Estatística computacional - Trabalho 1

Arthur Sonntag Kuchenbecker

July 2019

1 Introdução

O trabalho consiste em comparar os estimadores de mínimos quadrados e mínimos absolutos. A comparação será feita por meio de boxplots que mostram os erros da estimação para diferentes tamanhos de amostra.

Além disso, uma seção será dedicada a estudar os impactos - nos dois estimadores utilizados - causados pela existência de outliers.

2 Criação dos dados

Os dados utilizados são formados por um vetor $\vec{x}$ de variáveis independentes construído da seguinte forma:

$$x_1 = 0 \quad \text{e} \quad x_i = x_{i-1} + \frac{10}{n} \tag{1}$$

em que n é o tamanho da amostra.

E por um vetor $\vec{y}$ de variáveis dependentes, dado por:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2}$$

em que $i = 1, 2, \ldots, n$ e ϵ_i é o i-ésimo elemento do vetor $\vec{\epsilon}$ (vetor de erros) que possui distribuição normal com média 0 e desvio-padrão σ, ou distribuição t de student com v graus de liberdade (os parâmetros serão estimados para ambos os casos).

$$\epsilon \sim N(0, \sigma) \tag{3}$$

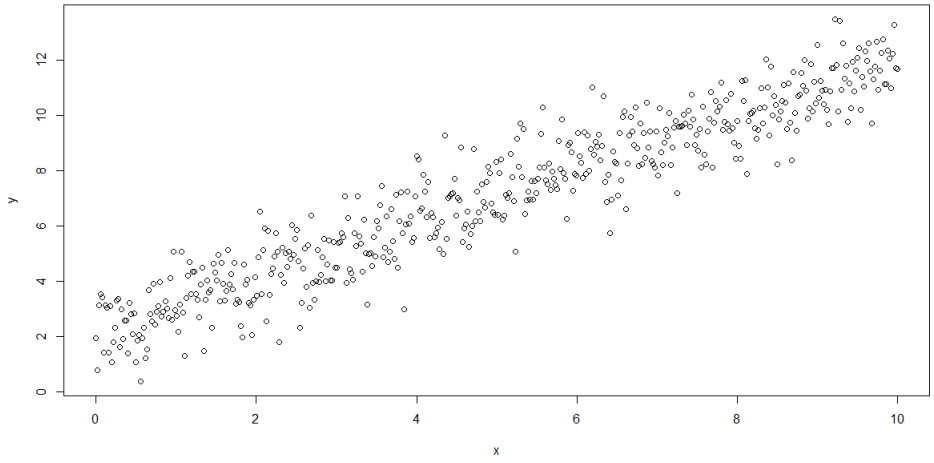
ou

$$\epsilon \sim t(v) \tag{4}$$

Ainda, β_0 e β_1 são os parâmetros da equação linear e mantidos fixos ao longo de todo o exercício em 2 e 1, respectivamente.

$$\beta_0 = 2, \quad \beta_1 = 1 \tag{5}$$

De forma que o gráfico de x por y apresenta a seguinte forma (no caso em que $\epsilon \sim N(0, \sigma)$ e o tamanho da amostra é igual a 500)



3 Tamanho da amostra e parâmetros

O exercício será realizado para três tamanhos diferentes de amostra:

$$n = 20, 100, 500 \tag{6}$$

Para a distribuição normal dos erros, o desvio-padrão (σ) é igual a 1 e para a distribuição t, o número de graus de liberdade (v) é igual a 4.

4 Estimação

Utilizando da estrutura apresentada nas seções anteriores, o exercício consiste em obter 10 mil amostras aleatórias para cada um dos três tamanhos e estimar β_1 via mínimos quadrados ordinários (MQO) e minimos desvios absolutos (MDA).

