Estatística Aplicada a Recursos Hídricos

Docente: Rachid Muleia

(rachid.muleia@uem.mz)

Mestrado em Gestão de Recursos Hídricos - DGEO/UEM

Tema: Regressão Linear

Ano lectivo: 2023

Análise de regressão

- É uma metodologia estatística que utiliza a relação estatística entre duas ou mais variáveis quantitativas de forma que uma variável (variável resposta) possa ser estimada ou prevista através de outras variáveis (variáveis explicativas)
- Pode ser usada para previsão, estimativa, teste de hipótese e modelagem de relações causais
- Pode ser usada em situações onde a variável dependente é difícil, cara ou impossível de medir, mas seus valores podem ser previstos a partir de outra variável facilmente mensurável à qual esteja funcionalmente relacionada
- Algumas relações (entre duas ou mais variáveis) são faceis de se verificar ou estudar, principalmente quando se está diante de fenómenos determinísticos. Quando há aleatoriedade a volta do fenómeno, o estabelecimento de uma relação estatística exige mais cuidado na análise

Relação funcional vs Relação estatística

Relação funcional

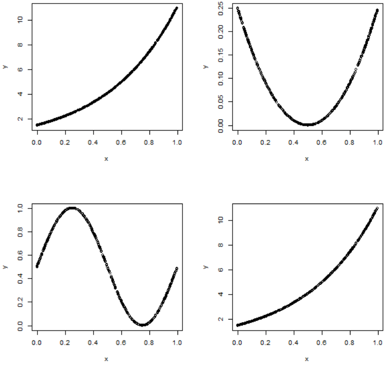
- A relação funcional entre duas variáveis é expressa através de uma equação matemática, $Y = f(X)$, onde é uma função conhecida. Ex: $Y = 2X$ ou $Y = X^2$

Relação estatística

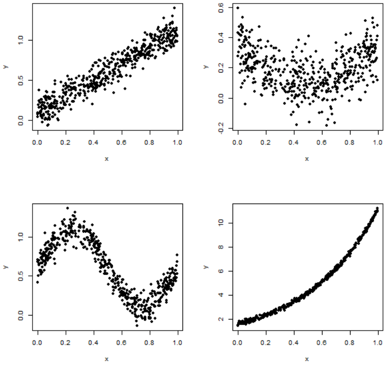
- Numa relação estatística as variáveis estão associadas a uma distribuição de probabilidade: $Y = f(X) + \epsilon$. O ϵ representa o erro que se comete ao se tentar aproximar Y por $f(X)$
- A variável Y é designada de variável dependente ou variável resposta, e a variável X é designada de variável independente

Relação funcional vs Relação estatística

Relação determinística



Relação estatística

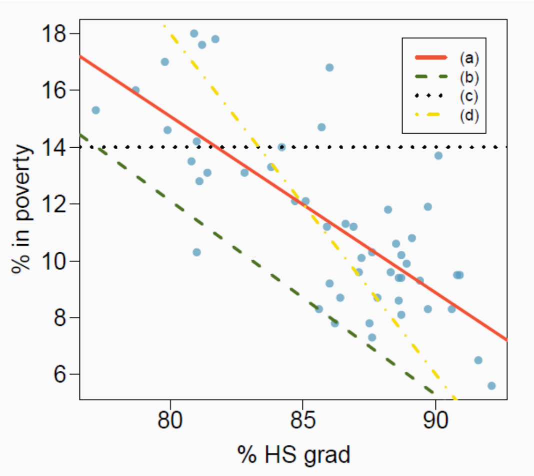


Exemplo de aplicação

O diagrama de dispersão abaixo mostra a relação entre a taxa de graduação no ensino secundário e a % de residentes que vivem abaixo da linha da pobreza em 50 estados dos EUA

Exemplo de aplicação: melhor estimativa - inspecção visual

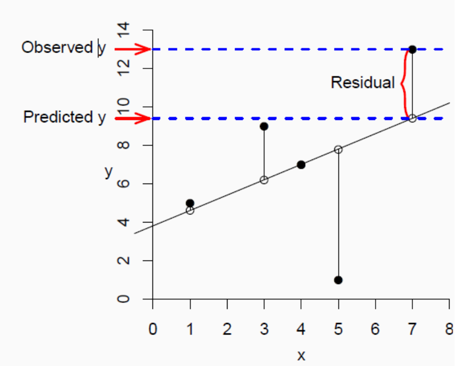
Qual é recta que melhor descreve a relação entre % de graduados e % de residentes que vivem abaixo da linha da pobreza?



Resíduos (erro de predição/previsão)

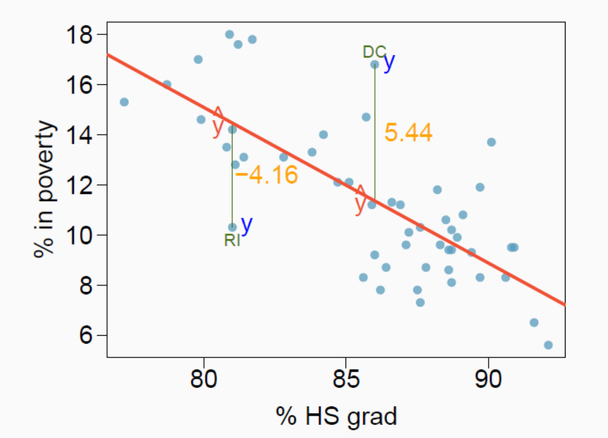
O resíduo de uma determinada observação ($i -$ ésima) é dada por

$$\underset{\text{(Resíduo)}}{e_i} = \underset{\text{(Valor observado de } y\text{)}}{y_i} - \underset{\text{(Valor previsto } \hat{y}_i\text{)}}{\hat{y}_i}$$



- O é a distância vertical de um ponto até a recta
- Os pontos acima da recta terão resíduo positivo e abaixo da recta valores negativos

Resíduos (cont.)



