Estatística Computacional

Norah Jones

2024-10-02

Índice

1 Introdução: Problemas de interesse

Encontrar soluções de equações não lineares onde não é possível obter uma solução analítica;

Obter integrais que apresentam uma forma complicada que inviabiliza encontrar uma solução analítica;

Gerar artificialmente amostras a partir de modelos estatísticos;

Aplicar a metodologia estudada na resolução de problemas de inferência.

2 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

Parte I

Solução Numérica de Equações

Motivação - O Estimador de Máxima Verossimilhança

No que segue o termo densidade, significa ou uma densidade de probabilidade (caso absolutamente contínuo) ou uma função de probabilidade (caso discreto).

Sejam $X_1, \ \ldots, \ X_n \overset{iid}{\sim} f(.|\theta), \ \theta \in \ \Theta$, onde $f(.|\theta)$ é uma densidade, θ é um parâmetro que desejamos estimar e Θ é o espaço paramétrico;

Suponha que observamos os valores $x_1, \ \ldots, \ x_n$. A função de verossimilhança é definida por:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta), \ \theta \in \Theta$$

A função de log-verossimilhança é dada por:

$$l(\theta) = log L(\theta) = \sum_{i=1}^{n} log f(x_i|\theta), \ \theta \ \in \Theta$$

Seja $\hat{\theta} \ \in \Theta$ um valor do parâmetro que maximiza a função de verossimilhança, ou seja, tal que

$$L(\hat{\theta}) \ \geq L(\hat{\theta}), \text{ para todo } \theta \ \in \ \Theta$$

Então dizemos que $\hat{\theta}$ é uma estimativa de máxima verossimilhança de θ.

A interpretação no caso discreto: é mais provável que $\hat{\theta}$ tenha gerado os dados $x_1, \ \ldots, \ x_n$

Como $\hat{\theta}$ depende da amostra, escrevemos $\hat{\theta}(x_1, \ \ldots, x_n)$. Neste caso, $\hat{\theta}(X_1, \ \ldots, \ X_n)$ é o estimador de máxima verossimilhança (EMV).

Nota

Para cada amostra observada $\mathbf{x} = \ (x_1, \ \ldots, x_n)$
A definição nos diz que $\hat{\theta}(\mathbf{x})$ é um ponto de máximo global. Podemos ter nenhum ou mais de um máximo global
Suponha que Θ é um intervalo e que o ponto $\hat{\theta}$ é um ponto interior de Θ que é ponto de máximo de L, podendo ser um máximo local. Se L tem derivada em $\hat{\theta}$, então $L'(\hat{\theta}) = 0$. Ou seja, $\hat{\theta}$ é um ponto estacionário de L (também dizemos $\hat{\theta}$ é um zero da função L'). Este resultado é conhecido no Cálculo como Teorema de Fermat para Pontos Estacionários.
Ou seja, sob as condições acima, se $\hat{\theta}$ for EMV, então a derivada de L se anula neste ponto. A recíproca pode não ser verdadeira.

Assim, em muitos casos encontrados na prática, encontrar o EMV é um problema relacionado a encontrar soluções em θ para a equação $L'(\hat{\theta}) = 0$ ou $l'(\theta) = 0$.

Por exemplo, considere uma amostra aleatória $X_1, \dots, X_n$ proveniente de uma distribuição $exp(\theta)$. Assim, cada X_i tem densidade de probabilidade.

$$f(x|\theta) = \begin{cases} \theta \exp(-\theta x), & \text{se} \quad x > 0 \\ 0, & \text{se} \quad x \leq 0. \end{cases}$$

