

Regressão linear múltipla

- Intuitivamente, pode-se melhorar modelo (predição) se **incluirmos novas variáveis** independentes no modelo de regressão (embora “parcimoniosamente”).
- Os conceitos e técnicas adoptados são uma **extensão natural** do que foi apresentado no **capítulo anterior**.
- Na RLM assume-se que existe uma relação linear entre uma **variável Y** (var. dependente) e **k variáveis independentes**, x_j ($j=1,2,\dots,k$), ou variáveis explicativas ou *regressores*.
- As **condições subjacentes são análogas** às da RLS.

(c) Eduardo Esteves, 2009

Modelo de regressão linear

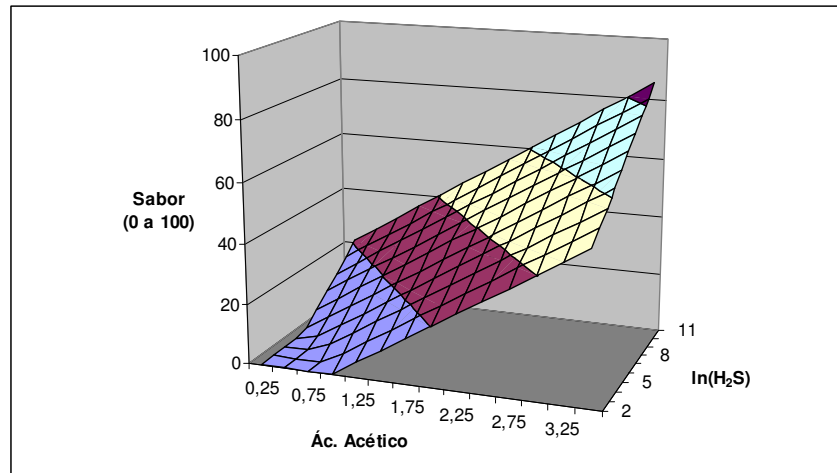
- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$
- Os $p=(k+1)$ parâmetros, β_j , são os coeficientes (parciais) de regressão e ε é o erro aleatório.
- Este modelo descreve um hiperplano no espaço k -dimensional dos regressores $\{x_j\}$...

- “Matricialmente”: $Y = X\beta + \varepsilon$

Vector coluna	Matriz	Vector coluna	Vector coluna
$(n \times 1)$	$(n \times p)$	$(p \times 1)$	$(n \times 1)$

(c) Eduardo Esteves, 2009

...hiperplano no espaço k -dimensional dos regressores $\{x_j\}$...



(c) Eduardo Esteves, 2009

Outras regressões múltiplas...

- Regressão quadrática/polinomial
- $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ em que $x_2 = x_1^2$

- Utilização de variáveis *mudas* (*dummy*)
- $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ em que $x_2 = (0,1)$, $x_3 = (0,1)$

- Regressão com interações (produtos cruzados)
- $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ em que $x_3 = x_1 \cdot x_2$

(c) Eduardo Esteves, 2009

MMQ...

- Dados amostrais...

- Minimizando $\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

- Estimar $\hat{\beta} = B = (X^T X)^{-1} X^T Y$

- Onde se obtém: $\hat{Y} = X\hat{\beta}$ (em que $e = Y - \hat{Y}$)

(c) Eduardo Esteves, 2009

Significância do modelo de regressão (ANOVA)

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0 \text{ versus } H_1 : \beta_j \neq 0 \quad j = 1, \dots, k$$

$$SQ_T = SQ_R + SQ_E$$

e.t.

Fonte de variação	Graus de liberdade ¹²	Soma de Quadrados	Média quadrática	F_0
Regressão ou do modelo	k	SQ_R	MQ_R	$\frac{MQ_R}{MQ_E}$
Erro ou residual	$n - p$	SQ_E	MQ_E	
Total	$n - 1$	SQ_T		

Tabela 2.4 – Tabela da ANOVA para a regressão linear múltipla.

