Estatística não paramétrica

Aula 29

Manoel Santos-Neto
Atualização: 05 de dezembro de 2023

O que você irá aprender nesta aula?

1. Estimação Robusta.

Introdução

Desde 1960, muitos esforços teóricos foram dedicados ao desenvolvimento de procedimentos estatísticos que sejam resistentes a pequenos desvios das suposições, ou seja, robustos em relação a valores atípicos e estáveis em relação a pequenos desvios do modelo paramétrico assumido. De fato, é bem conhecido que os procedimentos ótimos clássicos não se comportam bem sob ligeiras violações das rigorosas suposições do modelo.

Para lidar com os problemas decorrentes da violação das suposições paramétricas, as estatísticas robustas podem ser vistas como uma extensão das estatísticas paramétricas clássicas, considerando diretamente as divergências dos modelos. Enquanto os modelos paramétricos podem ser uma boa aproximação da verdadeira situação subjacente, as estatísticas robustas não pressupõem que o modelo seja exatamente correto. Um procedimento robusto, conforme declarado por Huber e Ronchetti (2009), deve, portanto, apresentar as seguintes características:

1. Deve estimar eficientemente o modelo assumido.
2. Deve ser confiável e razoavelmente eficiente sob pequenas divergências do modelo (por exemplo, quando a distribuição está em uma vizinhança do modelo assumido).
3. Divergências maiores do modelo não devem afetar excessivamente o procedimento de estimação.

Introdução

Isso significa que as estatísticas robustas são projetadas não apenas para lidar com a incerteza inerente aos modelos paramétricos, mas também para manter sua eficácia quando as condições do modelo são levemente violadas ou até mesmo quando há desvios substanciais. Esse enfoque torna as estatísticas robustas uma ferramenta valiosa em situações do mundo real, onde os dados podem não se conformar estritamente aos pressupostos teóricos.

Muitas vezes, acredita-se que procedimentos robustos podem ser evitados ao seguir o seguinte procedimento em duas etapas:

1. Limpar os dados usando alguma regra para rejeição de valores atípicos.
2. Aplicar procedimentos ótimos clássicos nos dados "limpos".

Infelizmente, tais procedimentos não podem substituir métodos robustos, como discutido por Huber e Ronchetti (2009), pelas seguintes razões:

- Raramente é possível separar as duas etapas. Por exemplo, em um cenário de regressão paramétrica, é difícil reconhecer valores atípicos sem estimativas confiáveis (ou seja, robustas) dos parâmetros do modelo.

Introdução

- Os dados "limpos" não corresponderão ao modelo assumido, pois haverá erros estatísticos de ambos os tipos (aceitação falsa e rejeição falsa). Portanto, em geral, a teoria clássica não é aplicável à amostra limpa.
- Empiricamente, os melhores procedimentos de rejeição não alcançam o desempenho dos melhores procedimentos robustos. Estes últimos são aparentemente superiores porque conseguem fazer uma transição mais suave entre a aceitação completa e a rejeição completa de uma observação usando procedimentos de ponderação, como indicado por Hampel et al. (1987).
- Estudos empíricos também mostraram que muitos dos métodos clássicos de rejeição são incapazes de lidar com múltiplos valores atípicos. De fato, é possível que um segundo valor atípico "mascare" o efeito do primeiro, de modo que nenhum deles seja rejeitado.

Estimador Clássico de Mínimos Quadrados

Podemos definir o modelo de regressão linear padrão da seguinte forma. Sejam $(\mathbf{x}_i, y_i) : i = 1, \ldots, n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, tal que:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + u_i,$$

em que $y_i \in \mathbb{R}$ é a i-ésima observação, $\mathbf{x}_i \in \mathbb{R}^p$ é a i-ésima linha da matrix de planejamento $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}$ é uma vetor $p \times 1$ de parâmetros desconhecidos, $u_i \in \mathbb{R}$ é o i-ésimo erro.

A estimativa de mínimos quadrados $\hat{\beta}_{MQ}$ de β pode ser expressa como um M-estimador (são obtidos como o mínimo de somas de funções dos dados). Os estimadores de mínimos quadrados são um exemplo da classe maior de M-estimadores. A definição de M-estimadores foi motivada pela estatística robusta, que trouxe novos tipos de M-estimadores, definidos pela equação de estimação:

$$\sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta) x_i = 0. \tag{1}$$

Esta estimativa é ótima sob as seguintes suposições:

- u_i são normalmente distribuídos;

Estimador Clássico de Mínimos Quadrados

- $E[u_i] = 0, \text{ para } i = 1, \ldots, n,$
- $\text{Cov}(u_1, \ldots, u_n) = \sigma^2 I_n,$

Em outras palavras, a estimação de mínimos quadrados é ótima apenas quando os erros são normalmente distribuídos. Pequenos desvios da suposição de normalidade para os erros resultam em uma considerável perda de eficiência do estimador de mínimos quadrados (para maiores detalhes ver: Hampel et al. (1987), Huber (1973) e Hampel (1973)).

