Exercício: regressão múltipla

Disciplina: Modelagem Estatística
Instrutor: Luiz Max Carvalho
Monitor: Isaque Pim

Março/2023

Notação: Como convenção adotamos $\mathbb{R} = (-\infty, \infty)$, $\mathbb{R}_+ = (0, \infty)$ e $\mathbb{N} = \{1, 2, \ldots\}$.

Motivação: O modelo de regressão é extremamente útil como ferramenta explanatória e preditiva, mas até agora nos limitamos à situação em que temos apenas uma variável independente. Em aplicações reais, em geral dispomos de muitas covariáveis que podem, em princípio, estar relacionadas à variável dependente (desfecho). Nos exercícios que se seguem, vamos estabelecer alguns resultados técnicos importantes para a execução e entendimento do modelo linear com múltiplas covariáveis.

Tome $\boldsymbol{X}$ uma matriz real $n \times P$ e $\boldsymbol{Y} = \{Y_1, \dots, Y_n\}^T \in \mathbb{R}^n$ um vetor contendo os valores da variável dependente.

Nosso modelo é

$$E[Y_i] =: \mu_i(\boldsymbol{\beta}) = \tilde{\boldsymbol{X}}_i^T \boldsymbol{\beta},$$

onde $\boldsymbol{\beta} \in \mathbb{R}^{P+1}$ é o vetor de coeficientes e parâmetro de interesse e $\tilde{\boldsymbol{X}}$ é uma matriz obtida adicionando uma coluna de uns, $\boldsymbol{X_0} = \{1, \dots, 1\}^T$, a $\boldsymbol{X}$. Para completar a especificação do modelo, vamos assumir que os erros em torno do preditor linear são normalmente distribuídos com variância comum:

$$\begin{aligned} Y_i &= \mu_i(\boldsymbol{\beta}) + \epsilon_i \\ \epsilon_i &\overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma^2), \end{aligned}$$

com $\sigma^2 \in \mathbb{R}_+$ desconhecida.

Questões

1. Escreva a log-verossimilhança e deduza seu gradiente e a sua derivada segunda (hessiana);

2. Com base nos cálculos do item anterior, mostre a forma do estimador de máxima verossimilhança para $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$;

3. Mostre que $\hat{\boldsymbol{\beta}}$ é não-viesado;

4. Considere um outro estimador <u>não-viesado</u> de $\boldsymbol{\beta}$: $\tilde{\boldsymbol{\beta}} = \boldsymbol{M}\boldsymbol{y}$, onde

$$\boldsymbol{M} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{D},$$

   e $\boldsymbol{D}$ é uma matriz $P \times n$ cujas entradas são não-zero.

   Mostre que $\boldsymbol{R} := \text{Var}(\tilde{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}})$ é positiva-definida.

   Dica: Compute $E[\tilde{\boldsymbol{\beta}}]$ e considere o que deve valer para $\boldsymbol{D}$ sob a premissa de que $\tilde{\boldsymbol{\beta}}$ é não-viesado.

   Comentário: Ao resolver o último item, você terá mostrado que o estimador de máxima verossimilhança (e também o estimador de mínimos quadrados) é o melhor estimador linear não-viesado (*best linear unbiased linear estimator, BLUE*). Em particular esta é a versão de Gauss[1] do famoso teorema de Gauss-Markov.

[1] Carl Friedrich Gauss (1777-1855), matemático alemão conhecido como o Príncipe dos Matemáticos.