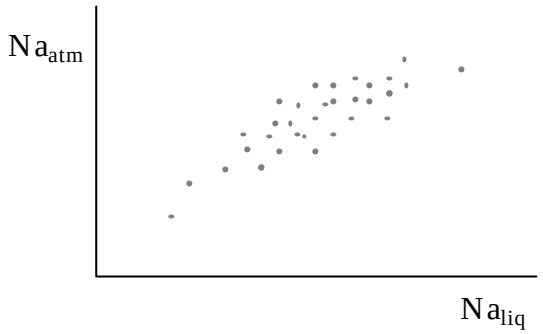
REGRESSÃO LINEAR



À variável Y cujo comportamento se pretende estudar dá-se o nome de variável dependente.

O comportamento desta variável depende de outras variáveis X_j chamadas variáveis independentes.

A modelação do comportamento da variável dependente (ou da esperança da variável dependente) quando os valores das variáveis independentes variam é um problema comum.

Um grupo importante de modelos é constituído pelos modelos lineares nos parâmetros:

$$Y = b + \sum_j a_j \, X_j$$

⇑ ⇑

parâmetros

Outro grupo importante de modelos é o grupo dos modelos linearizáveis.

MODELO LINEAR SIMPLES
(2 Variáveis)

$$E\{Y|X = x\} = \alpha + \beta\, x$$

α – ordenada na origem; valor de $E\{Y|x = 0\}$

β – tangente da recta

Cada valor individual pode ser tomado como:

$$y_i = \alpha + \beta\, x_i + \varepsilon_i$$

Os erros aleatórios (ou desvios) têm esperança nula e uma variância comum σ^2 e são independentes.

REGRESSÃO LINEAR

Objectivo

Encontrar $a = a^*$ e $b = b^*$ tal que, de acordo com determinado critério, $y_i^* = a + bx_i$ seja suficientemente próximo de y_i.

Critério dos mínimos quadrados

Minimizar $$SSE = \sum_i \left(y_i - y_i^*\right)^2 = \sum_i e_i^2$$

$$\sum_i e_i^2 = \sum_i \left(y_i - a + bx_i\right)^2$$

Derivando em ordem a a e a b e igualando a zero vem:

Equações normais $$\begin{cases} na + b\sum x_i = \sum y_i \\ a\sum x_i + b\sum x_i{}^2 = \sum x_i y_i \end{cases}$$

$$\begin{cases} a = m_y - bm_x \\ b = \dfrac{\sum\left(x_i - m_x\right)\left(y_i - m_y\right)}{\sum\left(x_i - m_x\right)^2} \end{cases}$$

ESTIMAÇÃO, PREVISÃO E RESÍDUOS

Cada valor y_i^* pode ser usado como:

- Estimar $E\{Y|X = x_i\}$
- prever o valor de Y quando a variável X toma um determinado valor x_i.

Diferentes graus de confiança e diferentes variâncias estão associados a estes dois significados de y_i^*.

Os resíduos medem a qualidade do ajustamento:

$e_i = y_i - y_i^*$ $\Rightarrow$ medida da discrepância entre os dados e o modelo.

VARIAÇÃO DA VARIÁVEL DEPENDENTE

Cada valor y_i^* pode ser tomado como a parte do valor real y_i tomado em conta pelo modelo enquanto e_i reflecte a parte não explicada pelo modelo.

$$y_i = y_i^* + e_i$$

$$\sum y_i^2 = \sum \left(y_i^* - e_i \right)^2$$

$$\sum y_i^2 = \sum y_i^{*2} + \sum e_i^2 - \underbrace{2 \sum y_i^* e_i}_{\Downarrow \; 0}$$

$$SST_u = SSM + SSE$$

Subtraindo o efeito da média vem

$$SST_u - n\, m_y^2 = (SSM - n\, m_y^2) + SSE$$

$$\boxed{SST = SSR + SSE}$$

$$SST = \sum_{i=1}^{n} (y_i - m_y)^2$$

$$SSR = \sum_{i=1}^{n} (y_i^* - m_y)^2$$

$$SSE = \sum_{i=1}^{n} (y_i - y_i^*)^2$$

COEFICIENTE DE DETERMINAÇÃO E CORRELAÇÃO

Coeficiente de determinação

$$R^2 = \frac{SSR}{SST}$$

O coeficiente de determinação mede a proporção da variabilidade da variável dependente explicada pelo modelo.