- A % de residentes abaixo da linha da pobreza em DC é 5.44% a mais do que o previsto
- EM RI a % é 4.16% a menos do que o previsto

Recta com melhor ajustamento

- O objectivo é encontrar uma recta que melhor se ajusta aos dados (ao diagrama de dispersão), $y = a + bx$, que tenha os menores resíduos

- O método mais popular para encontrar a recta que melhor se ajusta aos dados é o métodos dos mínimos quadrados ordinário (MQO)

$$e_1 + e_2 + \cdots + e_n = \sum_{i=1} n(y_i - \hat{y}_i)^2 = \sum_{i=1} n(y_i - a - bx)^2$$

- O intercepto a e a declive b podem ser calculadas usando as seguintes formulas

$$b = \text{declive} = r \cdot \frac{s_y}{s_x}$$
$$a = \text{intercepto} = \bar{y} - \text{declive} \cdot \bar{x}$$

Exemplo: Pobreza vs Graduados

- O objectivo é encontrar uma recta que melhor se ajusta aos dados (ao diagrama de dispersão), $y = a + bx$, que tenha os menores resíduos

- O método mais popular para encontrar a recta que melhor se ajusta aos dados é o métodos dos mínimos quadrados ordinário (MQO)

$$e_1 + e_2 + \cdots + e_n = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx)^2$$

- O intercepto a e a declive b podem ser calculadas usando as seguintes formulas

$$b = \text{declive} = r \cdot \frac{s_y}{s_x}$$
$$a = \text{intercepto} = \bar{y} - \text{declive} \cdot \bar{x}$$

Exemplo : Graduados vs Pobreza

A tabela abaixo mostra algumas estatisticas necessárias para as variáveis % de graduados e % de pobres

	% Grad x	% Pobreza y
Média	$\bar{x} = 86.01$	$\bar{y} = 11.35$
Desvio-padrão	$s_x = 3.37$	$s_y = 3.1$
correlação	$r = -0.75$	

- Considerando os dados da tabela, o intercepto e a declive para a recta (equação) de regressão linear são

$$b = \text{declive} = r \cdot \frac{s_y}{s_x} = -0.75 \times \frac{3.1}{3.37} = -0.62$$

$$a = \text{intercepto} = \bar{y} - \text{declive} \cdot \bar{x} = 11.35 - (-0.62) \times 86.01 = 64.68$$

Interpretação dos coeficientes do modelo de regressão - declive

- A declive representa a variação esperada na variável resposta quando a variável independente aumenta em uma unidade

$$\% \quad \widehat{pobreza} = 64.68 - 0.62 \times \% \quad \widehat{graduados}$$

- Para cada ponto percentual adicional na taxa de graduação, esperaríamos que a percentagem de pessoas que vivem na pobreza fosse inferior (reduzisse), em média, em 0.62 pontos percentuais.

- Para estados com taxa de graduação no HS de 90%, suas taxas de pobreza são 6.2% mais baixas, em média, do que aqueles estados com taxa de graduação no HS de 80%.

Interpretação dos coeficientes do modelo de regressão - intercepto

- O intercepto é o valor previsto da resposta quando $x = 0$, o que pode não ter um significado prático quando $x = 0$ não é um valor possível

$$\% \quad \widehat{pobreza} = 64.68 - 0.62 \times \% \quad \widehat{graduados}$$

- Espera-se que os estados sem graduados em HS tenham, em média,tenha 64.68% dos residentes que vivem abaixo da linha da pobreza. Isto não é possível

Variância dos residuos e dos valores previstos

Lembre-se que

$$\underset{\text{(Resíduo)}}{e_i} = \underset{\text{(Valor observado de } y\text{)}}{y_i} - \underset{\text{(Valor previsto } \hat{y}_i\text{)}}{\hat{y}_i}$$

Pode-se mostrar que :

$$\underset{\text{(Variância de } y\text{)}}{\underset{||}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \underset{\text{(Variabilidade de y explicada por } x\text{)}}{\underset{||}{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}} + \underset{\text{(Variabilidade de } y \text{ não explicada por } x\text{)}}{\underset{||}{\frac{1}{n-1}\sum_{i=1}^{n} e_i}}$$

Coeficiente de determinação, R^2

Pode-se igualmente mostrar que,

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{variância dos valores previstos } \hat{y}_i}{\text{Variância dos valores observados } y_i}$$

Isto é,

$$\begin{aligned} R^2 &= \text{O quadrado do coeficiente de correlação} \\ &= \text{proporção da variância em } y \text{ explicada} \\ &\quad \text{pela variável independente} \end{aligned}$$

O resto da variabilidade é explicada por variáveis não incluídas no modelo ou por aleatoriedade inerente aos dados