Regressão Linear - IA

Marcus Magalhães
Vinícius Rodrigues

May 2019

1 Introdução

Ao se estudar uma base de dados, um dos interesses é determinar se há alguma relação entre as suas variáveis. Por exemplo, na base de dados Boston podemos verificar se existe uma relação entre o numero de quartos de um imóvel e seu preço, entre distância de centros comerciais e poluição do ar, etc.
Na probabilidade e estatística, essa relação entre duas ou mais variáveis é chamada de correlação e regressão. Se o estudo trata apenas duas variáveis, tem-se correlação e regressão simples, se conter mais de duas variáveis tem-se correlação e regressão múltipla.
A correlação resume o grau de interdependência linear entre as variáveis. Já a analise de regressão determina uma equação que descreve o relacionamento entre as variáveis. A regressão e a correlação tratam apenas do relacionamento do tipo linear entre duas variáveis.

2 Correlação

Correlação é a dependência entre duas variáveis de uma população. Informalmente correlação é sinônimo de dependência. Formalmente variáveis são dependentes se não satisfizerem a propriedade matemática da independência probabilística.
Uma das maneiras de medir o grau e o sinal da correlação entre variáveis é a partir do Coeficiente de Correlação Linear de Pearson, definido por:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

O valor de r está no intervalo de -1 a 1. Se r = 1, indica uma correlação linear positiva perfeita. Se r = -1, então há uma correlação linear negativa perfeita.

Valor de r (+ ou -)	Interpretação
0.00 a 0.39	Muita fraca
0.40 a 0.80	Moderada
0.81 a 1.00	Forte

Table 1: Interpretação para o coeficiente r

2.1 Causalidade

A expressão correlação não implica causalidade significa que correlação não pode ser usada para a relação causal entre as variáveis. Por exemplo, a quantidade de queimaduras de sol pode estar fortemente correlacionada ao número de óculos de sol vendidos em uma cidade litorânea, mas nenhum fenômeno é provavelmente a causa do outro.

3 Regressão Linear Simples

A regressão linear simples determina uma equação linear que descreve o relacionamento entre duas variáveis. A partir do modelo gerado é possível então estimar o valor da variável y (variável alvo) a partir de uma variável independente x.

3.1 Equação da Regressão Linear

A equação da regressão linear é da forma

$$y_i = \beta X_i + \alpha + \epsilon$$

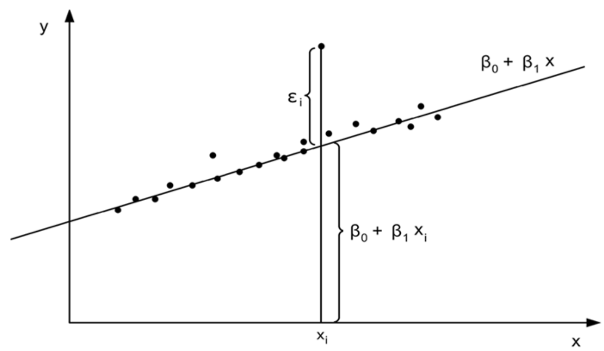
, onde:
y_i: Variável explicada (dependente); representa o que o modelo tentará prever;
α: É uma constante, que representa a interceptação da reta com o eixo vertical;
β: Representa a inclinação em relação à variável explicativa;
X_i: Variável explicativa (independente);
ϵ_i: Representa todos os factores residuais mais os possíveis erros de medição;

3.2 Determinação dos parâmetros da equação

Existem diversas maneiras de determinar os parâmetros da equação de regressão linear, a mais utilizada é usando o método dos Mínimos Quadrados (MMQ), também conhecido como método dos mínimos quadrados ordinários (MQO).

3.2.1 Métodos dos Mínimos Quadrados

Considere o conjunto S de n pares de valores (x_i, y_i), $i = 1, 2, 3, ...n$, que correspondem a pontos de um gráfico. Agora suponha que trassemos uma reta arbitrária $x\beta_0 + \beta_1$ que passe por esses pontos. No valor x_i da variável independente, o valor predito por esta reta é $\beta_0 + \beta_1 x_i$, enquanto valor observado é y_i. Os desvios (erros) da predição é $\epsilon_i = y_i - [\beta_1 + \beta_0 x_i]$.



Sendo assim, a melhor reta que descreve os pares de pontos do conjunto S, é aquela que minimiza a seguinte equação:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [y_i - \beta_1 + \beta_1 x_i]^2. \tag{1}$$

Para encontrar o minimo dessa equação, precisamos derivá-la em relação a β_0 e β_1, e igualar as equações resultante a 0. Assim:

$$\begin{aligned} \frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0} &= -2 \sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i = 0 \\ \frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{aligned} \tag{2}$$

Simplificando, obtemos as equações denominadas Equações Normais de Mínimos Quadrados.

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \tag{3}$$

Agora a ultima etapa para determinar o valor dos parâmetros β_0 e β_1 que minimizam L, é resolver o sistema de equações (3).

Resolvendo o sistema de equações (3), obtemos:

$$\begin{aligned} \beta_0 &= \bar{Y} - \beta_1 \bar{X} \\ \beta_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2} \end{aligned} \tag{4}$$

3.3 A variância em torno da linha de regressão

Assim como se pode definir uma variância (ou desvio padrão) de um conjunto de pontos em torno de seu valor médio Y , também se pode definir uma variância (ou desvio padrão) de um conjunto de pontos ordenados y_i em torno da sua linha de regressão $\hat{y}$. Esta quantidade, denotada por S_{XY}^2 , é definida como:

$$S_{XY}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

3.4 Coeficiente de determinação R^2

O coeficiente R^2 determina o quanto o modelo consegue explicar os valores observados. Seu valor é dado pela razão entre a soma dos quadrados dos resíduos (SQ_{res}) e a soma total dos quadrados (SQ_{tot}).

$$R^2 = \frac{SQ_{res}}{SQ_{tot}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{5}$$

4 Aplicação

A econometria é uma das áreas que mais utilizam modelos matemáticos para buscar entender as relações entre suas variáveis econômicas. Na econometria moderna, outras ferramentas estatísticas são frequentemente usadas, mas a regressão linear é, ainda, o ponto de partida mais frequente para a análise.

4.1 Referências

http://sisne.org/Disciplinas/Grad/ProbEstat2/aula18.pdf.
http://www.portalaction.com.br/analise-de-regressao/regressao-linear-simples.
http://www.pucrs.br/ciencias/viali/graduacao/engenharias/material/apostilas/Apostila_5.pdf