

Regressão Linear no RStudio

Saulo Morellato

Introdução

Objetivo Geral

Estabelecer uma função que descreva a relação entre uma variável contínua Y (variável resposta) e uma ou mais variáveis de apoio X_1, X_2, \dots, X_p (covariáveis) na forma

$$Y = f(X_1, X_2, \dots, X_p) + \epsilon$$

sendo ϵ um erro aleatório.

Erro Aleatório

Possíveis explicações para a presença do erro aleatório no modelo são:

- Caráter vago da teoria;
- Falta de dados disponíveis;
- Caráter aleatório da natureza;
- Escolha equivocada para a forma funcional.

Regressão Linear

Objetivo

- Na Análise de Regressão Linear o objetivo é identificar uma equação linear que permita descrever o comportamento da variável resposta Y utilizando valores conhecidos das covariáveis X_1, X_2, \dots, X_p .
- Ou seja, considera-se que a função $f(\cdot)$ tenha uma forma linear.

Regressão Linear Simples

Temos apenas uma covariável no modelo. Um exemplo seria tentar modelar o quanto as despesas com propaganda influenciam nas vendas de um determinado produto.

- Variável resposta: *vendas*
- Covariável: *gasto com propagandas*

Regressão Linear Múltipla

Temos apenas duas ou mais covariáveis no modelo. Um exemplo seria tentar modelar o quanto as características de um imóvel influenciam no preço de venda do mesmo.

- Variável resposta: *preço do imóvel*
- Covariáveis: *área, no de quartos, no de banheiros, idade,...*

Descrição do Modelo

- A forma funcional é linear;
- Considera-se apenas uma covariável;
- Desse modo, temos

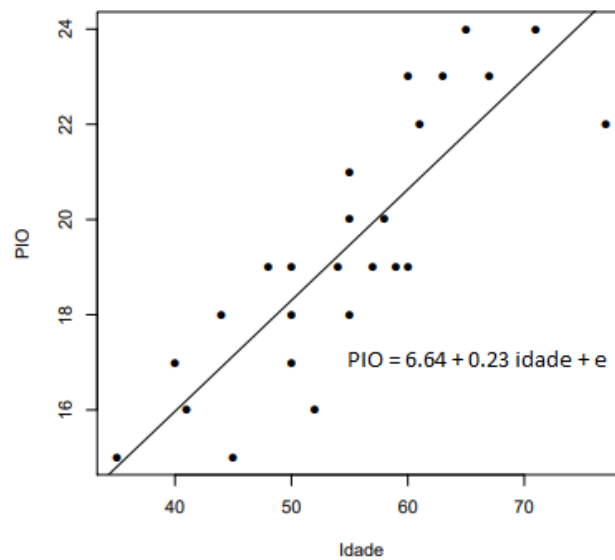
$$Y = \beta_0 + \beta_1 X + \epsilon$$

- β_0 e β_1 são valores desconhecidos (parâmetros) da reta que relaciona X e Y ;
- β_0 é chamado de intercepto; e
- β_1 é o coeficiente angular.

- **Exemplo:** (Clínica Oftalmológica) Acredita-se que a *pressão intra-ocular* é explicada (depende) da *idade*. Relembrando:

| Indivíduo | Idade | PIO | Indivíduo | Idade | PIO |
|-----------|-------|-----|-----------|-------|-----|
| 1 | 35 | 15 | 14 | 55 | 20 |
| 2 | 40 | 17 | 15 | 57 | 19 |
| 3 | 41 | 16 | 16 | 58 | 20 |
| 4 | 44 | 18 | 17 | 59 | 19 |
| 5 | 45 | 15 | 18 | 60 | 23 |
| 6 | 48 | 19 | 19 | 60 | 19 |
| 7 | 50 | 19 | 20 | 61 | 22 |
| 8 | 50 | 18 | 21 | 63 | 23 |
| 9 | 50 | 17 | 22 | 65 | 24 |
| 10 | 52 | 16 | 23 | 67 | 23 |
| 11 | 54 | 19 | 24 | 71 | 24 |
| 12 | 55 | 18 | 25 | 77 | 22 |
| 13 | 55 | 21 | | | |

