Inferência para CS
Tópico 11 - Eficiência e Otimalidade do MLE Multivariado

Renato Martins Assunção

DCC - UFMG

MLE multivariado

- Vamos considerar o caso em que o parâmetro $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ é um vetor com $k > 1$.
- O MLE $\widehat{\boldsymbol{\theta}} = (\widehat{\theta_1}, \ldots, \widehat{\theta_k})$ é também um vetor de dimensão k.
- O elemento $\widehat{\theta_i}$ é um estimador de θ_i (a entrada i do vetor $\boldsymbol{\theta}$.
- O MLE $\widehat{\boldsymbol{\theta}}$ é obtido resolvendo-se o sistema de equações não-lineares que resulta de igualar o gradiente $\frac{\partial \ell}{\partial \boldsymbol{\theta}}$ da log-verossimilhança $\ell(\boldsymbol{\theta})$ a zero:

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \partial \ell / \partial \theta_1 \\ \partial \ell / \partial \theta_2 \\ \vdots \\ \partial \ell / \partial \theta_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

- Queremos estudar o comportamento estatístico do MLE.
- Veremos que ele é um estimador ótimo de $\boldsymbol{\theta}$.

MLE multivariado

- O MLE $\widehat{\boldsymbol{\theta}} = (\widehat{\theta_1}, \ldots, \widehat{\theta_k})$ é um vetor ALEATÓRIO.
- Ele depende dos dados estocásticos e por isto varia de amostra para amostra.
- Precisamos estudar este comportamento aleatório do MLE.
- Como veremos, o comportamento é similar ao do caso unidimensional.
- Se n não é muito pequeno, de forma aproximada, teremos o seguinte:
  - O MLE $\widehat{\boldsymbol{\theta}}$ é aproximadamente não viciado para estimar o vetor $\boldsymbol{\theta}$. Isto é, $\mathbb{E}(\widehat{\boldsymbol{\theta}}) \approx \boldsymbol{\theta}$.
  - Aproximadamente, a matriz de variância e covariância de $\widehat{\boldsymbol{\theta}}$ é, num certo sentido matemático, a "menor matriz" possível para um estimador não-viciado, a matriz de informação de Fisher $I(\boldsymbol{\theta})$.
  - O MLE $\widehat{\boldsymbol{\theta}}$ possui distribuição gaussiana multivariada aproximadamente.

Matriz de informação de Fisher

- A matriz de variância e covariância de $\widehat{\boldsymbol{\theta}}$ está associada à matriz de informação de Fisher $I(\boldsymbol{\theta})$.
- Definição: A matriz de informação de Fisher $I(\boldsymbol{\theta})$ é uma matriz $k \times k$ cujo elemento ij é dado por

$$[I(\boldsymbol{\theta})]_{ij} = \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_i} \; \frac{\partial \ell}{\partial \theta_j}\right)$$

  onde $\ell = \ell(\boldsymbol{\theta})$ é a log-verossimilhança de $\boldsymbol{\theta}$.
- Por exemplo, no caso em que $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ teremos

$$I(\boldsymbol{\theta}) = \begin{bmatrix} \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_1}\right)^2 & \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_1} \; \frac{\partial \ell}{\partial \theta_2}\right) & \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_1} \; \frac{\partial \ell}{\partial \theta_3}\right) \\ \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_2} \; \frac{\partial \ell}{\partial \theta_1}\right) & \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_2}\right)^2 & \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_2} \; \frac{\partial \ell}{\partial \theta_3}\right) \\ \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_3} \; \frac{\partial \ell}{\partial \theta_1}\right) & \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_3} \; \frac{\partial \ell}{\partial \theta_2}\right) & \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_3}\right)^2 \end{bmatrix}$$

Exemplo de $I(\boldsymbol{\theta})$

- Suponha que $Y_1, Y_2, \ldots, Y_n$ sejam i.i.d. $N(\mu, \sigma^2)$.
- Temos $\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R}^2$
- A log-verossimilhança de $\boldsymbol{\theta}$ baseada numa amostra é dada por

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - -\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2$$

- O vetor gradiente $k \times 1$ da log-verossimilhança $\ell(\boldsymbol{\theta})$ é dado por

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \partial\ell/\partial\mu \\ \partial\ell/\partial\sigma^2 \end{bmatrix} = \begin{bmatrix} \sum_i (y_i - \mu)/\sigma^2 \\ -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_i (y_i - \mu)^2 \end{bmatrix}$$

- A partir deste vetor, podemos obter a matriz de informação de Fisher tomando os seus produtos cruzados.