Se $f_0 > f_{[1-\alpha; k; (n-p)]}$ ou se valor- $p < \alpha \Rightarrow$ Rej. H_0

(c) Eduardo Esteves, 2009

O R^2 e o R^2 ajustado

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T}$$

$$R^2_{\text{ajust.}} = 1 - \frac{\frac{SQ_E}{n-1}}{\frac{SQ_T}{n-p}} = 1 - \left(\frac{n-1}{n-p} \right) (1 - R^2)$$

Este coeficiente dá uma melhor ideia da proporção de variação de Y explicada pelo modelo de regressão uma vez que tem em conta o n° de regressores (i.e. variáveis independentes).

(c) Eduardo Esteves, 2009

Regressão estandardizada

- Para ter uma ideia da contribuição relativa dos regressores para a explicação da variação de Y usar a equação de regressão estandardizada, *e.g.*

- $y' = \beta'_1 x'_1 + \beta'_2 x'_2$

- Os coeficientes obtêm-se a partir de

- $\beta'_j = \beta_j \frac{s_{x_j}}{s_y}$

(c) Eduardo Esteves, 2009

Teste(s) de significância de β_j

Hipóteses: $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$

$$(e.t.) \quad T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}} \quad C = (X^T X)^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix}$$

Região de rejeição: $|t_0| > t_{\alpha/2} [n - p]$ ou *valor-p* $< \alpha$

Se H_0 não for rejeitada, isto indica que o regressor x_j pode ser “eliminado” do modelo...

(c) Eduardo Esteves, 2009

Seleção de variáveis/modelos

Para além da inclusão simultânea todos os regressores:

- Método exaustivo (*all possible regressions*)...
- Método progressivo (*forward selection*)...
- Método regressivo (*backward selection*)...
- Método *passo-a-passo* (*stepwise*)

(c) Eduardo Esteves, 2009

IC e IP

- IC para os coeficientes de regressão:

$$\hat{\beta}_j \pm t_{[1-\alpha/2; n-2]} \cdot se(\hat{\beta}_j)$$

- IC para \hat{Y} para um dado \mathbf{x}_0 :

$$\hat{\mu}_{Y|x_0} \pm t_{[1-\alpha/2; n-p]} \sqrt{\hat{\sigma}^2 \mathbf{x}_0^T \mathbf{C} \mathbf{x}_0}$$

- IP para \hat{Y} para um dado \mathbf{x}_0 :

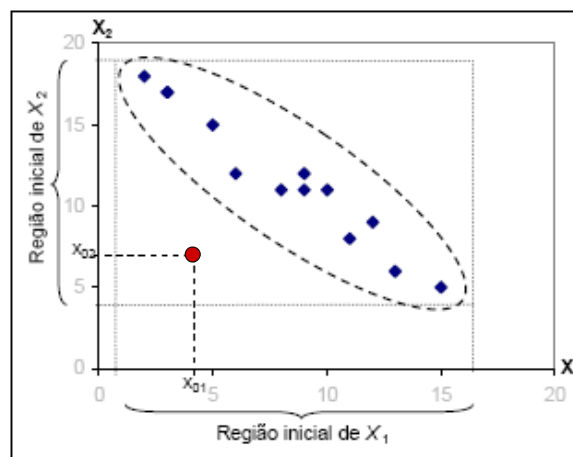
$$\hat{y}_0 \pm t_{[1-\alpha/2; n-p]} \sqrt{\hat{\sigma}^2 \left[1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0 \right]}$$

$$\hat{\mu}_{Y|x_0} = \mathbf{x}_0^T \hat{\boldsymbol{\beta}} = \hat{y}_0$$

$$\mathbf{x}_0^T = [1 \quad x_{01} \quad x_{02} \quad \dots \quad x_{0k}]$$

(c) Eduardo Esteves, 2009

Extrapolação: *Atenção!*



(c) Eduardo Esteves, 2009

Colinearidade: *Atenção!*

Cuidado com valores $|R| > 0,75$.

