Estatística II

Licenciatura em Gestão

2.º Ano/1.º Semestre

2023/2024

Aulas Teóricas N.ºs 18 e 19 (Semana 10)

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 Teóricas (Semanas 1 a 5)	Aulas Teóricas (Semanas 5 a 7)	Aulas Teóricas (Semanas 7 a 9)	Aulas Teóricas (Semanas 10 a 13)
• **Capítulo 1**: Estimação	• **Capítulo 2**: Testes de Hipóteses	• **Capítulo 3**: Modelo de Regressão Linear	• **Capítulo 4**: Complementos ao Modelo de Regressão Linear

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.

7ª semana (31/10 a 02/11)

T12 - Teste de hipóteses

Testes em universos normais com amostras emparelhadas. Exemplo. Teste de hipóteses para grandes amostras. Aplicação ao universo de Bernoulli (média e diferença de médias). Exemplos.

T13 - Modelo de regressão linear

Introdução; modelo linear e linearizável; exemplos; Hipóteses básicas; estimação dos coeficientes da regressão pelos Mínimos Quadrados. Exemplo.

8ª semana (07/11 e 09/11)

T14 - Modelo de Regressão Linea (MRL)r

Interpretação dos parâmetros da regressão; exemplos; Resíduos MQ e regressão ajustada; Propriedades dos estimadores MQ dos coeficientes da regressão; Estimador não enviesado da variância da variável residual; Exemplo.

T15 - Modelo de regressão Linear

Coeficiente de determinação e sua interpretação. Hipótese adicional (H6) e inferência estatística sobre o modelo; Inferência sobre um parâmetro beta. Exemplos

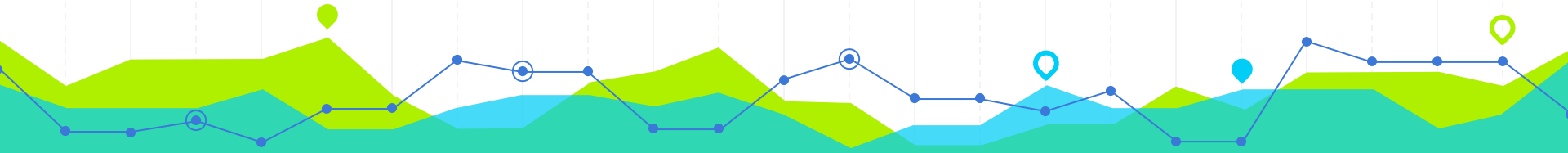
9ª semana (14/11 e 16/11)

T16 - Modelo de Regressão Linear

Mais exemplos de inferência sobre um parâmetro beta; Inferência sobre uma combinação linear de betas; exemplos.

T17 - Modelo de Regressão Linear

Teste de nulidade conjunta de vários coeficientes; exemplo; Teste F à significância global da regressão; Teste de um conjunto de restrições lineares; exemplo.



Teste de Hipóteses de Ajustamento do Q-Q

1

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")

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")

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

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

Estatística de teste

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

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 $\hat{p}_{\circ j}$: $\chi^2_{(m-k-1)}$

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=$(-8,67)^2/53,67+0,33^2/53,67$
$+8,33^2/53,67 = 2,695$

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-α 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)