Exemplo de $I(\boldsymbol{\theta})$

- Temos

$$
\begin{aligned}
I(\boldsymbol{\theta}) &= \mathbb{E}\left[\begin{array}{cc} \left(\frac{\partial \ell}{\partial \mu}\right)^2 & \left(\frac{\partial \ell}{\partial \mu}\ \frac{\partial \ell}{\partial \sigma^2}\right) \\ \left(\frac{\partial \ell}{\partial \sigma^2}\ \frac{\partial \ell}{\partial \mu}\right) & \left(\frac{\partial \ell}{\partial \sigma^2}\right)^2 \end{array}\right] \\
&= \mathbb{E}\left[\begin{array}{cc} \left(\sum_i \frac{Y_i-\mu}{\sigma^2}\right)^2 & \left(\sum_i \frac{Y_i-\mu}{\sigma^2}\right)\left(-\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{\sum_i (Y_i-\mu)^2}{(\sigma^2)^2}\right) \\ \left(-\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{\sum_i (Y_i-\mu)^2}{(\sigma^2)^2}\right)\left(\sum_i \frac{Y_i-\mu}{\sigma^2}\right) & \left(-\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{\sum_i (Y_i-\mu)^2}{(\sigma^2)^2}\right)^2 \end{array}\right]
\end{aligned}
$$

- Um cálculo de probabilidade laborioso permite obter cada uma das esperanças presentes na matriz $I(\boldsymbol{\theta})$ resultando no seguinte:

$$
I(\boldsymbol{\theta}) = \left[\begin{array}{cc} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{array}\right]
$$

Uma fórmula alternativa para $I(\boldsymbol{\theta})$

- De forma similar ao caso unidimensional, podemos calcular a informação de Fisher de forma alternativa, usando a matriz Hessiana (a matriz de derivadas parciais de segunda ordem de $\ell(\boldsymbol{\theta})$):

$$[I(\boldsymbol{\theta})]_{ij} = -\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}\right)$$

- Por exemplo, no caso em que $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$, por esta fórmula alternativa, temos

$$I(\boldsymbol{\theta}) = -\begin{bmatrix} \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_1^2}\right) & \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\right) & \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_3}\right) \\ \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1}\right) & \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_2^2}\right) & \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_3}\right) \\ \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_3 \partial \theta_1}\right) & \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_3 \partial \theta_2}\right) & \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_3^2}\right) \end{bmatrix}$$

- Note o sinal negativo na frente da matriz. Além disso, podemos passar o $\mathbb{E}$ para fora da matriz: esperança de matriz é a esperança de cada entrada da matriz.

Matriz de informação de Fisher

- Considerando o exemplo anterior de v.a.'s i.i.d. com distribuição $N(\mu, \sigma^2)$, a partir do vetor gradiente obtemos a matriz Hessiana abaixo:

$$\begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{\sum_i (y_i - \mu)}{(\sigma^2)^2} \\ -\frac{\sum_i (y_i - \mu)}{(\sigma^2)^2} & \frac{n}{2}\frac{1}{(\sigma^2)^2} - \frac{\sum_i (y_i - \mu)^2}{(\sigma^2)^3} \end{bmatrix}$$