Gráfico do modelo



Aplicação em R

Carregando Pacotes

- Para exemplificar a estimação de um modelo de Regressão Linear em R vamos utilizar conjunto de dados `dados_imoveis.csv`.
- Para isso primeiramente vamos carregar os pacotes necessários.

```
library(tidyverse) # para organizar os dados
library(gtsummary) # para organizar resultados em tabela
library(mixlm)     # para selecao de variaveis (stepwise)
```

Carregando e Manipulando Dados

- Carregue o arquivo `dados_imoveis.csv` utilizando o comando `read.csv()`.

```
dados<- read.csv("dados_imoveis.csv", header=TRUE)
```

- Dê uma olhada superficial na estrutura dos dados usando o comando `glimpse()`.

```
glimpse(dados)
```

Rows: 20

Columns: 5

```
$ preco    <int> 110000, 210000, 135000, 240000, 130000, 180000, 170000, 125000~
$ area     <int> 140, 165, 145, 230, 120, 195, 200, 105, 115, 110, 210, 250, 80~
$ idade    <int> 8, 9, 25, 6, 13, 34, 2, 24, 26, 35, 10, 12, 40, 28, 12, 6, 26,~
$ quartos  <int> 3, 4, 4, 5, 3, 5, 3, 3, 3, 4, 2, 5, 2, 4, 3, 5, 6, 6, 4, 2
$ piscina  <chr> "nao", "sim", "nao", "sim", "nao", "nao", "sim", "nao", "sim",~
```

- Transforme a variável `piscina` em fator, em seguida Verifique as estatísticas descritivas utilizando o comando `summary()`.

```
dados$piscina<- as.factor(dados$piscina)
summary(dados)
```

| preco | area | idade | quartos | piscina |
|----------------|---------------|---------------|-------------|---------|
| Min. : 75000 | Min. : 80.0 | Min. : 2.00 | Min. : 2.0 | nao: 7 |
| 1st Qu.:133750 | 1st Qu.:123.8 | 1st Qu.: 8.75 | 1st Qu.:3.0 | sim:13 |
| Median :175000 | Median :180.0 | Median :12.50 | Median :4.0 | |
| Mean :175000 | Mean :178.8 | Mean :17.40 | Mean :3.8 | |
| 3rd Qu.:211250 | 3rd Qu.:213.8 | 3rd Qu.:26.00 | 3rd Qu.:5.0 | |
| Max. :300000 | Max. :305.0 | Max. :40.00 | Max. :6.0 | |

Ajustando a Regressão Linear

- Para ajustar/estimar um modelo de Regressão Linear devemos utilizar o comando `lm()`, ao qual devemos fornecer as seguintes informações: fórmula e dados.
- Neste primeiro modelo consideremos que a variável **preco** dependa apenas de **area**.

```
modelo1<- lm(preco ~ area, data=dados)
```

Modelo 2

- Caso eu queira considerar um segundo modelo no qual **preco** dependa de todas as demais variável do conjunto de dados devemos utilizar os seguintes comandos:

```
modelo2<- lm(preco ~ area + idade + quartos + piscina, data=dados)
#modelo2<- lm(preco ~ . , data=dados)      # comando alternativo
```

Visualizando os Modelos

- A função `summary()` pode ser utilizada para visualizarmos um resumo do modelo estimado em forma de tabela.
- Deve-se observar que este comando apresenta os resultados de uma forma um pouco poluída, porém bem completa.
- Apliquemos este comando para o `modelo2`.

Visualização - summary()

```
summary(modelo2)
```

Call:

```
lm(formula = preco ~ area + idade + quartos + piscina, data = dados)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-23897 -12990  -2961   10454   43914
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   49867.3     19127.2   2.607  0.01982 *
area           268.7       141.6    1.898  0.07716 .
idade        -631.1       485.6   -1.300  0.21330
quartos       21647.1     5721.1    3.784  0.00180 **
piscina(sim) -19440.1     6394.2   -3.040  0.00827 **
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

s: 19060 on 15 degrees of freedom

Multiple R-squared: 0.9114,

Adjusted R-squared: 0.8878

F-statistic: 38.59 on 4 and 15 DF, p-value: 9.971e-08

Visualização - tbl_regression()

- Para uma visualização mais interpretável dos resultados, podemos utilizar o comando `tbl_regression()`