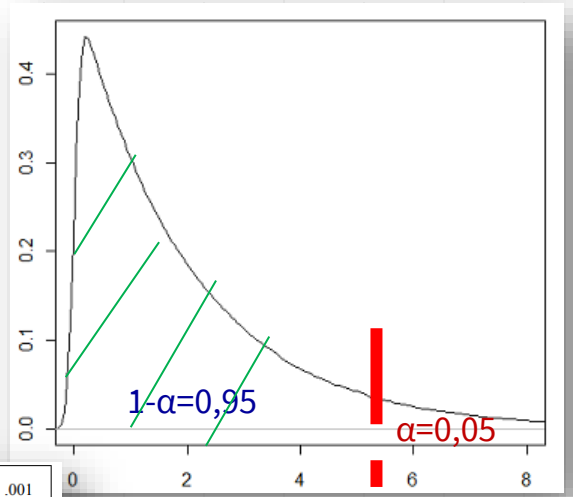
Área total é igual a 1

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

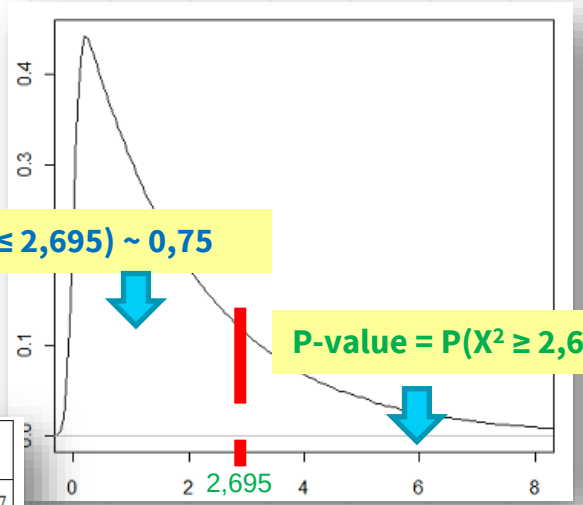
Cálculo do Valor-p quando a Estatística de Teste tem Distribuição Qui-Quadrado

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

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

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

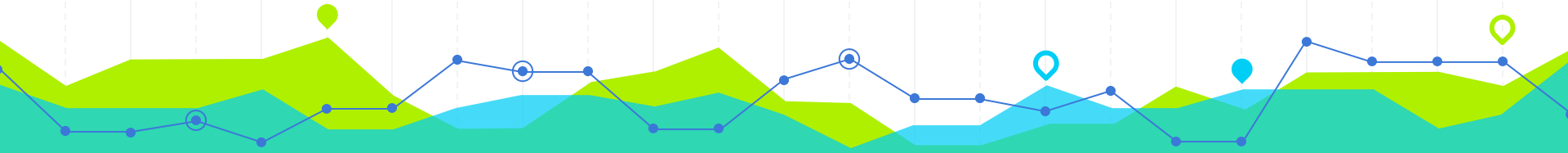
Área total é igual a 1



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

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

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



Teste de Hipóteses de Independência do Q-Q

Hipóteses, Estatística de Teste e Decisão

2

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

Hipóteses

H_0: As variáveis são independentes

Versus

H_1: As variáveis não são independentes

Estatística de teste

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

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

Frequências esperadas

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

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

Tabela da distribuição do Qui-Quadrado

Decisão

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

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

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			Total
			Fábrica A	Fábrica B	Fábrica C	
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

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

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

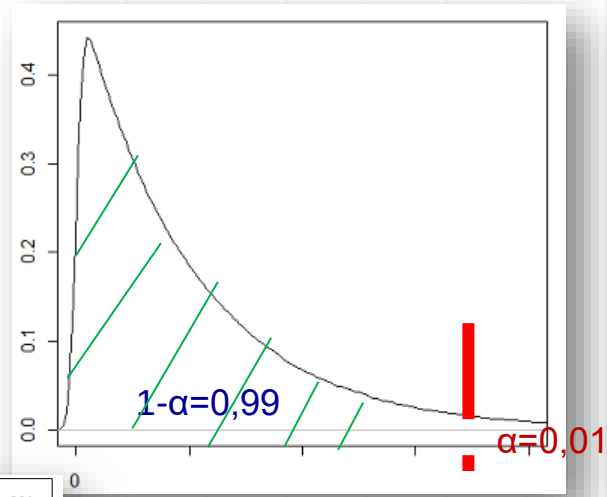
Nível de confiança (1-α=0,99)

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)



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

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

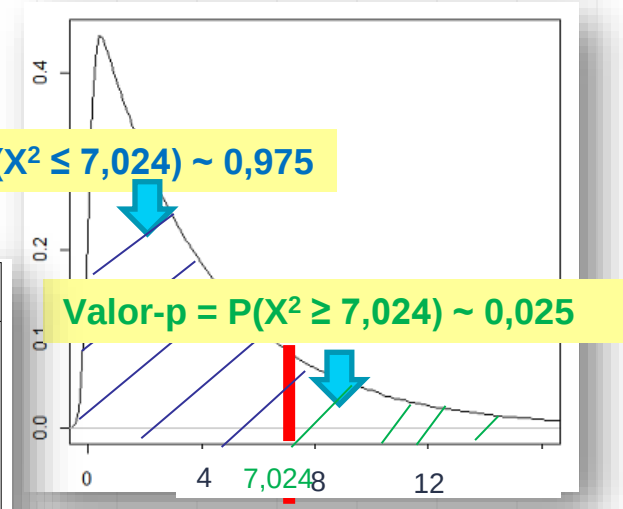
Cálculo do Valor-p quando a Estatística de Teste tem Distribuição Qui-Quadrado