- Para obter $I(\boldsymbol{\theta})$, substituímos as instâncias $\boldsymbol{y}$ pelas v.a.'s $\boldsymbol{Y}$ e tomamos o negativo da esperança da matriz hessiana.
- Resulta que, neste exemplo, tomar a esperança das expressões da matriz Hessiana é muito simples. $I(\boldsymbol{\theta})$.
- O resultado é o mesmo de antes:

$$I(\boldsymbol{\theta}) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

Forma vetorial de $I(\boldsymbol{\theta})$

- A matriz de informação $I(\boldsymbol{\theta})$ pode ser definida termo a termo (como fizemos) ou de forma vetorial.
- A primeira fórmula para $I(\boldsymbol{\theta})$, que especifica o elemento ij da matriz como $[I(\boldsymbol{\theta})]_{ij} = \mathbb{E}\left(\partial\ell/\partial\theta_i \cdot \partial\ell/\partial\theta_j\right)$ pode ser escrita de forma vetorial como a esperança da matriz que resulta ao multiplicar duas vezes o vetor gradiente de $\ell(\boldsymbol{\theta})$ de forma transposta:

$$
\begin{aligned}
I(\boldsymbol{\theta}) &= \mathbb{E}\left(\left[\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right] \cdot \left[\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right]'\right) \\
&= \mathbb{E}\left(\begin{bmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \frac{\partial \ell}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ell}{\partial \theta_k} \end{bmatrix} \cdot \left[\frac{\partial \ell}{\partial \theta_1}, \frac{\partial \ell}{\partial \theta_2}, \ldots, \frac{\partial \ell}{\partial \theta_k}\right]\right)
\end{aligned}
$$

Exemplo de $I(\boldsymbol{\theta})$

- Considerando o exemplo de n v.a.'s i.i.d. com distribuição $N(\mu, \sigma^2)$ e $\boldsymbol{\theta} = (\mu, \sigma^2)$ temos

$$
\begin{aligned}
I(\boldsymbol{\theta}) &= \mathbb{E}\left[\begin{array}{c} \partial\ell/\partial\mu \\ \partial\ell/\partial\sigma^2 \end{array} \right] \cdot \left[\partial\ell/\partial\mu, \partial\ell/\partial\sigma^2\right] \\
&= \mathbb{E}\left[\begin{array}{c} \sum_i (y_i - \mu)/\sigma^2 \\ -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_i (y_i - \mu)^2 \end{array} \right] \cdot \left[\sum_i (y_i - \mu)/\sigma^2, -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_i (y_i - \mu)^2 \right]
\end{aligned}
$$

Outra forma vetorial de $I(\boldsymbol{\theta})$

- Além da forma vetorial dada por

$$I(\boldsymbol{\theta}) = \mathbb{E}\left(\left[\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right] \cdot \left[\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right]'\right)$$

  podemos obter $I(\boldsymbol{\theta})$ usando uma forma vetorial para expressar a matriz hessiana de derivadas segundas.

- Especificamente, temos

$$I(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \frac{\partial \ell}{\partial \boldsymbol{\theta}}\right]$$

Desigualdade de Cramer-Rao multivariada

- Quando T é um estimador não viciado de $\theta \in \mathbb{R}$ vimos que $\text{MSE}(T) = \mathbb{V}(T) \geq 1/I(\theta)$.
- Temos uma versão multivariada deste resultado para $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$.
- Seja $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{Y}) = (T_1, \ldots, T_k) \in \mathbb{R}^k$ um estimador k-dimensional não-viciado de $\boldsymbol{\theta}$ (isto é, $\mathbb{E}(\boldsymbol{T}) = \boldsymbol{\theta}$).
- Então, para todo i, temos

$$\text{MSE}(T_i) = \mathbb{V}(T_i) \geq \left[I^{-1}(\boldsymbol{\theta})\right]_{ii}$$

que é o elemento ii da INVERSA da matriz de informação de Fisher.

Desigualdade de Cramer-Rao multivariada (OPCIONAL)

- O resultado multivariado que apresentamos é apenas parcial.
- O resultado completo é um pouco mais sutil e útil em alguns problemas estatísticos de estimação. e testes de hipóteses em experimentos planejados.
- Considere qualquer combinação linear dos k elementos do estimador não viciado $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{Y})$:

$$W = c_1 T_1 + c_2 T_2 + \ldots + c_k T_k = (c_1, \ldots, c_k)' \cdot (T_1, \ldots, T_k) = \boldsymbol{c}' \boldsymbol{T}$$

onde $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$ são constantes conhecidas.

