsymbol{\beta}}\right)}{\partial \underset{\sim}{\boldsymbol{\beta}}} = -2\,\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{y}} + 2\,\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}}\,\underset{\sim}{\boldsymbol{\beta}}$$

Index of /wp-content/uploads/2014/02 (hedibert.org)

Estimação do MRLM: MMQ

Curiosidade

E, igualando o resultado anterior a zero, vem que

$$-2\,\underset{\sim}{X}'\underset{\sim}{y} + 2\,\underset{\sim}{X}'\underset{\sim}{X}\,\underset{\sim}{\hat{\beta}} = \underset{\sim}{0} \Leftrightarrow \underset{\sim}{X}'\underset{\sim}{X}\,\underset{\sim}{\hat{\beta}} = \underset{\sim}{X}'\underset{\sim}{y}$$

que é o sistema de equações normais na forma matricial.

Para encontrarmos o resultado de interesse, precisaremos supor que a matriz X'X admite inversa (ou seja, precisaremos supor que X'X é não-singular). Para tanto, assumiremos que os regressores não apresentam relação linear perfeita.

Index of /wp-content/uploads/2014/02 (hedibert.org)

Estimação do MRLM: MMQ

Assim, assumindo que X'X é não-singular, a solução do sistema de equações normais é dada por

$$\boxed{\hat{\underset{\sim}{\beta}} = \left(\underset{\sim}{X}'\underset{\sim}{X}\right)^{-1}\underset{\sim}{X}'\underset{\sim}{y}}$$

Nota:
X´é a matriz transposta da matriz X

que é o vetor de estimadores de mínimos quadrados do vetor de parâmetros de interesse.

Index of /wp-content/uploads/2014/02 (hedibert.org)

MRLS: MMQ em Notação Matricial

Regressão Linear Simples

Dada a equação:

$$Y_i = \alpha + \beta X_i + e_i$$

Que representa o sistema:

$$Y_1 = \alpha + \beta X_1 + e_1$$
$$Y_2 = \alpha + \beta X_2 + e_2$$
$$\cdots$$
$$Y_n = \alpha + \beta X_n + e_n$$

Para obter os estimadores de MQO:

$$EQT = \sum \hat{e}_i^2$$

e

$$\frac{\partial EQT}{\partial \hat{\alpha}} = 0 \Rightarrow \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\frac{\partial EQT}{\partial \hat{\beta}} = 0 \Rightarrow \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

A equivalente matricial:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Que representa o sistema:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix} = \alpha \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix} + \beta \begin{pmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{pmatrix} \Rightarrow \underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix}}_{\mathbf{y}_{n\times 1}} = \underbrace{\begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdots & \cdots \\ 1 & X_n \end{pmatrix}}_{\mathbf{X}_{n\times p}} \underbrace{\begin{pmatrix} \alpha \\ \beta \end{pmatrix}}_{\boldsymbol{\beta}_{p\times 1}} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{pmatrix}}_{\mathbf{e}_{n\times 1}}$$

Para obter os estimadores de MQO:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \Rightarrow \hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} \Rightarrow EQT = \hat{\mathbf{e}}^T \hat{\mathbf{e}}$$

onde $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}$

Então:

$$\frac{\partial EQT}{\partial \hat{\boldsymbol{\beta}}} = 0 \Rightarrow \boxed{\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})}$$

Nota:
p = k+1, sendo
p = nº de parâmetros a estimar
k = nº de variáveis

Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

Regressão Linear Múltipla

MRLM: MMQO em Notação Matricial

Dada a equação:

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e_i$$

Que representa o sistema:

$$Y_1 = \alpha + \beta_1 X_{1_1} + \beta_2 X_{2_1} + \ldots + \beta_k X_{k_1} + e_1$$
$$Y_2 = \alpha + \beta_1 X_{1_2} + \beta_2 X_{2_2} + \ldots + \beta_k X_{k_2} + e_2$$
$$\ldots$$
$$Y_n = \alpha + \beta_1 X_{1_n} + \beta_2 X_{2_n} + \ldots + \beta_k X_{k_n} + e_n$$

Para obter os estimadores de MQO:

$$EQT = \sum \hat{e}_i$$

e

$$\frac{\partial EQT}{\partial \hat{\alpha}} = 0 \Rightarrow \hat{\alpha} = \ldots$$
$$\ldots$$
$$\frac{\partial EQT}{\partial \hat{\beta}_k} = 0 \Rightarrow \hat{\beta}_k = \ldots$$

