Semana 2.1 - Regressão Linear

Prof. Raphael Y. de Camargo
Centro de Matemática, Computação e Cognição (CMCC)
Universidade Federal do ABC

UFABC

Regressão Linear

Primeira técnica que aprendemos em aprendizado de máquina / estatístico

Em uma campanha de marketing podemos nos perguntar:

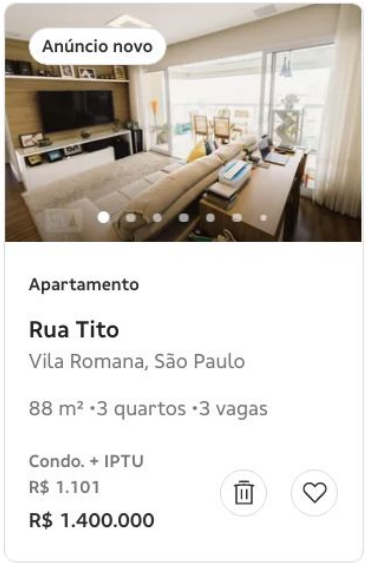
- Existe uma relação entre gastos em propaganda e venda?
- Quão forte é esta relação?
- Quais tipos de mídia contribuem para as vendas
- Como estimar o efeito de cada mídia nas vendas
- Podemos predizer vendas futuras?
- A relação é linear?
- Existem sinergias entre os mídias?

https://www.template.net/business/marketing/sample-product-marketing-template/

Preços de Imóveis

Um outro exemplo é se podemos estimar o preço de imóveis
a partir de informações como número de quartos, idade do imóvel,
renda de moradores no bairro e outro:

- Existe relação entre cada característica e o preço do imóvel?
- Quão forte é esta relação?
- Quais tipos de características contribuem para os preços?
- Como estimar o efeito de cada característica nos preços?
- Podemos prever o preço de um imóvel a partir destas características?
- A relação é linear?
- Existem sinergias entre as características?



https://www.template.net/business/marketing/sample-product-marketing-template/

Parte I
Regressão Linear Simples

Regressão Linear Simples

Fazemos a suposição que existe uma relação linear simples entre um preditor X e a resposta Y

$$Y \approx \beta_0 + \beta_1 X.$$

Por exemplo, podemos modelar que as vendas de um produto são proporcionais ao valor gasto em propaganda na TV

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

β_0: vendas independentes da propaganda
β_1: vendas resultantes da propaganda

Para o caso de imóveis

valor ≅ β_0:+ β_1 x metragem

Determinando os Coeficientes

Precisamos estimar os valores dos parâmetros β_0 e β_1, a partir de um conjunto de exemplos $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$

Exemplo das vendas:

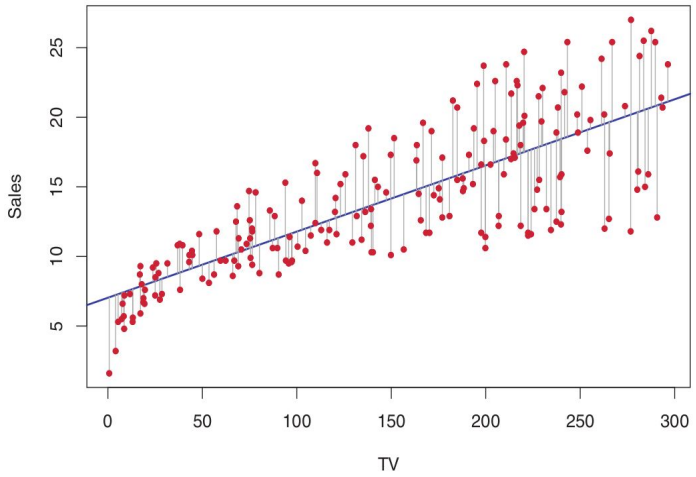
- Para cada ponto, calculamos $e_i = y_i - \hat{y}_i$
- A soma dos quadrados dos resíduos (RSS) é

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

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

- Basta agora fazermos a derivada parcial de RSS com relação a β_0 e β_1 ser igual a 0 e chegamos em:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Acurácia dos Coeficientes

Os coeficientes são estimados com relação a uma amostra de dados fornecida.
Eles mudam para diferentes amostras.

Para o cálculo da média μ de um conjunto de dados, o **erro padrão** SE da estimativa é dado por:

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n},$$

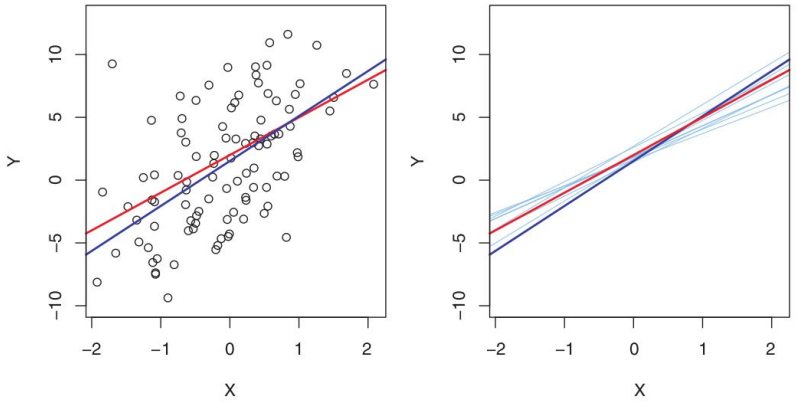
Para o caso da regressão linear, os erros são:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