O espaço paramétrico é $\Theta = (0, \infty)$. Temos então a seguinte função de log-verassimilhança:

$$\ell(\theta) = n \log \theta - \theta \sum_{i=1}^{n} x_i, \;\; \theta > 0,$$

implicando em

$$\ell'(\theta) = n/\theta - \sum_{i=1}^{n} x_i, \;\; \theta > 0.$$

A solução da equação $\ell'(\theta) = 0$ é dada por $\hat{\theta} = n/\sum_{i=1}^{n} x_i$

```
theta = 5
amostra = rexp(100, rate = theta); amostra
```

```
  [1] 0.9819075856 0.3108157889 0.1077715725 0.1509913161 0.2214408448
  [6] 0.1128802529 0.0695914217 0.1430017840 0.3271275244 0.1516832099
 [11] 0.3623297386 0.1080711435 0.1198464817 0.6156150192 0.0105449191
 [16] 0.0637782078 0.1279208741 0.6511004289 0.0424126700 0.1905779742
 [21] 0.2289686710 0.1197245879 0.0108151906 0.0208573886 0.3454194378
 [26] 0.0890529274 0.1498659069 0.3298726216 0.0614424960 0.0441634457
 [31] 0.0181366200 0.1207487039 0.1969123967 0.8447289075 0.2285378680
 [36] 0.0488103906 0.5109166761 0.1945086909 0.0182187333 0.2289024340
 [41] 0.2021307845 0.0483779819 0.0317972994 0.1542048489 0.2284189850
 [46] 0.0259389349 0.0756173474 0.0593180646 0.1795119539 0.1805451056
 [51] 0.1089347670 0.1447679678 0.2638754776 0.0215398431 0.0036816819
 [56] 0.0979081987 0.1059629615 0.0724566999 0.2224534795 0.0282022174
 [61] 0.0099313338 0.0128306695 0.0093737623 0.2763081971 0.0191507578
 [66] 0.2410981348 1.4398879013 0.1985839741 0.3858103152 0.1486711245
 [71] 0.3376318971 0.5398133285 0.1112120789 0.2705229213 0.4506878640
 [76] 0.0712205910 0.2828466508 0.0358083900 0.0167620294 0.0183636387
 [81] 0.0783255087 0.1114569452 0.3961382590 0.0837270088 0.3013804176
```

```
 [86] 0.4743980836 0.0402212932 0.1020970868 0.4559471275 0.1861280639
 [91] 0.0260237844 0.0134339258 0.3642277077 0.0151143121 0.1437257986
 [96] 0.0006289349 0.0855163206 0.0200865351 0.1300846416 0.0534388816
```

```r
loglik <- function(theta = NULL) {
    return(length(amostra)*log(theta) - theta*sum(amostra))
}

score_exp = function(theta = NULL) {
    n = length(amostra)

    return(n/theta - sum(amostra))
}

theta_x <- seq(from = 1, to = 10, by = 0.1)  # Sequência de valores de -10 a 10 com incremen

y <- loglik(theta = theta_x); y
```

```
 [1] -18.894274 -11.252683  -4.440973   1.673871   7.195241  12.205100
 [7]  16.769525  20.942560  24.768974  28.286269  31.526171  34.515760
[13]  37.278334  39.834083  42.200617  44.393389  46.426033  48.310638
[19]  50.057975  51.677680  53.178408  54.567963  55.853405  57.041144
[25]  58.137013  59.146339  60.073999  60.924469  61.701867  62.409988
[31]  63.052342  63.632175  64.152503  64.616126  65.025650  65.383508
[37]  65.691972  65.953165  66.169078  66.341580  66.472423  66.563258
[43]  66.615640  66.631032  66.610818  66.556304  66.468727  66.349258
[49]  66.199005  66.019021  65.810305  65.573808  65.310433  65.021039
[55]  64.706448  64.367439  64.004759  63.619119  63.211200  62.781653
[61]  62.331099  61.860135  61.369332  60.859237  60.330375  59.783250
[67]  59.218345  58.636126  58.037039  57.421514  56.789965  56.142790
[73]  55.480372  54.803080  54.111272  53.405290  52.685467  51.952122
[79]  51.205564  50.446092  49.673995  48.889551  48.093031  47.284695
[85]  46.464797  45.633580  44.791283  43.938134  43.074357  42.200167
[91]  41.315773
```

```r
maximo_estimado = optimise(loglik, interval = c(1, 10), maximum = T)
y_score = score_exp(theta = theta_x)
```

```r
#Plotando a log-verossimilhança
plot(theta_x, y, type = "l", col = "blue", lwd = 2,
     main = paste("Gráfico da log-verossimilhança de: f(x) =", theta, "exp(-", theta, "x)"),
```

```
     xlab = "x", ylab = "f(x)")
abline(v=theta, col="blue")
abline(v=maximo_estimado, col="red")

#Plotando a função score
plot(theta_x, y_score, type = "l", col = "blue", lwd = 2,
     main = paste("Gráfico da funcao score de: f(x) =", theta, "exp(-", theta, "x)"),
     xlab = "x", ylab = "f(x)")
abline(h=0, col="red")
abline(v=maximo_estimado, col="red")
```

