Regressão Linear Simples

Machine Learning
Prof. Neylson Crepalde

Pense nos dados *Advertisement*

Algumas perguntas importantes:

1. Existe uma relação entre o investimento em propaganda e vendas?
2. Quão forte é a relação entre o investimento em propaganda e vendas?
3. Qual *media* contribui para as vendas?
4. Quão precisa é a estimação do efeito de cada *media* sobre as vendas?
5. Quão precisamente podemos prever novas vendas?
6. A relação é linear?
7. Existe sinergia entre as diversas *media* de progaganda?

Na regressão linear, podemos predizer o valor de uma variável quantitativa *Y* (dependente) a partir de um preditor *X*. Matematicamente, podemos representar essa relação da seguinte maneira:

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

onde $\approx$ representa "é aproximadamente modelado como". Beta0 e Beta1 são constantes do modelo que representam, respectivamente o *intercepto* e o *slope* (inclinação). Esses coeficientes (ou parâmetros, no caso da população, estimadores, no caso da amostra) são desconhecidos e serão estimados. A partir deles, podemos prever novas vendas.

Ciências de Dados

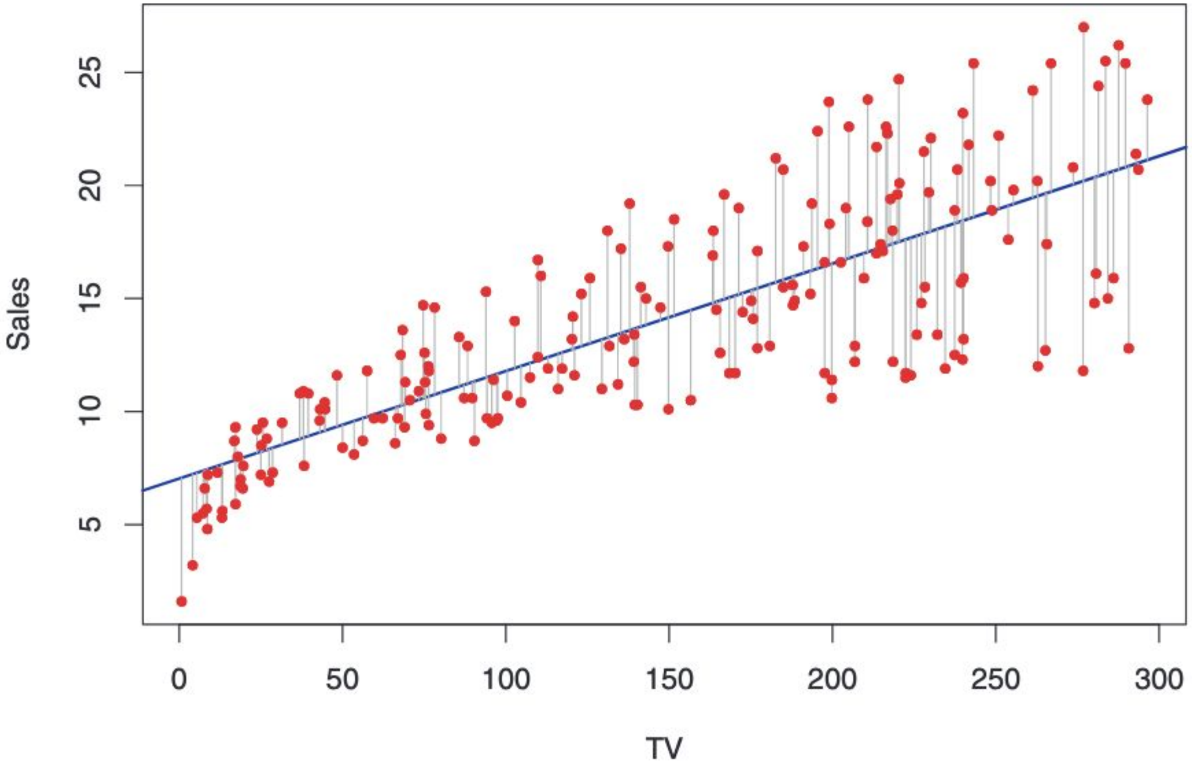

FIGURE 3.1. *For the* Advertising *data, the least squares fit for the regression of* sales *onto* TV *is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.*

Fonte: Gareth *et al*, 2013.

O método dos mínimos quadrados ordinários visa minimizar a soma dos quadrados dos erros (*Residual Sum of Squares*).

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

A partir desse método de estimação, os coeficientes podem ser estimados por:

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

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

Acurácia do modelo

Para mensurar o quanto nossas estimativas estão próximas ou não dos parâmetros populacionais, usamos o Erro Padrão (*Standard Error*).