A equivalente matricial:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Que representa o sistema:

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix}}_{\mathbf{y}_{n\times 1}} = \underbrace{\begin{pmatrix} 1 & X_{1_1} & X_{2_1} & \ldots & X_{k_1} \\ 1 & X_{1_2} & X_{2_2} & \ldots & X_{k_2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{1_n} & X_{2_n} & \ldots & X_{k_n} \end{pmatrix}}_{\mathbf{X}_{n\times p}} \underbrace{\begin{pmatrix} \alpha \\ \beta_1 \\ \ldots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}_{p\times 1}} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix}}_{\mathbf{e}_{n\times 1}}$$

Para obter os estimadores de MQO:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \Rightarrow \hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} \Rightarrow EQT = \hat{\mathbf{e}}^T \hat{\mathbf{e}}$$

Então:

$$\frac{\partial EQT}{\partial \hat{\boldsymbol{\beta}}} = 0 \Rightarrow \boxed{\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})}$$

Nota:
p = k+1, sendo
p = nº de parâmetros a estimar
k = nº de variáveis

Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

Estimadores dos MMQ dos Coeficientes do MRLM

MODELO REGRESSÃO LINEAR $y_t = \beta_1 + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t$, $t = 1,2,\ldots,n.$

Formulário

EMQ (estimadores dos mínimos quadrados)

Caso particular: Regressão Linear Simples

Caso geral	Caso particular: $y_t = \beta_1 + \beta_2 x_t + u_t$
$b = (X^T X)^{-1} X^T Y$ $\hat{u}_t = y_t - \hat{y}_t$ $s^2 = \dfrac{\sum \hat{u}_t^2}{(n-k)}$ $\hat{Cov}(b \mid X) = s^2 (X^T X)^{-1}$	$b_1 = \bar{y} - b_2 \bar{x}$; $\hat{Var}(b_1 \mid X) = \dfrac{s^2 \sum x_t^2}{n \sum x_t^2 - \left(\sum x_t\right)^2}$ $b_2 = \dfrac{n \sum x_t y_t - \sum x_t \sum y_t}{n \sum x_t^2 - \left(\sum x_t\right)^2}$; $\hat{Var}(b_2 \mid X) = \dfrac{n s^2}{n \sum x_t^2 - \left(\sum x_t\right)^2}$ $s^2 = \dfrac{\sum \hat{u}_t^2}{(n-2)}$

Nota:
k = nº de parâmetros a estimar
k-1 = nº de variáveis

Murteira et al (2015)
Capítulo 10

1. Para cada uma das relações apresentados nas alíneas seguintes, indique as que são intrinsecamente lineares. Nos casos afirmativos, defina as variáveis e os parâmetros da relação linear respectiva.
   a) $z = \alpha + \beta e^{2w}$;
   b) $z = \alpha + \beta w + \gamma \ln(w)$ ($w > 0$);
   c) $z = 6\alpha + w^{\beta}$ ($w > 0$);
   d) $z = \alpha + \beta e^{w} + \gamma e^{-w}$.

Exercício 1

$z = \ln(w)$

Modelo linear: $y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K$

1) $z = \alpha + \beta e^{2w}$

Modelo intrinsecamente linear:

$y = z \qquad \beta_1 = \alpha$

$x_2 = e^{2w} \qquad \beta_2 = \beta$

Exercício 1

b) $z = \alpha + \beta W + \gamma \ln(W)$, $W > 0$

Modelo intrinsecamente linear:

$y = z$ $\qquad\qquad$ $\beta_1 = \alpha$

$x_2 = W$ $\qquad\qquad$ $\beta_2 = \beta$

$x_3 = \ln(W)$ $\qquad\qquad$ $\beta_3 = \gamma$

Exercício 1

c) Modelo não é intrinsecamente linear.

d) $z = \alpha + \beta e^{w} + \gamma e^{-w}$

Modelo intrinsecamente linear:

$y = z$ $\qquad\qquad$ $\beta_1 = \alpha$

$x_2 = e^{w}$ $\qquad\qquad$ $\beta_2 = \beta$

$x_3 = e^{-w}$ $\qquad\qquad$ $\beta_3 = \gamma$

7

Modelo de Regressão Linear Múltipla

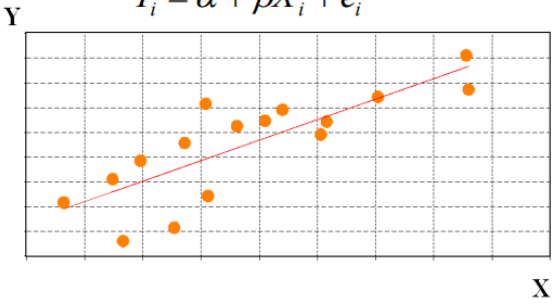
Interpretação dos Coeficientes da Reta de Regressão

MRLM: Interpretação dos Coeficientes

Caso particular: MRLM com apenas duas variáveis regressoras

Regressão Linear Simples:

$$Y_i = \alpha + \beta X_i + e_i$$



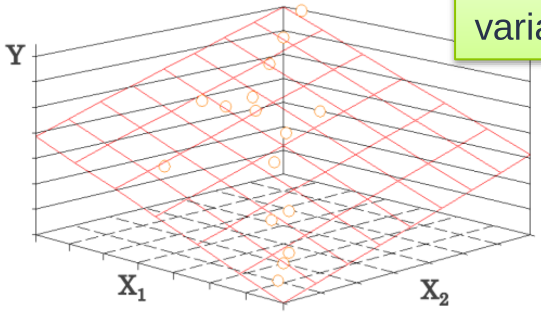
Temos que:

$E[Y / X = 0] = \alpha$ Valor esperado de *Y* quando *X* é nulo.

$\dfrac{dY}{dX} = \beta$ Variação marginal esperada em *Y* para cada variação unitária em *X*.

Regressão Linear Múltipla:

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i$$



Temos que:

$E[Y / X_1 = 0, X_2 = 0] = \alpha$ Valor esperado de *Y* quando ambos X_1 e X_2 são nulos.

$\dfrac{\partial Y}{\partial X_1} = \beta_1$ Variação marginal esperada em *Y* para cada variação unitária em X_1, *mantendo X_2 constante*.

$\dfrac{\partial Y}{\partial X_2} = \beta_2$ Variação marginal esperada em *Y* para cada variação unitária em X_2, *mantendo X_1 constante*.

Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

MRLM: Interpretação dos Coeficientes...

Caso geral: MRLM com k variáveis regressoras

Regressão Múltipla

Em um modelo de regressão múltipla, a variável dependente (*Y*) será determinada por mais de uma variável independente (*X*). Genericamente, um modelo de regressão linear múltipla com *k* variáveis independentes e *p* parâmetros ($p=k+1$) pode ser representado por:

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \ldots + \beta_k X_{k_i} + e_i$$

Onde:

α é o valor esperado de Y quando todos as variáveis independentes forem nulas;

β_1 é a variação esperada em Y dado um incremento unitário em X_1, mantendo-se constantes todas as demais variáveis independentes;

...

β_k é a variação esperada em Y dado um incremento unitário em X_k, mantendo-se constantes todas as demais variáveis independentes;

e_i é o erro não explicado pelo modelo;

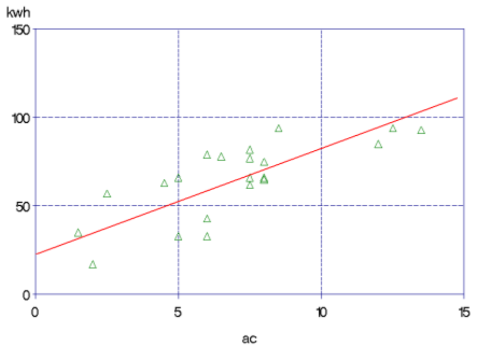
Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

MRLM: Exemplo 1 - Interpretação dos Coeficientes

Seja a relação para consumo de energia (*Kwh*), horas de ar condicionado ligado (*AC*) e horas de secadora ligada (*SEC*):

Regressão Linear Simples

$$Kwh_i = \alpha + \beta AC_i + e_i$$

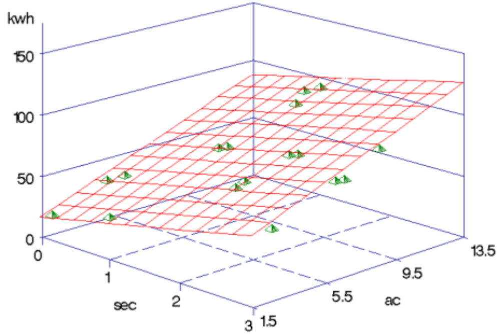


O coeficiente α indicará o consumo esperado de energia quando o ar condicionado permanecer desligado.

