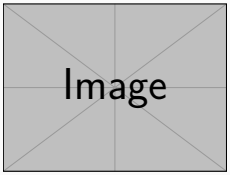
Regresssão Linear - Introdução

Mário Olímpio de Menezes

Maio – 2020

About me

Me



- Físico
- Professor e Pesquisador
- momenezes
- mariomenezes

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 *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*.

Modelos

Um estatístico britânico chamado George Box escreveu:

> “Todos os modelos são errados, alguns são úteis.”

O que é um modelo

- Um modelo é uma abstração da realidade;
- Necessariamente deixa de lados alguns aspectos *menos relevantes*;
- É uma simplificação proposital para um propósito específico.

Estamos vivendo uma explosão de modelos:

- Modelos Epidemiológicos
- Modelos de crescimento da economia
- Modelos financeiros
- Modelos, modelos e modelos

- Uma das coisas mais sensíveis e importantes quando se começa é a escolha do tipo correto de análise estatística. A escolha depende:
  - da natureza dos dados
  - da questão que se quer responder, entre outras coisas.
- A chave é entender que tipo de variável *resposta* você tem e saber a natureza de suas variáveis *explicativas*.
  - A variável *resposta* é coisa com a qual você está trabalhando:
    - é a variável cuja variação você está tentando entender!
    - é a variável que você está tentando prever;
    - é a que vai no eixo *y* do gráfico.
  - A variável *explicativa* vai no eixo *x* do gráfico.
    - você está interessado em entender como a variação da variável *resposta* está associada com a variação da variável *explicativa*.

- Você também precisa considerar o *modo* que as variáveis na sua análise medem o que elas se propõem a medir.
- Uma medida contínua é uma variável do tipo altura ou peso que pode assumir valores com números reais.
- Uma variável categórica é um fator com dois ou mais níveis:
  - sexo é um fator com dois níveis (masculino e feminino)
  - cor pode ser um fator com sete níveis (vermelho, laranja, amarelo, verde, azul, índigo e violeta)
- Portanto, é essencial responder às seguintes questões:
  - Qual das variáveis é **a variável resposta**?
  - Quais são as variáveis explicativas?
  - As variáveis explicativas são contínuas ou categóricas, ou uma mistura de ambas?
  - Que tipo de variável resposta temos:
    - é uma medida contínua? uma contagem? uma proporção? um tempo (ocasião) de morte? ou uma categoria?

Variáveis

Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type
ritz	2014	3.35	5.59	27000	Petrol
sx4	2013	4.75	9.54	43000	Diesel
ciaz	2017	7.25	9.85	6900	Petrol

age	sex	bmi	children	smoker	region	charges
19	female	27.90	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.552
28	male	33.00	3	no	southeast	4449.462

LOW	LWT	RACE
0	182	2
0	155	3
0	105	1

Codificação de Variáveis

- Uma coisa que é preciso atenção nos *datasets* é a codificação de variáveis categóricas utilizando-se números, como no último *dataset* mostrado.
- A variável **RACE** é categórica, mas está codificada com números (1,2,3, . . .).
- Ao se manipular estes dados é preciso cuidado para que esta variável seja tratada como categórica e não como numérica.

Escalas de Mensuração

- Nominal – usada para rotular variáveis, sem valor quantitativo
- Ordinal – nesta escala, a ordem dos valores é o que é importante e significante, mas as diferenças entre os níveis não é conhecida.
- Intervalar – é uma escala numérica, na qual nós conhecemos tanto a ordem como a diferença exata entre os valores. A escala intervalar não tem um zero significativo, isto é, o zero não significa “ausência”, mas é simplesmente mais um intervalo na escala. Isto limita as operações possíveis com esta escala.
- Razão – esta é a escala completa: conhecemos a ordem, a diferença exata entre os valores e ela tem um zero significativo, ou seja, o zero significa realmente “ausência” daquela característica, o que nos permite realizar todas as operações.

Método Estatístico Apropriado

Algumas *chaves* simples para a escolha do método estatístico apropriado

As variáveis explicativas

1. Todas as variáveis explicativas são contínuas ⟹ **Regressão**
    - É possível realizar uma Regressão com variáveis explicativas contínuas e categóricas (transformando)
2. Todas as variáveis explicativas são categóricas ⟹ **Análise de Variância (ANOVA)**
3. Variáveis explicativas são tanto contínuas como categóricas ⟹ **Análise de Covariância (ANCOVA)**

A variável resposta

1. Contínua ⟹ **Regressão Normal, ANOVA ou ANCOVA**
2. Proporção ⟹ **Regressão Logística**
3. Contagem ⟹ **Modelos log-linear**
4. Binária ⟹ **Análise logística binária**
5. Tempo na morte ⟹ **Análise de sobrevivência**

Objetivo da Modelagem Estatística

- Determinar os valores dos parâmetros em um modelo específico que *levam ao melhor ajuste do modelo aos dados*
- Os dados são *sacrosantos*; eles nos dizem o que realmente aconteceu sob determinadas circunstâncias.
  - É um erro comum dizer “os dados foram ajustados ao modelo” como se os dados fossem flexíveis, e nós tivéssemos uma estrutura clara do modelo.
  - É o contrário: o que se procura é o modelo **mínimo adequado** que descreva os dados.
  - O modelo é ajustado aos dados; não o contrário!
- O melhor modelo é o que produz o mínimo de variação não explicada (o *mínimo desvio dos resíduos*), sujeito à restrição de que todos os parâmetros no modelo devem ser estatisticamente significantes