- É fácil mostrar que W é estimador não-viciado para $c_1\theta_1 + c_2\theta_2 + \ldots + c_k\theta_k = \boldsymbol{c}'\,\boldsymbol{\theta}$.
- Então temos

$$\text{MSE}(W) = \mathbb{V}(W) = \mathbb{V}(\boldsymbol{c}'\,\boldsymbol{T}) \geq \boldsymbol{c}'\; I(\boldsymbol{\theta})\;\; \boldsymbol{c}$$

MLE: Caso Multivariado

- Suponha que o parâmetro $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ é um vetor com k componentes
- O MLE de $\boldsymbol{\theta}$ é um vetor aleatório $\widehat{\boldsymbol{\theta}}$ de dimensão k.
- O resultado anterior sobre o MLE é válido em multi-dimensões:

$$\widehat{\boldsymbol{\theta}} \approx N_k \left(\boldsymbol{\theta}, I^{-1}(\boldsymbol{\theta})\right)$$

- Se n não é pequeno, temos aproximadamente:
  - O MLE possui distribuição de probabilidade normal multivariada;
  - é aproximadamente não-viciado (isto é, $\mathbb{E}(\widehat{\boldsymbol{\theta}}) \approx \boldsymbol{\theta}$);
  - A matriz de covariância do MLE é aproximadamente igual ao inverso de $I(\boldsymbol{\theta})$, que é um limite ótimo para estimadores não-viciados.

MLE em regressão linear

- No modelo de regressão linear múltipla, temos v.a.'s $Y_1, \ldots, Y_n$ que são independentes mas não são i.d.: $Y_i \sim N(\boldsymbol{x}_i' \ \boldsymbol{\beta}, \sigma^2)$.
- A versão matricial do modelo de regressão linear é

$$\boldsymbol{Y} = \boldsymbol{X} \ \boldsymbol{\beta} + \varepsilon$$

- onde $\boldsymbol{Y}$ é vetor $n \times 1$, a matriz de desenho $\boldsymbol{X}$ é de dimensão $n \times p$ com $p-1$ variáveis independentes e a coluna de 1's.
- O vetor $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ é composto de v.a.'s i.i.d. $N(0, \sigma^2)$.
- O parâmetro $\boldsymbol{\theta}$ é $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2) = (\beta_0, \ldots, \beta_{p-1}, \sigma^2)$
- Queremos estimar $\boldsymbol{\beta}$ e σ^2.

MLE de $\boldsymbol{\beta}$ coincide com mínimos quadrados

- O MLE $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ coincide com o estimador de mínimos quadrados.
- Para ver isto, considere a log-verossimilhança $\ell(\boldsymbol{\theta})$.
- As v.a.'s Y_i são independentes com distribuição $N(\boldsymbol{x}_i' \ \boldsymbol{\beta}, \sigma^2)$.
- Então

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i' \ \boldsymbol{\beta})^2\right)\right) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (y_i - \boldsymbol{x}_i' \ \boldsymbol{\beta})^2 \end{aligned}$$

- É fácil perceber que, para qualquer valor fixo de σ^2, o vetor $\boldsymbol{\beta}$ que maximiza $\ell(\boldsymbol{\theta})$ será aquele que minimiza a soma de quadrados $\sum_i (y_i - \boldsymbol{x}_i' \ \boldsymbol{\beta})^2$.
- Assim, o MLE de $\boldsymbol{\beta}$ é o vetor que minimiza a distância entre o vetor $\boldsymbol{Y}$ e a combinação linear $\boldsymbol{X}\,\boldsymbol{\beta}$:

MLE de $\boldsymbol{\beta}$ coincide com mínimos quadrados

- Já apredemos que a solucão deste último problema é a combinação que produz a projeção ortogonal de $\boldsymbol{Y}$ no espaço vetorial das combinações lineares das colunas de $\boldsymbol{X}$:

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta_0}, \ldots, \widehat{\beta_p})' = (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{Y}$$

- $\widehat{\boldsymbol{\beta}}$ é uma função do vetor de dados aleatórios $\boldsymbol{Y}$ e da matriz de regressores $\boldsymbol{X}$ (que é considerada uma matriz de constantes conhecida).
- Se escrevermos a matriz $k \times n$ dada por $(\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t$ por A podemos ver que $\widehat{\boldsymbol{\beta}} = A\, \boldsymbol{Y}$.
- Assim, cada elemento de $\widehat{\boldsymbol{\beta}}$ é uma combinação linear dos elementos do vetor $\boldsymbol{Y}$.

Resíduos

- O modelo prediz ou estima o valor de Y_i usando $\widehat{\boldsymbol{\beta}}$.
- O vetor $n \times 1$ com os valores preditos pelo modelo para os valores realmente observados $\boldsymbol{Y}$ são dados por

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\ \boldsymbol{\beta} = \boldsymbol{X}\ (\boldsymbol{X}^t\,\boldsymbol{X})^{-1}\boldsymbol{X}^t\,\boldsymbol{Y}$$

- A diferença entre $\boldsymbol{Y}$ e a predição $\widehat{\boldsymbol{Y}}$ forma o vetor de resíduos ou vetor de erros de predição $\boldsymbol{r} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$.
- O i-ésimo elemento do vetor de resíduos é dado por

$$r_i = Y_i - \boldsymbol{x}_i'\,\widehat{\boldsymbol{\beta}}$$

- Gráficos e análises com os resíduos são uma excelente maneira de identificar falhas nas suposições do modelo de regressão linear

Resíduos

- Seja *RSS* a soma de quadrados dos resíduos (residual sum of squares)

$$RSS = || \boldsymbol{Y} - \widehat{\boldsymbol{Y}}||^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}})^2$$

- O MLE de σ^2 é dado pela média dos resíduos ao quadrado:

$$\widehat{\sigma^2} = \frac{RSS}{n}$$

- Para verificar isto, veja que tomando $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ teremos a log-verossimilhança $\ell(\boldsymbol{\theta})$ igual a

$$\ell(\widehat{\boldsymbol{\beta}}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}RSS$$

que é maximizada se tomarmos $\sigma^2 = RSS/n$ (basta derivar e igualar a zero).

$\widehat{\boldsymbol{\beta}}$ é não-viciado

- Revisão de probabilidade: A é matriz de constantes e $\boldsymbol{Y}$ é vetor aleatório de dimensões compatíveis. Então o vetor aleatório $A\,\boldsymbol{Y}$ tem um vetor esperado igual a $\mathbb{E}(A\,\boldsymbol{Y}) = A\mathbb{E}(\boldsymbol{Y})$ e a matriz de covariância é $\mathbb{V}(A\,\boldsymbol{Y}) = A\mathbb{V}(\boldsymbol{Y})A^t$.
- O MLE $\widehat{\boldsymbol{\beta}}$ é não-viciado para $\boldsymbol{\beta}$ pois

$$\begin{aligned}
\mathbb{E}(\widehat{\boldsymbol{\beta}}) &= \mathbb{E}((\boldsymbol{X}^t\,\boldsymbol{X})^{-1}\boldsymbol{X}^t\,\boldsymbol{Y}) \\
&= (\boldsymbol{X}^t\,\boldsymbol{X})^{-1}\boldsymbol{X}^t\,\mathbb{E}(\boldsymbol{Y}) \\
&= (\boldsymbol{X}^t\,\boldsymbol{X})^{-1}\boldsymbol{X}^t\mathbb{E}(\boldsymbol{X}\;\boldsymbol{\beta}+\varepsilon) \\
&= (\boldsymbol{X}^t\,\boldsymbol{X})^{-1}\boldsymbol{X}^t(\boldsymbol{X}\;\boldsymbol{\beta}+\mathbb{E}(\varepsilon)) \\
&= \boldsymbol{\beta}+\mathbf{0} = \boldsymbol{\beta}
\end{aligned}$$