Estimadores robusto para modelos de regressão

O "Estimador de Huber", introduzido em Huber (1973), foi um dos primeiros métodos de estimação robustos aplicados a modelos lineares. Basicamente, este estimador é uma versão ponderada da estimativa de mínimos quadrados com pesos da forma:

$$w_i = \min\left(1, \frac{c}{|r_i|}\right),$$

em r_i é o i-ésimo resíduo e c é uma constante positiva que controla a compensação entre robustez e eficiência.

Huber propôs um M-estimador $\hat{\beta}_H$ de β definido pela equação de estimação:

$$\sum_{i=1}^{n} \psi_c(y_i - \mathbf{x}_i^\top \beta) x_i = 0,$$

em que $\psi_c(\cdot)$ corresponde à função de peso de Huber:

$$w(x) = \begin{cases} 1, & \text{se } |x| \leq k; \\ \frac{k}{|x|}, & \text{se } |x| > k, \end{cases}$$

Estimadores robusto para modelos de regressão

e, portanto, é definida como:

$$\psi_c(r) = \begin{cases} r, & \text{se } |r| \leq c; \\ c \cdot \text{sign}(r), & \text{se } |r| > c, \end{cases}$$

em que $\text{sign}(x)$ retorna valores -1, 0 e 1 para valores de x negativo, zero e positivo, respectivamente.

No entanto, o estimador de Huber não consegue lidar com problemas causados por pontos atípicos na matriz de planejamento (ou de covariáveis) $\mathbf{X}$. Para detalhes de outros M-estimadores ver Thiago Macêdo (2014). Além do M-estimador de Huber, o autor apresenta os M-estimadores de Tukey bisquare e de Hampel. Estes estimadores podem ser obtidos usando a função `rlm()` do pacote `MASS` do `R`. Também tem a função `lmrob()` do pacote `robustbase` que usa o método proposto por Koller e Stahel (2011). Por fim, existe a função `lmRob()` do pacote `robust`.

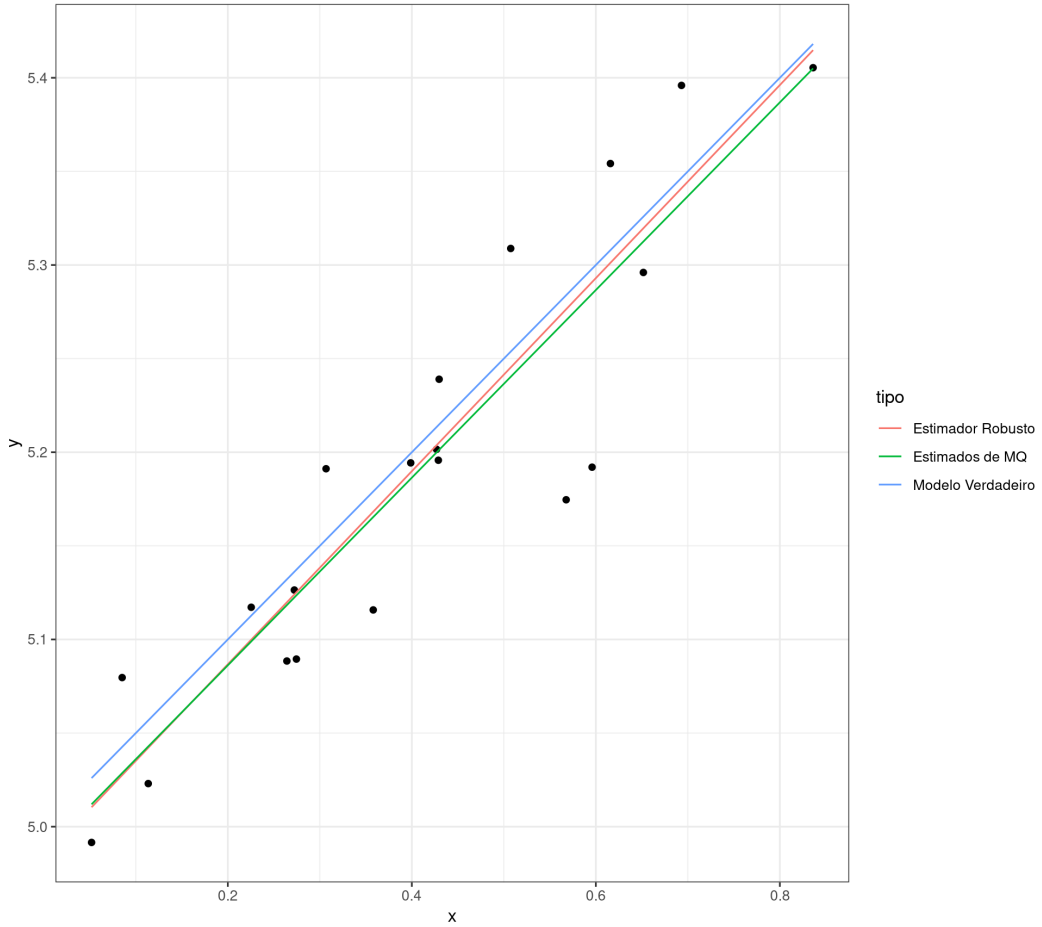
Aplicação de Estimação Robusta

```r
library(MASS)
library(ggpubr)
# Tamanho da amostra
n <- 20

# Parametros do modelo
beta0 <- 5    # Intercepto
beta1 <- 0.5   # inclinacao
sigma <- .05     # variancia

#construindo variavel resposta sem perturbacao
set.seed(10)
x <- runif(n)
y <- beta0 + beta1*x + rnorm(n,0,sigma)

y_true <- beta0 + beta1*x
```