O coeficiente β indicará o consumo de energia adicional esperado para cada hora adicional com ar condicionado ligado.

Regressão Linear Múltipla

$$Kwh_i = \alpha + \beta_1 AC_i + \beta_2 SEC_i + e_i$$



O coeficiente α indicará o consumo esperado de energia quando ambos ar condicionado e secadora permanecerem desligados.

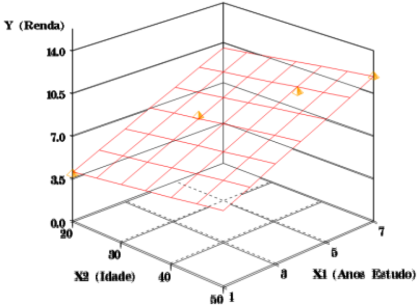
O coeficiente β_1 indicará o aumento no consumo de energia esperado para cada hora adicional com ar condicionado ligado, mantendo-se constante o tempo de uso da secadora. Analogamente, O coeficiente β_2 indicará efeito isolado de uma hora adicional com a secadora ligada sobre o consumo esperado de energia.

Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

MRLM: Exemplo 2 - Estimação dos Coeficientes e Interpretação

Seja a relação entre renda familiar em SM (Y), anos de estudo (X_1) e idade (X_2) do responsável pela família:

Y (Renda)	X_1 (Anos Estudo)	X_2 (Idade)
4	1	20
8	4	30
10	6	40
12	7	50



$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i \Rightarrow \mathbf{y} = \mathbf{X\beta} + \mathbf{e}$$

A função de regressão amostral será dada por:

$$\mathbf{y} = \mathbf{X\hat{\beta}} + \mathbf{\hat{e}} \Rightarrow \underbrace{\begin{pmatrix} 4 \\ 8 \\ 10 \\ 12 \end{pmatrix}}_{\mathbf{y}_{4\times1}} = \underbrace{\begin{pmatrix} 1 & 1 & 20 \\ 1 & 4 & 30 \\ 1 & 6 & 40 \\ 1 & 7 & 50 \end{pmatrix}}_{\mathbf{X}_{4\times3}} \underbrace{\begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}}_{\hat{\boldsymbol{\beta}}_{3\times1}} + \underbrace{\begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \hat{e}_3 \\ \hat{e}_4 \end{pmatrix}}_{\hat{\mathbf{e}}_{4\times1}}$$

E as estimativas de MQO:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y}) = \begin{pmatrix} 4 & 18 & 140 \\ 18 & 102 & 730 \\ 140 & 730 & 5400 \end{pmatrix}^{-1} \begin{pmatrix} 34 \\ 180 \\ 1320 \end{pmatrix} = \begin{pmatrix} 1{,}9 \\ 1 \\ 0{,}06 \end{pmatrix}$$

Maia, Alexandre Gori (2017). Econometria: conceitos e aplicações. Cap. 6.

Espera-se, para cada ano adicional de estudo do responsável pela família, um aumento de 1 SM e, para cada ano de idade adicional, um aumento de 0,06 SM na renda familiar.

O departamento de RH da empresa TEMCO objetiva estudar o comportamento dos salários dos funcionários dos mais diversos setores da empresa.

Para tanto, o gerente de RH, baseando-se numa amostra aleatória de 46 empregados, coletou informações sobre as seguintes variáveis:

- **id** – número cadastral do funcionário;
- **salario** – anual, em dólares;
- **anosemp** – tempo (em anos) na empresa;
- **expprev** – experiência anterior (em anos);
- **educ** – anos de estudo após o segundo grau;
- **sexo** – (feminino = 0, masculino = 1);
- **dept** – departamento no qual atua (Compras = 1, Engenharia = 2, Propaganda = 3, Vendas = 4);
- **super** – número de empregados sob responsabilidade do empregado.