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

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

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



$P(X^2 \leq 7{,}024) \sim 0{,}975$

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

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

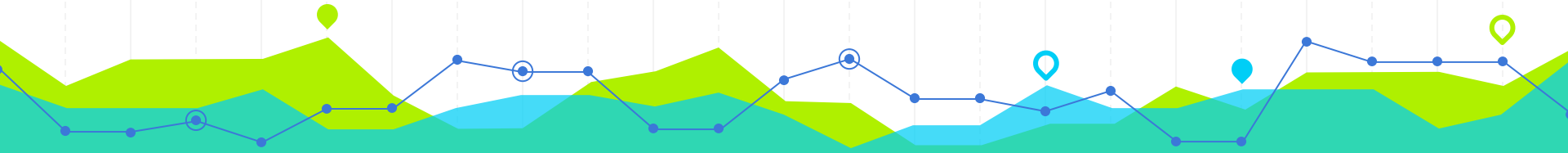
Verificação das condições de aplicabilidade

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

Contaminado * Fábrica Tabulação cruzada

			Fábrica			Total
			Fábrica A	Fábrica B	Fábrica C	
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

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.



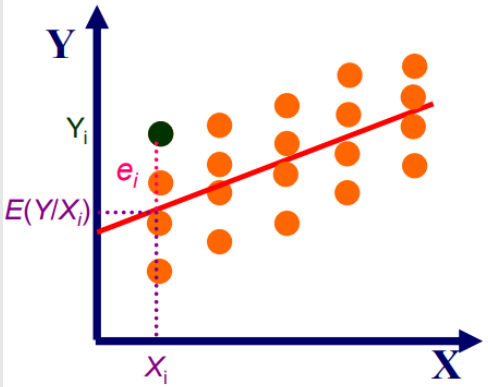
Modelo de Regressão Linear Múltipla

3

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*

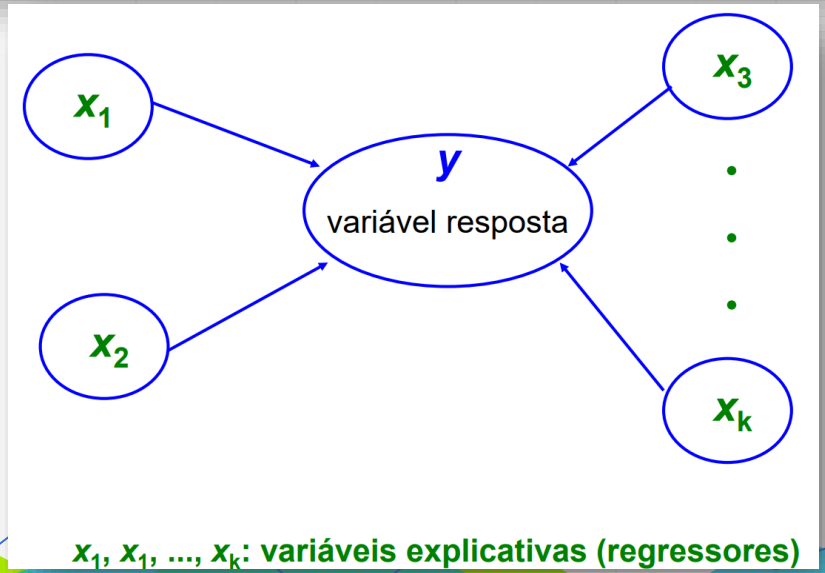
$E(Y/X_i)$ é a esperança condicional de *Y* e representa o valor esperado de *Y* para o *i*-ésimo valor de *X*

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.