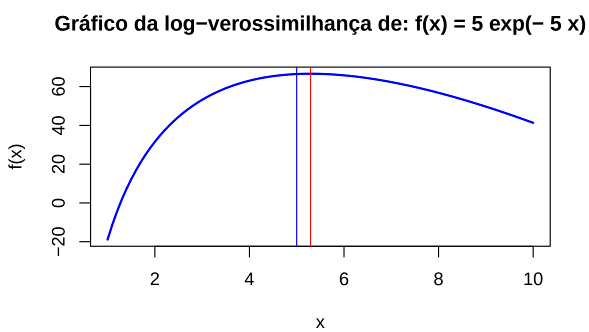

Figura 2.1: log-verossimilhança1

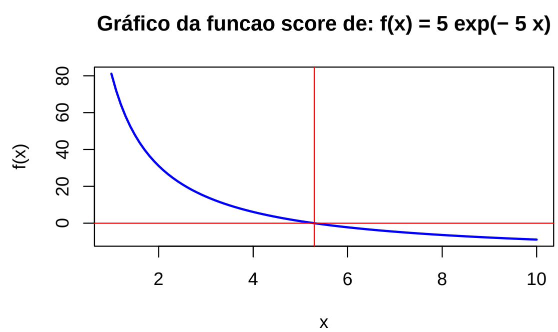

Figura 2.2: Função score

A função $S(\theta) = \ell'(\theta)$, $\theta \in \Theta$, é denominada função escore. Assim, em geral, encontrar o EMV é equivalente a resolver em θ a equação $S(\theta) = 0$.

No exemplo acima obtivemos uma solução analítica para esta equação. Mas em algumas situações isto não é possível.

Considere o seguinte exemplo (Bolfarine e Sandoval, 2010): sejam $X_1, \dots, X_n \overset{iid}{\sim} f(.|\theta)$, onde

$$f(x|\theta) = \begin{cases} \frac{1}{2}(1+\theta x), & \text{se} \quad x > 0 \\ 0, & \text{se} \quad x \leq 0. \end{cases} \tag{2.1}$$

Onde $-1 < \theta < 1$ verifique que $f(\cdot|\theta)$ é uma densidade.

Então é possível mostrar que

$$S(\theta) = \sum_{i=1}^{n} \frac{x_i}{1+\theta x_i}.$$

No entanto, não há solução analítica para $S(\theta) = 0$. Os gráficos a seguir mostram o comportamento das funções de log-verossimilhança (à esquerda) e escore (à direita). A amostra utilizada foi gerada artificialmente a partir do modelo (Equação 2.1), com $n = 40$ e $\theta = -0,35$ (como gerar esta amostra? voltaremos em breve a este assunto).

3 Método da Bisseção

asdajhduahd

4 Summary

In summary, this book has no content whatsoever.

References

Knuth, Donald E. 1984. «Literate Programming». *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.