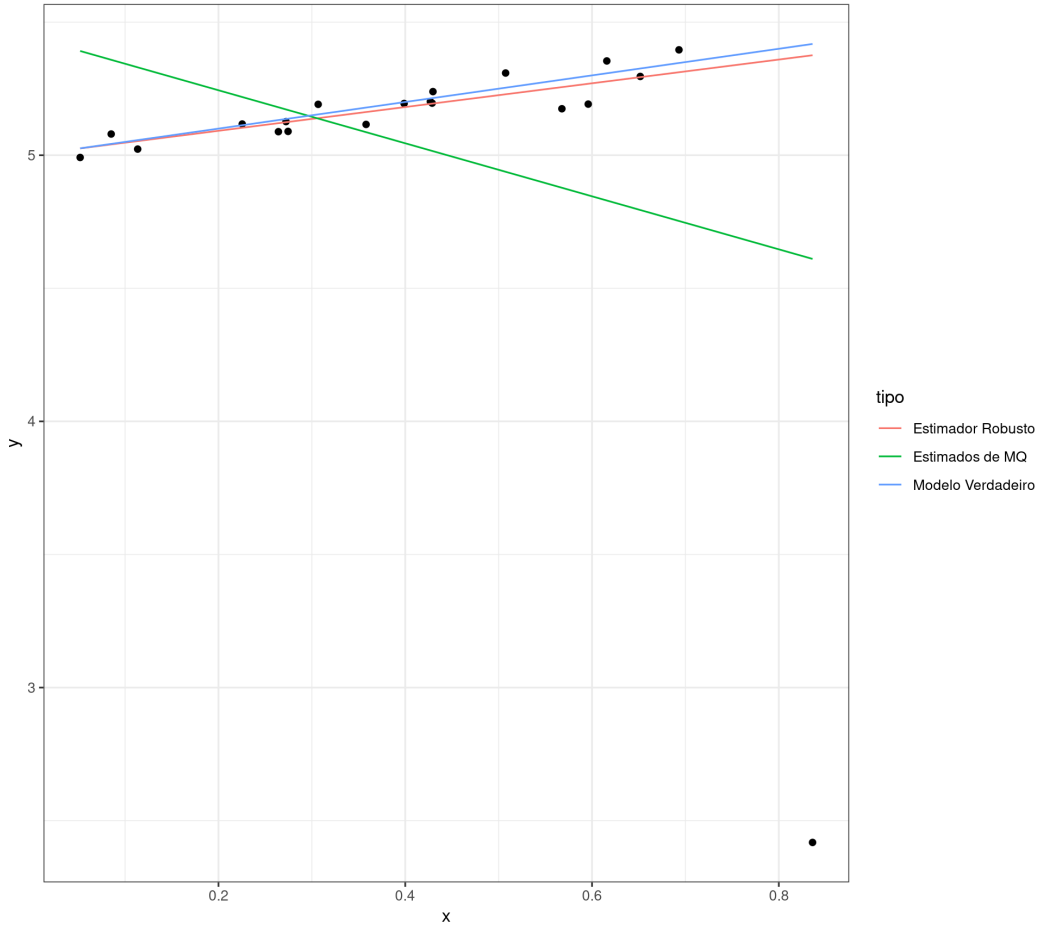
Aplicação de Estimação Robusta

```r
library(MASS)
library(ggpubr)
# Tamanho da amostra
n <- 20

# Parametros do modelo
beta0 <- 5    # Intercepto
beta1 <- 0.5   # inclinacao
sigma <- .05     # variancia

#construindo variavel resposta sem perturbacao
set.seed(10)
x <- runif(n)
y <- beta0 + beta1*x + rnorm(n,0,sigma)

y_true <- beta0 + beta1*x
```



Exercício

Ajuste o conjunto de dados `data(starsCYG, package = "robustbase")` considerando os estimadores de mínimos quadrados e robusto.

```
library(ggplot2)
data(starsCYG, package = "robustbase")
ggplot(starsCYG, aes(x = log.Te, y = log.light )) + geom_point() + theme_bw()
```