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,

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

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

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

Para determinarmos os estimadores de mínimos quadrados de $\beta_0, \beta_1, \ldots, \beta_k$, 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

$$S(\beta_0,\ \beta_1,\ \ldots,\beta_k)=\sum_{i=1}^{n}(\varepsilon_i)^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{1i}-\cdots-\beta_kx_{ki})^2$$

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

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

Estimação dos Coeficientes do MRLM: MMQ

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

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

$$\vdots$$

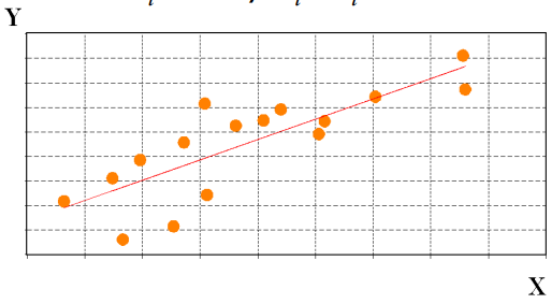
$$\frac{\partial}{\partial \beta_k} S(\beta_0, \beta_1, \ldots, \beta_k) = -2\sum_{i=1}^{n}\left[\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right) x_{ki}\right] = 0$$

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:

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

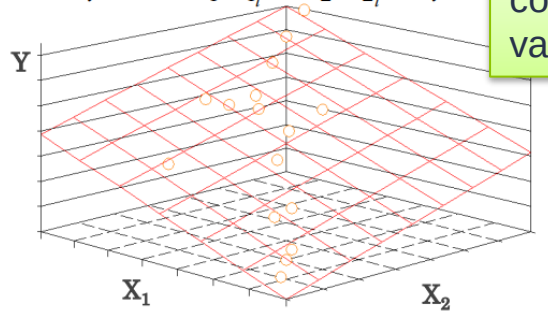
Minimizando EQT:

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

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

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

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

também pode ser escrita como

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \ldots + \beta_k x_{k1} + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \ldots + \beta_k x_{k2} + \varepsilon_2$$
$$y_3 = \beta_0 + \beta_1 x_{13} + \beta_2 x_{23} + \ldots + \beta_k x_{k3} + \varepsilon_3$$
$$\vdots$$
$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \ldots + \beta_k x_{kn} + \varepsilon_n$$

MRLM: Abordagem Matricial

As igualdades anteriores podem ser alocadas facilmente em dois vetores colunas (n×1), descritos a seguir:

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

MRLM: Abordagem Matricial

Ainda,

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

MRLM: Abordagem Matricial

Finalmente,

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

MRLM: Abordagem Matricial

Vamos definir:

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

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

15

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.

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,

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

Estimação do MRLM: MMQ

Curiosidade

Abrindo a expressão anterior, vem que

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

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

Estimação do MRLM: MMQ

Curiosidade

Como

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

são escalares e

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

então

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

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

Estimação do MRLM: MMQ

Curiosidade

Lembrando que objetivamos minimizar

$$S(\underset{\sim}{\beta}) = \underset{\sim}{y}'\underset{\sim}{y} - 2\underset{\sim}{y}'\underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\beta}'\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{\beta}$$

e, utilizando os resultados vistos anteriormente, temos que

$$\frac{\partial S(\underset{\sim}{\beta})}{\partial \underset{\sim}{\beta}} = -2\underset{\sim}{X}'\underset{\sim}{y} + 2\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{\beta}$$

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

Estimação do MRLM: MMQ

Curiosidade

E, igualando o resultado anterior a zero, vem que

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

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.

Estimação do MRLM: MMQ

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

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

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

Nota:
X´é a matriz transposta da matriz X

MRLS: MMQ em Notação Matricial

Regressão Linear Simples

Dada a equação:

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

Que representa o sistema:

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

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:

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

Para obter os estimadores de MQO:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \Rightarrow \hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} \Rightarrow EQT = \hat{\mathbf{e}}^T\hat{\mathbf{e}} \quad \text{onde} \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}$$

Então:

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

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:

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

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:

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

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

Formulário

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

EMQ (estimadores dos mínimos quadrados)

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

Caso particular: Regressão Linear Simples

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

Obrigada!

Questões?