		Correlations ^a						
		CHOL CHOLESTE ROL LEVEL	AGE AGE IN YEARS	CALCIUM	ACID URIC ACID	ALB	WEIGHT WEIGHT IN POUNDS	WTALB
CHOL CHOLESTEROL LEVEL	Pearson Correlation Sig. (2-tailed)	1	.365** .000	.255** .001	.274** .000	.057 .443	.146 .050	.148* .048
AGE AGE IN YEARS	Pearson Correlation Sig. (2-tailed)	.365** .000	1	-.009 .908	.209** .005	-.079 .292	.255** .001	.255** .001
CALCIUM	Pearson Correlation Sig. (2-tailed)	.255** .001	-.009 .908	1	.166* .026	.453** .000	.065 .388	.073 .331
ACID URIC ACID	Pearson Correlation Sig. (2-tailed)	.274** .000	.209** .005	.166* .026	1	.030 .690	.304** .000	.305** .000
ALB	Pearson Correlation Sig. (2-tailed)	.057 .443	-.079 .292	.453** .000	.030 .690	1	-.235** .002	-.218** .003
WEIGHT WEIGHT IN	Pearson Correlation Sig. (2-tailed)	.146 .050	.255** .001	.065 .388	.304** .000	-.235** .002	1	1.000** .000
WTALB	Pearson Correlation Sig. (2-tailed)	.148* .048	.255** .001	.073 .331	.305** .000	-.218** .003	1.000** .000	1

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

a. Listwise N=180

http://www-personal.umich.edu/~kwelch/510/2008/handouts/spss_multiple_regression_lecture_2008.doc

(c) Eduardo Esteves, 2009

Colinearidade: *Variance Inflation Factor**

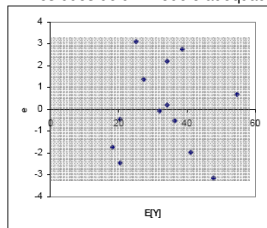
- Medida para diagnóstico da multicolinearidade
- $VIF = 1/(1 - R_i^2)$
- Quanto maior for VIF , maior a correlação múltipla entre variáveis independentes (i.e. mais grave a multicolinearidade).
- $VIF > 5$ ou $VIF > 10$ (depende dos autores) indicam problemas com estimação de β_1 devido à multicolinearidade.

* Maroco (2003) pp. 415-418.

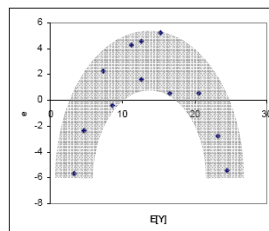
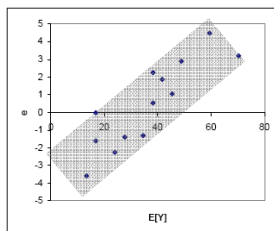
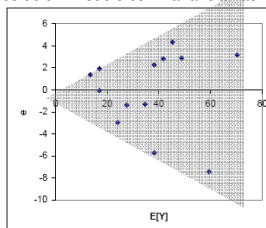
(c) Eduardo Esteves, 2009

Análise (gráfica) de resíduos

■ Resíduos de um modelo adequado.



■ Resíduos de um modelo com variância não-homogênea.



■ Resíduos resultantes dum erro de análise (e.g. falta b_0).

■ Resíduos de um modelo (linear) não-adequado.

(c) Eduardo Esteves, 2009

Exemplo

Ensaio	Humidade (%)	Doçura (mg/L)	Apreciação (0 a 100)
1	6	10,1	64
2	10	13,0	73
3	5	10,5	61
4	8	11,4	70
5	4	14,3	75
6	6	13,5	81
7	9	16,2	88
8	7	18,7	95
9	4	19,2	94
10	5	18,1	99
11	10	17,5	97

(c) Eduardo Esteves, 2009