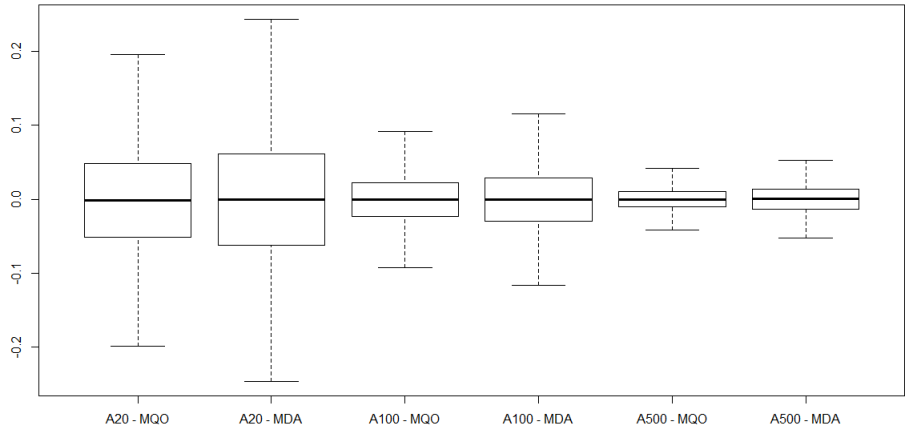
A estimação via MQO é feita pela função `lm()` do `R`. A estimação via MDA, por outro lado, é feita por meio da função abaixo.

```
mabs <- function(x,y){
     b = c(1,1)
     sda <- function(b,x,y){
     sum(abs(y-b[1]-b[2]*x))
   }
     res = optim(b,sda,x=x,y=y)
   res$par
}
```

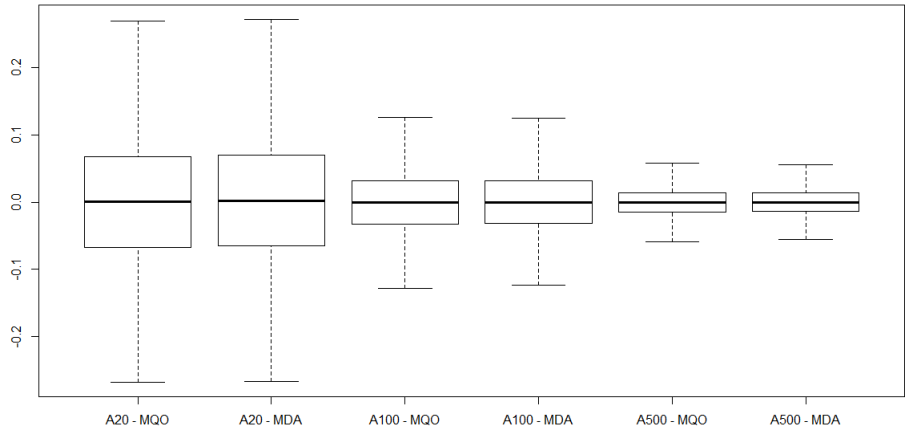
5 Resultados

Os resultados encontrados mostram que - para erros com distribuição normal - o estimador de MQO é menos viesado que o estimador de MDA para todos os tamanhos de amostra.

Os boxplots abaixo mostram a distribuição (10 mil valores) da diferença entre o valor real de β_1 e os estimadores calculados ($\hat{\beta_1}$) para os três tamanhos de amostra diferentes (20, 100 e 500).



Para o caso em que os erros do modelo linear seguem distribuição t de student, ambos os estimadores apresentam vies muito semelhante.



Outra conclusão derivada da análise dos boxplots é a de que quanto maior o tamanho da amostra, menor o vies dos estimadores.

6 Outliers

O último passo do exercício é analisar (para ambas as distribuições e para todos os tamanhos de amostra) qual o impacto da existência de outliers nas amostras.

No código, os outliers foram embutidos nas amostras da seguinte forma:

```
# Colocando os outliers
y[n[i]/2] = 5*y[n[i]/2]
y[n[i]*3/4] = 10*y[n[i]*3/4]
```

Ou seja, para cada tamanho de amostra (`n[i]`), foram embutidos outliers multiplicando a observação $\frac{n}{2}$ por 5 e a observação $\frac{n \times 3}{4}$ por 10.

Os boxplots abaixo mostram que, para ambas as ditribuições, o estimador de MDA é menos viesado na presença de outliers.