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

Os Erros Padrão podem ser usados para estimar os intervalos de confiança da seguinte maneira:

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

Ciências de Dados

Testes de hipótese

Os Erros Padrão podem ser usados para computar os testes de hipótese. Testamos a hipótese nula de que não há relação entre X e Y e a hipótese alternativa de que há relação entre X e Y. O teste de Hipótese é expresso da seguinte maneira:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0,$$

Para testar a hipótese nula, verificamos se o estimador é suficientemente distante de zero de modo que temos confiança de que ele é de fato diferente de zero. Para isso, computamos a estatística *t:*

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

que mede o número de desvios padrão que o estimador Beta1 está afastado de 0. Consequentemente, podemos computar a probabilidade de observar um número igual a *t* ou maior em valores absolutos assumindo Beta1 = 0. Chamamos essa probabilidade de *p-value*. De modo geral, interpretamos o *p-value* da seguinte maneira:

Um *p-value* pequeno indica uma baixa probabilidade de observar uma associação tão substancial entre entre o preditor e a dependente por acaso, na ausência de uma relação real entre X e Y. Desse modo, quando obtemos um *p-value* pequeno (< 0.05), rejeitamos a hipótese nula e inferimos que há associação entre X e Y. Do contrário, não podemos rejeitar a hipótese nula e consideramos que não há relação entre X e Y.

Lab: Regressão Simples com *Advertisement Data*

Acurácia do modelo

O RSE (*Residual Standard Error*) - Erro Padrão do Resíduo - é uma estimação do desvio padrão do erro. Ela pode ser definida por:

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

RSE é considerado uma medida de *falta de ajuste* do modelo. Quanto menor o RSE melhor o modelo se ajusta aos dados.

R2 representa a proporção da variância explica. É uma medida de quanto da variância de Y o modelo explica. R2 pode ser calculado assim:

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

onde:

$$\mathrm{TSS} = \sum (y_i - \bar{y})^2 \quad \& \quad \mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Multiple Linear Regression

Na prática, não é comum utilizarmos a regressão simples pois é conhecido que qualquer fenômeno que possamos escolher para investigação pode ter múltiplas correlações. Em face ao argumento de que poderíamos estimar várias regressões simples para testar o efeito de várias variáveis, podemos afirmar que essa abordagem não é satisfatória pois nenhum efeito estimado leva em conta as outras variáveis mensuradas. Por esse motivo, a estimação de correlações simples entre variáveis pode levar a um viés grande.

A regressão múltipla possui a seguinte equação:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

onde os vários X representam as diversas variáveis que vamos inserir no modelo. Cada preditor vai possuir o seu *slope*, ou seja, seu efeito sobre Y levando em conta, controlando, pelos demais preditores (*Ceteris Paribus*).

Ciências de Dados

Lab: Regressão Simples e Múltipla com *Advertisement Data*

Ciências de Dados

Algumas questões importantes

1. Pelo menos um dos preditores é útil para prever a variável dependente?
2. Todos os preditores ajudam a explicar Y ou apenas um subconjunto desses preditores?
3. Quão bem o modelo se ajusta aos dados?
4. Dado um set de preditores, qual resposta nós daríamos e quão acurada seria essa resposta?

1

Consideremos a hipótese nula

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

com relação à alternativa:

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

Podemos testá-la calculando a estatística de teste *F* que mede, em linhas gerais, quantas vezes o resultado do modelo é melhor que a média:

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)}$$

Ciências
de Dados

2 Decidindo sobre variáveis importantes

Parte do nosso curso será destinado exclusivamente a este problema. Idealmente nós tentaríamos todas as combinações de variáveis e verificaríamos qual dos modelos possui o melhor ajuste através de qualquer medida escolhida (AIC, BIC, R2, RSS). Entretanto, existe um total de 2^p modelos. Se há apenas, 2 variáveis, ótimo. 2^2 = 4 modelos. Entretanto, se há no banco 30 variáveis, temos um total de 2^30 = 1.073.741.824 modelos. IMPOSSÍVEL!!

Três abordagens são mais conhecidas:

1. *Forward Selection*: Começar com um modelo vazio e adicionar variáveis gradualmente testando o ajuste a cada adição;
2. *Backward Selection*: Começar com todas as variáveis e retirar gradualmente enquanto testa o ajuste;
3. *Mixed Selection*: Uma mistura de *forward* e *backward selection.*

Ciências de Dados

3 Ajuste

Observar:

- R2
- RSE

Ciências de Dados

4 Predições

Para prever a variável Y é necessário apenas realizar a equação de regressão uma vez que os estimadores Betas são conhecidos. Entretanto, há um erro associado aos estimadores. Portanto, para uma melhor acurácia dos resultados, é possível computar um intervalo de confiança. Normalmente utilizamos 95% de confiança.

É importante lembrar que o modelo linear assume uma relação linear entre as variáveis o que por vezes pode não ser verdade. Isso introduz um viés no modelo. Estudaremos posteriormente maneiras de corrigir esse viés. Por ora, assumiremos que a equação linear está correta.

Mesmo que os parâmetros reais fossem conhecidos, seria impossível uma predição completamente acertada tendo em vista o componente aleatório do modelo. Para saber qual é o tamanho da variação que Y predito pode ter, podemos computar os intervalos de predição.

Exercício!!

Preditores Qualitativos

Ciências
de Dados