```
tbl_regression(modelo2)
```

| Characteristic | Beta | 95% CI | p-value |
|----------------|--------|---------------|---------|
| area | 269 | -33, 570 | 0.077 |
| idade | -631 | -1,666, 404 | 0.2 |
| quartos | 21,647 | 9,453, 33,841 | 0.002 |

Interpretação

- Se a covariável é numérica, o acréscimo de 1 unidade nesta covariável espera-se um acréscimo de β unidades na variável resposta.
- Se a covariável é categórica, o fato de pertencer a certa classe espera-se um acréscimo de β unidades na variável resposta.
- O aumento de 1 unidade em **area** implica, em média, num acréscimo de 268.7 unidades monetárias no **preço**.
- O aumento de 1 unidade em **idade** implica, em média, num decréscimo de 631.1 unidades monetárias no **preço**.
- O aumento de 1 unidade em **quarto** implica, em média, num acréscimo de 21647.1 unidades monetárias no **preço**.
- O fato do imóvel possuir piscina implica, em média, num acréscimo de 38880.2 unidades monetárias no **preço**.

Modelos com Interação

- Em um modelo de regressão o Y pode depender linearmente de X_1 , X_2 e do produto X_1X_2 .
- Desse modo, temos o modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- No R, para considerar um modelo com interação podemos utilizar os seguintes comandos:

```
modelo3<- lm(preco ~ area + idade + area:idade, data=dados)
```

- Caso eu queira considerar todas as interações possíveis das minha covariáveis devo utilizar o seguinte comando:

```
modelo4<- lm(preco ~ .^2, data=dados)
```

Seleção de Variáveis

- Dentre todas as variáveis do conjunto de dados, quais são as que interferem significativamente em `preco`?
- Para determinar tais variáveis utilizamos o comando `backward()`.
- O argumento que devemos utilizar nesta função seria o modelo ajustado com todas as variáveis e/ou todas as interações.
- As variáveis que não são estatisticamente significativas são removidas do modelo.

```
melhor_modelo<- backward(modelo4, alpha=0.1)
```

Backward elimination, alpha-to-remove: 0.1

Full model: `preco ~ area + idade + quartos + piscina + area:idade + area:quartos + area:piscina + idade:quartos + idade:piscina + quartos:piscina`

| | Step | RSS | AIC | R2pred | Cp | F value | Pr(>F) |
|---------------|------|------------|--------|---------|--------|---------|--------|
| idade:piscina | 1 | 1096178359 | 376.39 | 0.89874 | 10.972 | 1.9721 | 0.1938 |

```
tbl_regression(melhor_modelo)
```

| Characteristic | Beta | 95% CI | p-value |
|-----------------|--------|----------------|---------|
| area | -1,467 | -2,243, -691 | 0.002 |
| idade | 147 | -2,253, 2,546 | 0.9 |
| quartos | 31,448 | -6,649, 69,544 | 0.10 |
| area * idade | 36 | 8.9, 64 | 0.015 |
| area * quartos | 218 | 112, 324 | <0.001 |
| idade * quartos | -2,063 | -3,509, -617 | 0.010 |

Predição

- Suponha que 2 imóveis estão para ser vendidos.
- Suponha ainda não haver preço de venda para estes 2 imóveis.
- Utilize as características destes, juntamente com o modelo estimado, para estimar seus preços.
- Carregue o arquivo `novos_imoveis.csv` utilizando o comando `read.csv()`.


```
novos<- read.csv("novos_imoveis.csv", header=TRUE)
```

- Faça a predição/estimção de preço para estes 2 imóveis utilizando o comando `predict()`.

```
preditos<- predict(modelo2, newdata=novos)
```

- Para uma melhor visualização vamos concatenar as informações dos novos imóveis com suas respectivas predições/estimções de preço.

```
cbind.data.frame(novos, preditos)
```

| | area | idade | quartos | piscina | preditos |
|---|------|-------|---------|---------|-----------|
| 1 | 100 | 5 | 2 | nao | 97432.01 |
| 2 | 200 | 31 | 4 | sim | 190063.05 |