EViews - [Group: UNTITLED Workfile: TEMCO::Temco\]

File Edit Object View Proc Quick Options Window Help

View Proc Object Print Name Freeze Default Sort Transpose Edit+/- Smpl+/- Title Sample

obs	ID	SALARIO	ANOSEMP	EXPPREV	EDUC	SEXO	DEPT	SUPER
1	972.0000	47536.00	15.00000	5.000000	6.000000	0.000000	3.000000	4.000000
2	539.0000	23654.00	0.000000	0.000000	0.000000	1.000000	3.000000	2.000000
3	649.0000	37548.00	19.00000	9.000000	4.000000	0.000000	3.000000	6.000000
4	824.0000	36578.00	4.000000	4.000000	8.000000	0.000000	3.000000	8.000000
5	649.0000	54679.00	20.00000	3.000000	6.000000	1.000000	3.000000	4.000000
6	624.0000	53234.00	25.00000	0.000000	6.000000	0.000000	3.000000	3.000000
7	891.0000	31425.00	7.000000	6.000000	5.000000	1.000000	3.000000	6.000000
8	974.0000	39743.00	9.000000	6.000000	5.000000	1.000000	2.000000	1.000000
9	648.0000	26452.00	1.000000	3.000000	2.000000	1.000000	2.000000	0.000000
10	321.0000	34632.00	5.000000	4.000000	4.000000	0.000000	2.000000	0.000000
11	264.0000	35631.00	6.000000	4.000000	4.000000	0.000000	2.000000	2.000000
12	291.0000	46211.00	14.00000	5.000000	6.000000	1.000000	2.000000	5.000000
13	267.0000	34231.00	6.000000	2.000000	6.000000	0.000000	2.000000	3.000000
14	548.0000	26548.00	5.000000	1.000000	0.000000	0.000000	2.000000	2.000000
15	555.0000	36512.00	6.000000	6.000000	4.000000	1.000000	2.000000	2.000000
16	366.0000	34869.00	7.000000	5.000000	4.000000	1.000000	2.000000	1.000000
17	246.0000	41255.00	9.000000	4.000000	6.000000	0.000000	2.000000	4.000000
18	215.0000	39331.00	9.000000	3.000000	6.000000	1.000000	2.000000	1.000000
19	814.0000	35487.00	8.000000	2.000000	2.000000	1.000000	2.000000	2.000000
20	212.0000	36487.00	6.000000	5.000000	2.000000	0.000000	2.000000	3.000000
21	526.0000	68425.00	25.00000	2.000000	12.00000	0.000000	2.000000	1.000000
22	778.0000	69246.00	22.00000	3.000000	10.00000	0.000000	2.000000	45.00000

Quadro 1 - Parte de uma planilha que contem informações sobre os empregados da empresa TEMCO.

Como parte do estudo, a gerente de RH propôs a estimação dos parâmetros do seguinte modelo de regressão múltipla:

$$salario = \beta_0 + \beta_1 educ + \beta_2 anosemp + \varepsilon$$

a) Em termos do problema, β_0 apresenta algum significado prático?

b) Qual o sinal esperado para β_1? E para β_2?

c) Encontre as estimativas dos parâmetros, via mínimos quadrados ordinários, escreva a equação estimada e interprete os resultados obtidos, em termos do problema de interesse.

Exercícios (a) e (b): Interpretação dos Parâmetros

Interpretação dos parâmetros do modelo proposto, em termos do problema:

- β_0 – salário médio dos funcionários da empresa TEMCO, que acabaram de entrar na empresa (ou que ainda não completaram um ano) e que não apresentam nenhum ano de escolaridade após o segundo grau;
- β_1 – efeito no salário médio dos funcionários da empresa TEMCO, dada a variação de um ano no tempo de escolaridade após o segundo grau, mantendo constante a variável *anosemp*; e
- β_2 – efeito no salário médio dos funcionários da empresa TEMCO, dada a variação de um ano no tempo de empresa, mantendo constante a variável *educ*.

