Regresssão Linear - Introdução

Mini-curso de Introdução a ML e AI

Mário Olímpio de Menezes

Regressão Linear

Análise de Regressão

- A Análise de Regressão é utilizada para se explicar ou modelar o relacionamento entre uma única variável Y, chamada de variável *resposta*, *de saída* ou *dependente* e uma ou mais variáveis *preditoras*, *de entrada* ou *explicativas*, $X_1, X_2, ..., X_p$.
- Quando $p = 1$, é chamada regressão simples;
  - Quando $p > 1$ é chamada regressão múltipla ou algumas vezes, regressão multivariada.
- A variável resposta deve ser uma variável contínua
- As variáveis explicativas podem ser contínuas, discretas ou categóricas.

Conceito Chave

- *Variação constante*:
  - Para cada **acréscimo de uma unidade** da variável explicativa, temos um **acréscimo constante na variável resposta.**

Regressão Linear – objetivos, história

A Análise de Regressão tem vários possíveis objetivos, incluindo:

- Predição de observações futuras
- Avaliação do efeito de, ou do relacionamento entre, as variáveis explicativas sobre a resposta
- Uma descrição geral da estrutura dos dados

História (em poucas linhas)

- Problemas do tipo regressão foram abordados primeiramente no início do século 19, e estavam relacionados ao uso da astronomia na navegação.
- Legendre desenvolveu o método dos mínimos quadrados em 1805.
- Gauss disse que o tinha desenvolvido alguns anos antes e mostrou, em 1809, que os mínimos quadrados eram a solução ótima quando os erros tem uma distribuição normal.
- A metodologia ficou restrita às ciências físicas até a parte final do século 19, quando em 1875, Francis Galton cunhou o termo *regressão à mediocridade*.

Modelo linear

Um **modelo linear** entre duas variáveis X e Y, é definido matematicamente como uma equação com dois parâmetros desconhecidos,

$$y = b_0 + b_1 x$$

que é uma estimativa da linha de regressão verdadeira da população:

$$\mu_y = \beta_0 + \beta_1 x$$

Esta linha de regressão descreve como a resposta média μ_y muda com x.

Os valores observados para y variam em torno da sua média μ_y e assumimos que tem o mesmo desvio padrão σ.

Os valores ajustados b_0 e b_1 estimam o verdadeiro *deslocamento* (*intercept*) e a inclinação da linha de regressão da população.

Para fins de simplificação, indicamos $Y \equiv \mu_y$ na fórmula:

$$Y = \beta_0 + \beta_1 X$$

Assim, dados n pares de valores, $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, se for admitido que Y é função linear de X, pode-se estabelecer uma regressão linear simples, cujo modelo estatístico é

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad i = 1, 2, \ldots, n$$

Regressão por Mínimos Quadrados Ordinários (OLS)

- Incluindo: regressão linear simples, regressão polinomial e regressão linear múltipla (multivariada)
- Para podermos interpretar corretamente os coeficientes de um modelo OLS, temos que satisfazer algumas hipóteses estatísticas:
  - *Normalidade* – Para valores fixos das variáveis independentes, a variável dependente é distribuida normalmente.
  - *Independência* – Os valores de Y_i são independentes uns dos outros.
  - *Linearidade* – A variável dependente está linearmente relacionada às variáveis independentes.
  - *Homocedasticidade* – A variância y é constante, ou seja, não varia com os níveis das variáveis independentes.
- Além disso:
  - A variável explicativa x é medida sem erro;
  - A diferença entre um valor medido de y e o valor predito pelo modelo para o mesmo valor de x é chamado de *resíduo*
    - Resíduos são medidos na escala de y, e são distribuídos normalmente.

Exemplo Regressão Linear Simples

Exemplo de Regressão Linear Simples

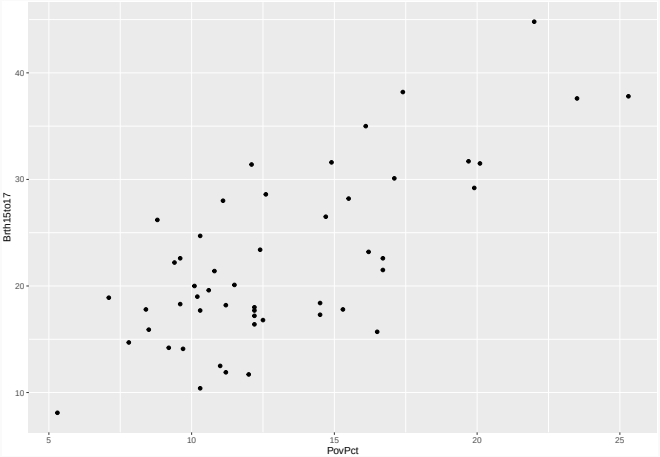
Dataset: Taxa de Nascimentos (mães entre 15 e 17 anos) e Níveis de Pobreza