- A melhor coisa a fazer é gastar um tempo substancial, logo de início, para entender os dados e o que eles mostram.
  - Isto vai ajudar a guiar o pensamento para a modelagem estatística mais apropriada.
- **Thinking with Data** – Max Shron
  - Scoping: Why Before How
    - *“Most people start working with data from exactly the wrong end. They begin with a data set, then apply their favorite tools and techniques to it. The result is narrow questions and shallow arguments. Starting with data, without first doing a lot of thinking, without having any structure, is a short road to simple questions and unsurprising results. We don’t want unsurprising – we want knowledge.”*

Checklist da Modelagem Estatística

Checklist

- Certificar-se de que o `data.frame` está correto em estrutura e conteúdo:
  - Todos os valores de cada variável estão na mesma coluna?
    - **tidy data**
  - Todos os zeros são realmente 0 ou deveriam ser `NA`?
  - Cada linha contém o mesmo número de entradas?
  - Existe algum nome de variável que contém espaço?
- Depois de carregar os dados, a Análise Exploratória de Dados é **essencial**

Checklist da Modelagem Estatística

Sobre o Modelo

- Algumas coisas básicas na escolha do modelo
  - Quais variáveis explicativas deveriam ser incluídas?
  - Transformação da variável resposta é necessária?
  - Interações deveriam ser incluídas?
  - Termos não lineares deveriam ser incluídos? ($X^2, X^3 ...$)
  - As variáveis explicativas deveriam ser transformadas?
- Tente utilizar o tipo mais simples de análise que seja apropriado para seus dados e para a questão que está tentando responder.
- Ajuste um modelo máximo e vá simplificando-o paulatinamente ao remover parâmetros.
- Faça o *diagnóstico do modelo*
- Por fim, documente tudo o que fizer, e explique cada um dos passos. Desta maneira você entenderá o que fez e porque fez quando retornar à sua análise 6 meses mais tarde!

Especificando o modelo

- Um modelo incorpora nosso entendimento mecanicista das variáveis explicativas envolvidas, e da maneira que elas estão relacionadas com a variável resposta.
- Buscamos um modelo **mínimo** por conta do princípio da *parcimônia*[1], e também um modelo **adequado**
- É muito importante entender que *não há **um** modelo*.
  - em muitos casos, haverá um grande número de modelos diferentes, uns mais plausíveis do que outros.
- É preciso determinar quais, se algum, dos modelos possíveis, são adequados
  - e depois, dos adequados, qual é o modelo *mínimo adequado*.
  - pode haver um conjunto de modelos que descrevem os dados igualmente bem (ou de modo igualmente probre se a variabilidade for grande)

[1] Procure por "Navalha de Ocam" para entender mais sobre este princípio da Parcimônia

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

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

Método dos 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

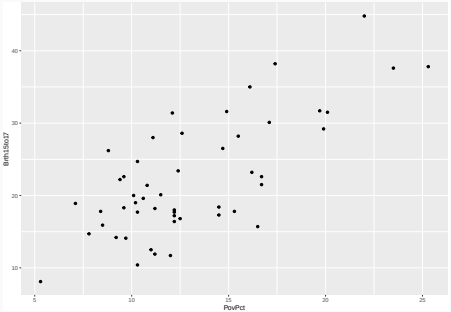
Dataset: Taxa de Nascimentos (mães entre 15 e 17 anos) e Níveis de Pobreza

- 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**.

Exemplo de Regressão Linear Simples

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

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

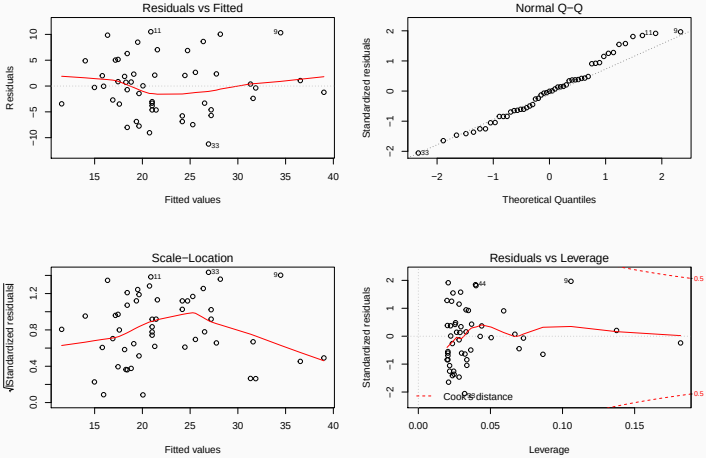


Modelo de Regressão Linear Simples

```
> 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 (`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)
> gvlma::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.
```

Interpretando os resultados do modelo

```
> 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
```

```
> summary(poverty_vs_teenbirthrate[, c("PovPct",
    "Brth15to17")])
     PovPct        Brth15to17
 Min.   : 5.30   Min.   : 8.10
 1st Qu.:10.25   1st Qu.:17.25
 Median :12.20   Median :20.00
 Mean   :13.12   Mean   :22.28
 3rd Qu.:15.80   3rd Qu.:28.10
 Max.   :25.30   Max.   :44.80
```

Interpretando os resultados do modelo

A cada aumento de uma unidade no índice de pobreza, temos um aumento de 1.3733 pontos percentuais na taxa de nascimentos por 1000 meninas de 15 a 17 anos.

No modelo que obtivemos, o *Intercepto* não tem significância estatística (`p-value` = 0.098). Portanto, o modelo final pode ser escrito como:

$$\text{Brth15to17} = 1.3733 \times \text{PovPct}$$

Próximo Bloco

Regressão Linear Multivariada