Exercícios (a) e (b): Interpretação dos Parâmetros

Dependent Variable: SALARIO
Method: Least Squares
Date: 08/26/12 Time: 15:45
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*EDUC+C(3)*ANOSEMP

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	23177.47	1769.732	13.09660	0.0000
C(2)	1916.489	379.2670	5.053139	0.0000
C(3)	672.3250	141.6725	4.745629	0.0000

R-squared	0.739927	Mean dependent var	39827.39
Adjusted R-squared	0.727830	S.D. dependent var	10999.24
S.E. of regression	5738.291	Akaike info criterion	20.21070
Sum squared resid	1.42E+09	Schwarz criterion	20.32996
Log likelihood	-461.8462	Hannan-Quinn criter.	20.25538
F-statistic	61.16907	Durbin-Watson stat	1.229794
Prob(F-statistic)	0.000000		

Exercício (c): Modelo Estimado

$$\hat{salário} = 23177{,}47 + 1916{,}49\,educ + 672{,}32\,anosemp$$

Pergunta: qual o salário médio estimado para pessoas com 3 anos de escolaridade após o 2º grau e com 5 anos na empresa?

$$\hat{salario} = 23.177{,}47 + 1.916{,}49 * 3 + 672{,}33 * 5$$

$$\hat{salario} = 32288{,}54$$

Murteira et al (2015)
Capítulo 10

10. A 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 est guir.

MODELO REGRESSÃO LINEAR $y_t = \beta_1 + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t$, $t = 1,2,\ldots,n$.

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

Murteira et al (2015)
Capítulo 10

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.

Exercício 10 a)

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u_t$$

y = consumo familiar de energia eléctrica (em u. m.)

x_2 = rendimento familiar

x_3 = nº de indivíduos na família

x_4 = area do fogo (em m^2)

$n = 10 \qquad K = 4$

- Nulidade conjunta: $F = \dfrac{R^2/(k-1)}{(1-R^2)/(n-k)} = \dfrac{VE/(k-1)}{VR/(n-k)} \sim F(k-1, n-k)$

Exercício 10 a)

a) $H_0: \beta_2 = \beta_3 = \beta_4 = 0$ vs $H_1: \exists j: \beta_j \neq 0 \ (j = 2,3,4) \ (\alpha = 0.05)$

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)} \sim F(K-1, n-K) = F(3,6)$$

$$f_{obs} = \frac{0.870462864 / 3}{(1-0.870462864)/6} = 13.4396$$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.932985993
R Square (R^2)	0.870462864
Adjusted R Square	0.805694295
Standard Error	4.571741129
Observations (n)	10

$$W_F = \{ f_{obs} : f_{obs} > F_{3,6,0.05} = 4.76 \}$$

$f_{obs} \in W_F$ logo rejeitamos H_0.

Os regressores incluídos no modelo são estatísticamente significativos (conjuntamente) ao vível de 5%.

Exercício 10 b) and c)

b) $\alpha = 0.05$

	Coefficients	Standard Error	t Stat	p-value		Estat. sig. a 5% ?
Intercept	12.47998124	9.282257233	1.344498534	0.227386	> 0.05	não
x_2	0.060527128	0.024073045	2.514311292	0.045637	< 0.05	sim
x_3	−2.81451648	2.679535456	−1.050374786	0.334	> 0.05	não
x_4	0.020535759	0.057861717	0.354910983	0.734798	> 0.05	não

Observando os valores-p dos testes-t à significância estatística individual dos coeficientes estimados, apenas o regressor referente ao rendimento familiar é estatisticamente significativo ao nível de 5%.

c) O modelo foi estimado com base num numero muito pequeno de observações (n=10), logo as estimativas resultantes e respetiva inferência não devem ser levadas muito a sério.

Exercício 10 d)

Inferência estatística do MRL, com $y_t \mid X \sim N(x_{t\bullet}\beta, \sigma^2)$:

- $q = \dfrac{\sum_{t=1}^{n} \hat{u}_t^2}{\sigma^2} = \dfrac{(n-k)s^2}{\sigma^2} \sim \chi^2(n-k)$
- $t_j = \dfrac{b_j - \beta_j}{s_{b_j}} \sim t(n-k)$ ou $F_j = \dfrac{(b_j - \beta_j)^2}{s_{b_j}^2} \sim F(1, n-k)$

d) $H_0: \beta_2 \leqslant 0 \qquad H_1: \beta_2 > 0 \qquad (\alpha = 0.05)$

ou $H_0: \beta_2 = 0$ (teste c/ a mesma região crítica)

$$T = \frac{b_2 - \beta_2}{s_{b_2}} \sim t(n-k) = t(6)$$

$$t_{obs} = \frac{0.060527128 - 0}{0.024073045} = 2.514311292$$

$$W_T = \{ t_{obs} : t_{obs} > t_{6,\,0.05} = 1.943 \}$$

$t_{obs} \in W_T$ logo rejeita-se H_0. O aumento do rendimento parece implicar, em média, um aumento do consumo de energia eléctrica.

