Tópicos Especiais em Computação I

Regressão Linear

Patrícia Lucas

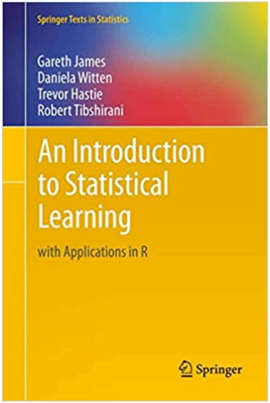
Bacharelado em Sistemas de Informação
IFNMG - Campus Salinas

Salinas
Março 2021

INSTITUTO FEDERAL
Norte de Minas Gerais
Campus Salinas

Versão 1.2021

Referência



Capítulo 3: Linear Regression

An Introduction to Statistical Learning: with Applications in R. G. James, D. Witten, T. Hastie, and R. Tibshirani. Springer, 2013.

É uma abordagem linear muito simples para prever uma resposta quantitativa Y com base em uma única variável preditora X.

Supõe-se que exista aproximadamente uma relação linear entre X e Y.

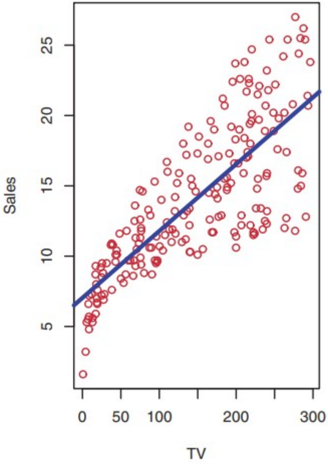
$$Y \approx \beta_0 + \beta_1 X \tag{1}$$

β_0 e β_1 são duas constantes desconhecidas que representam os termos de interceptação e inclinação no modelo linear.

Juntos, β_0 e β_1 são conhecidos como coeficientes ou parâmetros do modelo.

Exemplo:

$$Sales \approx \beta_0 + \beta_1 TV \tag{2}$$



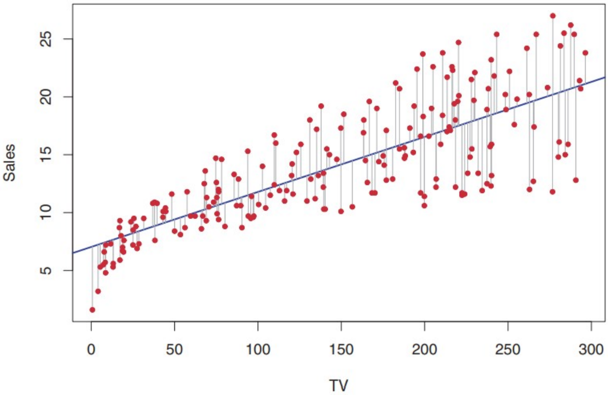
Na prática, β_0 e β_1 são desconhecidos. Portanto, antes de podermos usar o modelo para fazer previsões, precisamos usar dados para estimar os coeficientes.

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

Onde x são os atributos preditores (entradas) e y o atributo alvo (saídas).

Nosso objetivo é obter estimativas dos coeficientes β_0 e β_1 de modo que o modelo linear se ajuste bem aos dados disponíveis.

Em outras palavras, queremos encontrar uma reta cuja interceptação β_0 e inclinação β_1, de modo que a linha resultante seja o mais próxima possível dos n pontos de dados.



Existem várias maneiras de medir a proximidade, usaremos o **Método dos Mínimos Quadrados:**

Seja $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ a previsão para Y com base no i-ésimo valor de X.

Então $\epsilon_i = y_i - \hat{y}_i$ representa o i-ésimo erro, ou seja, a diferença entre o i-ésimo valor de resposta observado e o i-ésimo valor de resposta previsto pelo modelo linear.

Definimos a soma dos erros quadráticos (SEQ) como:

$$SQE = e_1^2 + e_2^2 + \ldots + e_n^2 \tag{3}$$

ou equivalente:

$$SQE = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \tag{4}$$

A abordagem dos mínimos quadrados escolhe $\hat{\beta}_0$ e $\hat{\beta}_1$ para minimizar a SEQ. Usando alguns cálculos, pode-se mostrar que os minimizadores são:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})}{\sum_{i=1}^{n}(x_i - \hat{x})^2} \tag{5}$$

$$\hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \hat{x} \tag{6}$$

onde $\hat{y}$ e $\hat{x}$ são as médias da amostra:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{7}$$

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{8}$$

Ajuste da regressão linear simples para os dados de Publicidade usando vendas como resposta e TV como preditor, onde $\hat{\beta}_0 = 7,03$ e $\hat{\beta}_1 = 0,0475$.