Matemática da regressão linear - Básico

Luiz Paulo Tavares

2023-09-04

```
rm(list = ls())
graphics.off()
set.seed(123)

pacman::p_load(tidyverse,
               stats)
```

Dependências computacionais

Da equação da reta à regressão linear

É intuitivo iniciar a exposição de regressão linear no $\mathbb{R}^2$ tomando como ponto de partida a equação da reta:

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x, y) \mid x, y \in \mathbb{R}\}$$

$$y = a + bx$$

Estatisticamente, uma dada regressão linear simples busca estimar a inclinação da reta, ou seja, estimar uma reta que minimiza os pontos entre x e y de intersecção dado uma determinada dispersão. Assim, pode-se especificar um modelo de regressão tomando uma variável dependente $Y \in \mathbb{R}^2$ dado um vetor $x = (x_1, ..., x_d) \in \mathbb{R}^d$:

$$\mathbb{R}^d = \mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R} = \{(x_1, x_2, \ldots, x_d) \mid x_1, x_2, \ldots, x_d \in \mathbb{R}\}$$

Em uma notação convencional (isto é, não matricial) e simplificada, temos:

$$Y_i = \beta_1 + \beta_2 X_i + \mu_i$$

Observe a adição de μ representando os resíduos da regressão. Os quais são em essência a diferenças entre os valores observados e estimados. Por sua vez, β_1 e β_2 representam o intercepto e coeficiente angular da equação da reta, respectivamente. Assim, estima-se dado uma amostra n com variabilidade:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\mu}_i$$

Os resíduos:

$$\hat{\mu}_i = Y - \hat{Y}_i$$

Portanto:

$$\hat{\mu_i} = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

Do exposto, desenvolve-se por consequência a busca por encontrar ou, melhor, minimizar os resíduos encontrando valores próximos de Y observado. Critério bem difundido na literatura, pelo menos desde Friedrich Gauss, é o método dos mínimos quadrados. O qual, como suguere, busca minimizar os resíduos (GUJARATI & PORTER, 2011). Tomando como ponto de partida:

$$\sum_{i=1}^{n} \mu_i^2 = \sum (Y_i - \hat{Y}_i)^2$$
$$= \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

Posteriormente, a minimização via MQO será retomada. Por enquanto, tenha como dado que minimizar os resíduos como exposto e chegar num métrica de ajuste do modelo estimado: coeficiente de determinado R²:

$$SQE = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

$$SQT = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

$$R\check{s} = \frac{SQE}{SQT}$$

Dispersão

Pode-se estimar considerando apenas uma variável indepedente com $n = 1000$ com $X \sim N(0, \sigma^2)$:

```
# Modelo de regresssão linear simples

y = rnorm(n = 1000)
x = rnorm(n = 1000, mean = 0, sd = 1)

bd_model = data.frame(depedente = y, explicativa = x)

model_lm = stats::lm(formula = depedente ~ explicativa, data = bd_model) %>% print()
```

```
##
## Call:
## stats::lm(formula = depedente ~ explicativa, data = bd_model)
##
## Coefficients:
## (Intercept)  explicativa
##     0.01252      0.08494
```

Assim, estima-se $\hat{\beta}_0$ e $\hat{\beta}_1$, ou seja, o intercepto e o coeficiente angular da reta de regressão linear, respectivamente. O interceto retornou aproximadamente 0.12 e, por outro lado, o coefiente angular 0.85. É interessante nota que matematicamente a inclinação da reta pode ser encontrada como segue:

$$\hat{\beta}_2 = \frac{rs_y}{s_x}$$

Ou seja, dado a correlação de Pearson r entre x e y multiplicado pelo desvio padrão s_y dividido pelo desvio padrão de s_x. Assim, temos o resultado de aproximadamente 0.85 para o coefiente angular encontrado anteriormente. Por questão de dúvida, em seguida define-se a correlação e desvio padrão:

$$Cov(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

$$r = \frac{cov(X, Y)}{s_x s_y}$$

e desvio padeão:

$$s = \sqrt{\frac{1}{N} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

```
b_2 = (stats::cor(x,y) * stats::sd(y)) / (stats::sd(x))
print(b_2)
```

```
## [1] 0.08493951
```

para b_zero E o intercepto

$$\beta_1 = \overline{y} - \beta_2 \overline{x}$$

```
#b_1 = base::mean(y) - (model_lm[["coefficients"]][["x"]] *base::mean(x))
#print(b_1)
```

Ou seja, uma alta dispersão, principalmente, seguindo um padrão não linear pode em muito prejudicar o ajuste de uma reta linear entre x e y de minimização. Indo além, pode-se tomar a dispersão como μ:

$$y(x) := E[Y|X = x]$$