Coeficiente de correlação

$$r = \sqrt{R^2} \quad \Rightarrow$$ Coeficiente de correlação entre y^* e y

$$r = r_{XY}$$

PRECISÃO DOS PARÂMETROS

Hipóteses: Os y_i são independentes e têm a mesma variância s^2.
Os e_i são gaussianos, independentes e têm a mesma variância σ^2.

Variância de *b*

$$Var(b) = \frac{s^2}{\sum(x_i - m_x)}$$

Variância de *a*

$$Var(a) = \left[\frac{1}{n} \quad \frac{m_x^2}{\sum(x_i - m_x)^2}\right]^2 s^2$$

PRECISÃO DOS ESTIMADORES E PREDICTORES

Variância de y_i^* (estimador da esperança condicional)

$$Var\left(y_0^*\right) = \left[\frac{1}{n} \frac{\left(x_0 - m_x\right)^2}{\sum\left(x_i - m_x\right)^2}\right]^2 s^2$$

Atinge o mínimo quando $x_0 = m_x = E(X)$.

Variância de y_i^* (predictor de y_i)

$$Var_p\left(y_0^*\right) = \left[1 + \frac{1}{n} \frac{\left(x_0 - m_x\right)^2}{\sum\left(x_i - m_x\right)^2}\right]^2 s^2 = Var(y_0^*) + s^2$$

MODELOS LINEARIZÁVEIS

Exponencial

$$y_i = a\, e^{bx_i} e_i$$

$$y_i^* = a^* + b\, x_i + e_i^* \quad \begin{cases} y_i^* = \ln(y_i) \\ a^* = \ln(a) \\ e_i^* = \ln(e_i) \end{cases}$$

Polinomial inverso

$$y_i = \frac{1}{a + b\, x_i + e_i}$$

$$y_i^* = b + a\frac{1}{x_i} + e_i^* \quad \begin{cases} y_i^* = \dfrac{1}{y} \end{cases}$$

REGRESSÃO MÚLTIPLA

$$Y = \boldsymbol{a} + \sum_j \boldsymbol{b}_j X_j$$

$$y_i = a + \sum_j b_j x_{ij} + \boldsymbol{e}_i$$

Em notação matricial vem:

$$\boldsymbol{y} = \boldsymbol{b}\,\boldsymbol{x} + \boldsymbol{e}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_1 \\ \vdots \\ \boldsymbol{e}_n \end{bmatrix}$$

Equações normais $\quad \boldsymbol{x'x\,b = x'\,y}$

$$\boxed{\boldsymbol{b} = (\boldsymbol{x'x})^{-1}\,\boldsymbol{x'\,y}}$$

No caso particular de *a*, vem:

$$\boxed{a = m_y - \sum b_j m_j}$$

REGRESSÃO MÚLTIPLA

AVALIAÇÃO DO MODELO

Coeficiente de determinação múltiplo

$$R^2 = \frac{SSR}{SST}$$

O coeficiente de determinação mede a proporção da variabilidade da variável dependente explicada pelo modelo.

Coeficiente de determinação ajustado

$$R^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$$

Coeficiente de correlação múltiplo

$$r = \sqrt{R^2} \quad \Rightarrow$$ Coeficiente de correlação entre y^* e y

REGRESSÃO MÚLTIPLA

AVALIAÇÃO DO MODELO

Estatística *F*

$$F = \frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} = \frac{MSR}{MSE}$$

Parâmetro que mede o significado do modelo.

Diagramas de dispersão

- y_i^* / y_i
- e_i^* / y_i
- e_i^* / x_{ij}

REGRESSÃO MÚLTIPLA

CASOS ESPECIAIS

Regressão passo a paso (*stepwise*)

- ascendente (*forward*)
- descendente (*backward*)

Regressão baseada na ACP