Quanto maior $(x_i - \bar{x})^2$, menores são os erros

$$\sigma^2 = \mathrm{Var}(\epsilon)$$

Para $\bar{x} = 0$, SE ($\hat{\beta_0}$) = SE ($\hat{\mu}$)



Vermelho: relação real Y = 2X + 3 + ϵ
Azul: relação estimada para a amostra

Azul Claro: relações estimadas para outras amostras da população

Intervalos de Confiança

Com o SE, podemos calcular os intervalos onde os $\hat{\beta}$ possuem 95% de chance de estar:

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

No exemplo da propaganda: β_0=[6.130, 7.935] e β_1=[0.042, 0.053]

É importante para sabermos o quanto podemos confiar no modelo.

O SE está também ligado aos **testes de hipótese:**

Hipótese nula (H_0): Não existe relação entre X e Y $\Leftrightarrow$ $\beta_1 = 0$ ***vs***

Hipótese alternativa (H_a): Existe uma relação entre X e Y $\Leftrightarrow$ $\beta_1 \neq 0$

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

p-valor: indica a probabilidade de termos encontrado $\beta_1 \neq 0$ de modo acidental

Calculado a partir da *estatística-t*

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

Acurácia do Modelo

Erro Residual Padrão (Residual Standard Error - RSE):

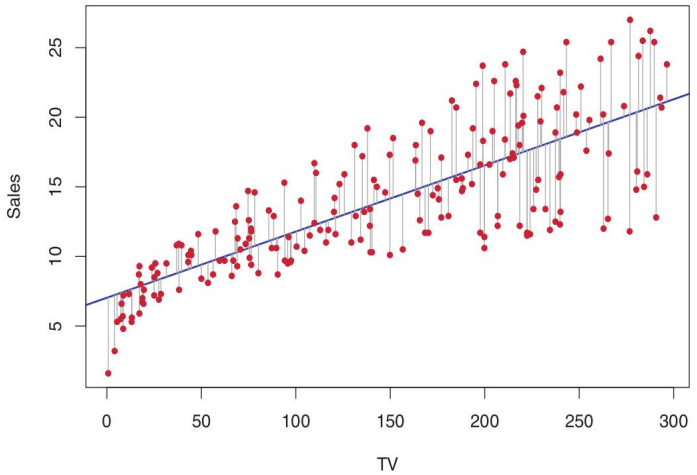
$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Podemos interpretá-lo como sendo a soma de quanto cada predição está longe do valor correto.

Os erros são elevados ao quadrado, de modo que predições muito erradas tem um grande impacto

No exemplo da propaganda na TV, o erro é de 3,26

O valor médio é de 14.000, de modo que o que erro percentual é de 23%



Acurácia do Modelo (cont.)

Estatística R²: $R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$

$$\text{TSS} = \sum (y_i - \bar{y})^2$$

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2, \quad e_i = y_i - \hat{y}_i$$

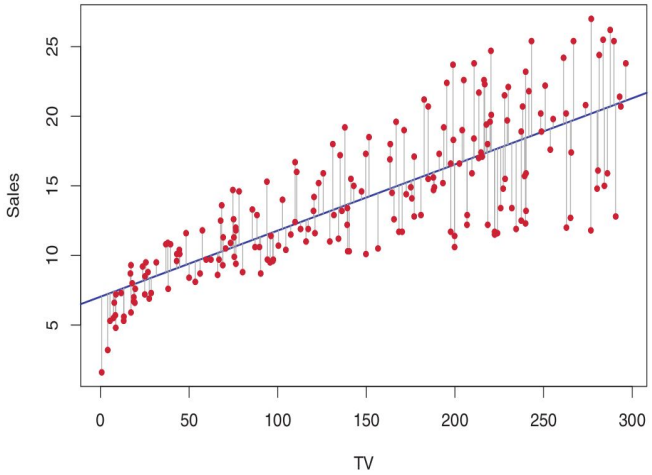
Total Sum of Squares (TSS): é a variância da variável alvo Y

Residual Sum of Squares (RSS): é o quanto da variabilidade de Y que o modelo ***não*** consegue explicar

R²: indica a proporção da variância de Y pode ser explicada pelo modelo.

Valores próximos de 1 indicam um alto poder de explicação

No exemplo da propaganda, este valor é 0.61



Exemplo no statsmodels

statsmodels: provê a qualidade do ajuste, p-valores e outras estatísticas

https://www.statsmodels.org/stable/index.html

Faremos inicialmente um exemplo sintético

Em seguida, usaremos os mesmos dados dos preços dos imóveis na Califórnia da semana 1