Murteira et al (2015)
Capítulo 10

14. Numa primeira tentativa para modelar a distribuição regional (18 distritos mais as duas regiões autónomas) das vendas de automóveis ligeiros de passageiros em Portugal em determinado ano obteve-se:

$$\hat{y}_t = -9767.66 + \underset{(0.0145)}{0.118}\, x_t \quad (t = 1, 2, \ldots, 20),$$

onde y representa as vendas em milhares de unidades e x a população activa de cada distrito, em milhares. Entre parênteses apresenta-se o erro padrão (estimativa do desvio padrão do estimador MQ do respectivo coeficiente de regressão).

a) Analise a adequação estatística do modelo, supondo que a dimensão dos testes a efectuar é $\alpha = 0.1$.

b) Construa um intervalo de confiança a 90% para o declive da regressão.

c) Teste a hipótese do declive ser igual a 0.10, supondo que $\alpha = 0.05$.

Exercício 14 a)

$y \equiv$ Vendas (em milhares de €)

$x \equiv$ População ativa do distrito (em milhares)

$$\hat{y}_t = \underset{}{-9767.66} + \underset{(0.0145)}{0.118}\, x_t \qquad (t = 1, \cdots, 20)$$

Exercício 14 a)

Inferência estatística do MRL, com $y_t \mid X \sim N(x_{t\bullet}\beta, \sigma^2)$:

- $q = \frac{\sum_{t=1}^{n} \hat{u}_t^2}{\sigma^2} = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2(n-k)$
- $t_j = \frac{b_j - \beta_j}{s_{b_j}} \sim t(n-k)$ ou $F_j = \frac{(b_j - \beta_j)^2}{s_{b_j}^2} \sim F(1, n-k)$

a) $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$ $(\alpha = 0.01)$

sob H_0, $T = \frac{b_2 - \beta_2}{s_{b_2}} = \frac{b_2}{s_{b_2}} \sim t(n-k) = t(18)$

$t_{obs} = \frac{0.118}{0.0145} = 8.1379$

$W_T = \{ t_{obs} : |t_{obs}| > t_{18, 0.005} = 2.878 \}$

$t_{obs} \in W_T \Rightarrow$ Rejeita-se H_0.

O efeito estimado do tamanho da população ativa no valor das vendas é estatisticamente significativo.

Exercício 14 b)

b) $1-\alpha = 0.9$ $\qquad \alpha = 0.1$ $\qquad \frac{\alpha}{2} = 0.05$

$IC_{90\%}$ para β_2 : $\quad b_2 \mp t_{18,0.05}\, s_{b_2} =$

$= 0.118 \mp 1.734 \times 0.0145 =$

$= (0.0929 ; 0.1431)$

Exercício 14 c)

Inferência estatística do MRL, com $y_t \mid X \sim N(x_{t\bullet}\beta, \sigma^2)$:

- $q = \dfrac{\sum_{t=1}^{n} \hat{u}_t^2}{\sigma^2} = \dfrac{(n-k)s^2}{\sigma^2} \sim \chi^2(n-k)$
- $t_j = \dfrac{b_j - \beta_j}{s_{b_j}} \sim t(n-k)$ ou $F_j = \dfrac{(b_j - \beta_j)^2}{s_{b_j}^2} \sim F(1, n-k)$

c) $H_0: \beta_2 = 0.1$ $\quad H_1: \beta_2 \neq 0.1$ $\quad (\alpha = 0.05)$

sob H_0, $T = \dfrac{b_2 - \beta_2}{s_{b_2}} = \dfrac{b_2 - 0.1}{s_{b_2}} \sim t(18)$

$$t_{obs} = \frac{0.118 - 0.1}{0.0145} = 1.2414$$

$$W_T = \{ t_{obs} : |t_{obs}| > t_{18, 0.025} = 2.101 \}$$

$t_{obs} \notin W_T$ logo não se rejeita H_0.

Obrigada!

Questões?