- Este dataset tem n = 51, (50 estados americanos mais o Distrito de Colúmbia). As variáveis são:
  - y = taxa de nascimentos por 1000 meninas de 15 a 17 anos no ano de 2002, e
  - x = taxa de pobreza, que é o percentual da população do estado vivendo em casas com rendas abaixo do nível de pobreza definido pelo governo federal (Fonte dos Dados: Mind On Statistics, 3rd edition, Utts and Heckard)
- Estamos interessados nas seguintes variáveis:
  - Brth15to17 – taxa de nascimento por 1000 meninas de 15 a 17 anos no ano de 2002 — **Variável Resposta**;
  - PovPct – taxa de pobreza — **Variável Explicativa**.

```
library(readr)
poverty_vs_teenbirthrate <- read_table2("~/datasets/poverty_vs_teenbirthrate.txt")
```

Exemplo de Regressão Linear Simples

```
library(ggplot2)
ggplot(data = poverty_vs_teenbirthrate, aes(x = PovPct, y = Brth15to17)) +
    geom_point()
```

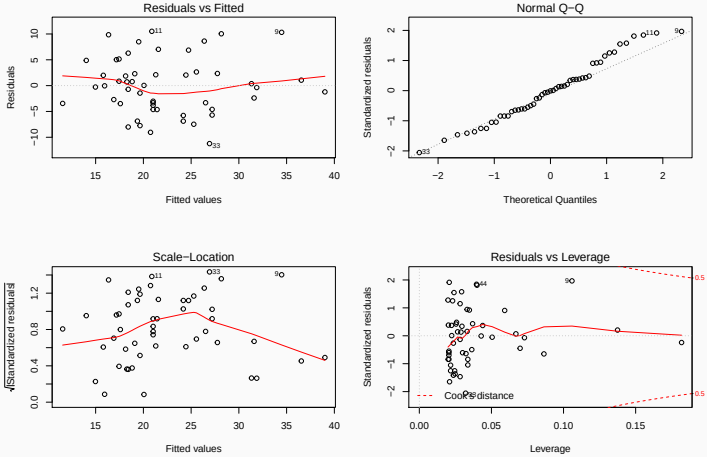


Modelo de Regressão Linear Simples

```
modpoverty <- lm(Brth15to17 ~ PovPct, data = poverty_vs_teenbirthrate)
summary(modpoverty)

Call:
lm(formula = Brth15to17 ~ PovPct, data = poverty_vs_teenbirthrate)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2275  -3.6554  -0.0407   2.4972  10.5152

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.2673     2.5297   1.687    0.098 .
PovPct        1.3733     0.1835   7.483 1.19e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.551 on 49 degrees of freedom
Multiple R-squared:  0.5333,    Adjusted R-squared:  0.5238
F-statistic:    56 on 1 and 49 DF,  p-value: 1.188e-09
```

Diagnóstico do Modelo

```
par(mfrow = c(2, 2))
plot(modpoverty)
```



Usando o pacote `gvlma`

O pacote `gvlma` é uma implementação do artigo de Pena & Slate called "Global Validation of Linear Model Assumptions" e nos permite verificar rapidamente por:

- Linearidade – o teste **Global Stat** testa a hipótese nula de que nosso modelo é uma combinação linear das preditoras.
- Heterocedasticidade – o teste correspondente testa a hipótese nula de que a variância dos nossos resíduos é relativamente constante.
- Normalidade – testa distorções na distribuição dos resíduos (*skewness* e *curtose*), para entendermos se os resíduos do modelo seguem uma distribuição normal. Se a hipótese nula é rejeitada, provavelmente é necessária uma transformação nos dados (p.explo, uma transformação **log**). Podemos observar isso visualmente no *QQ-Plot*.
- *Link Function* – testa se nossa variável dependente é realmente contínua, ou categórica. Se a hipótese nula é rejeitada ($\texttt{p-value} < 0.05$), é uma indicação de que deveríamos utilizar uma forma alternativa do modelo linear generalizado (p.explo, Regressão Logística ou Binomial, etc).

Diagnóstico do Modelo

```
library(gvlma)
diaggvlma <- gvlma(modpoverty)
display.gvlmatests(diaggvlma)

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = modpoverty)

                     Value p-value                Decision
Global Stat        2.9037  0.5741 Assumptions acceptable.
Skewness           0.3643  0.5461 Assumptions acceptable.
Kurtosis           0.9280  0.3354 Assumptions acceptable.
Link Function      0.8987  0.3431 Assumptions acceptable.
Heteroscedasticity 0.7127  0.3985 Assumptions acceptable.
```

Muito Obrigado!