Covariância de $\widehat{\beta}$

- É possível calcular a matriz de covariância de $\widehat{\beta}$.
- Chame $A = (\boldsymbol{X}^t \boldsymbol{X})^{-1}\boldsymbol{X}^t$ e então $\widehat{\beta} = A\,\boldsymbol{Y}$.
- Como as v.a.'s Y_i são independentes, sua matriz de covariância é $\sigma^2 I_n$, um múltiplo da matriz identidade de ordem n.
- Então, usando o resultado de probab, temos

$$\mathbb{V}(\widehat{\beta}) = A(\sigma^2 I_n)A^t = \sigma^2(\boldsymbol{X}^t \boldsymbol{X})^{-1}\boldsymbol{X}^t \;\; \boldsymbol{X}(\boldsymbol{X}^t \boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^t \boldsymbol{X})^{-1}$$

Propriedades de $\widehat{\sigma^2} = \frac{RSS}{n}$

- Quanto ao MLE de σ^2 no modelo de regressão linear, podemos mostrar que

$$\mathbb{E}(\widehat{\sigma^2}) = \mathbb{E}(\frac{RSS}{n}) = \frac{n-k-1}{n}\sigma^2$$

- Portanto o MLE $\widehat{\sigma^2}$ é viciado para estimar σ^2 mas seu vício desaparece quando n cresce.
- É comum usarmos um estimador não-viciado para σ^2 dados por

$$S^2 = \frac{RSS}{n-k-1}$$

- O uso deste estimador não-viciado (que não é o MLE) permite fazer vários cálculos exatos de probabiliadde envolvidos em intervalos de confiança e testes de hipótese (mais a frente no curso).
- Para o MLE $\widehat{\sigma^2} = RSS/n$ temos a variância $2\sigma^4(n-k-1)/n^2$.
- Além disso, a covariância entre $\widehat{\boldsymbol{\beta}}$ e $\widehat{\sigma^2}$ é o vetor $\mathbf{0}$.

$I(\boldsymbol{\theta})$

- Podemos voltar a log-verossmilhança $\ell(\boldsymbol{\theta})$ e calcular a matriz de informação de Fisher usando vários truques de probabilidade e álgebra linear.
- No final encontramos

$$I(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2}\boldsymbol{X}^t\boldsymbol{X} & \mathbf{0} \\ \mathbf{0}^t & \frac{n}{2\sigma^4} \end{pmatrix}$$

- A matriz inversa é igual a

$$I^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1} & \mathbf{0} \\ \mathbf{0}^t & \frac{2\sigma^4}{n} \end{pmatrix}$$

$I(\boldsymbol{\theta})$

- A cota-inferior de Cramér-Rao para um estimador nã-viciado de $\boldsymbol{\theta}$ é

$$I^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1} & \mathbf{0} \\ \mathbf{0}^t & \frac{2\sigma^4}{n} \end{pmatrix}$$

- Compare $I^{-1}(\boldsymbol{\theta})$ com a matriz de covariância exata do MLE

$$\mathbb{V}(\widehat{\boldsymbol{\theta}}) = \mathbb{V}(\widehat{\boldsymbol{\beta}}, \widehat{\sigma^2}) = \begin{pmatrix} \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1} & \mathbf{0} \\ \mathbf{0}^t & \frac{2\sigma^4(n-k-1)}{n^2} \end{pmatrix} \to I^{-1}(\boldsymbol{\theta})$$

- O MLE do parâmetro $\boldsymbol{\beta}$ é não-viciado para $\boldsymbol{\beta}$ e atinge a cota inferior de Cramér-Rao $\sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}$
- O MLE de σ^2 tem um vício que vai a zero com n e tem uma variância que aproxima-se rapidamente da cota inferior de Cramér-Rao.

MLE de regressão logística

- No caso da regressão logística, o MLE $\widehat{\theta}$ vai satisfazer a equação de verossimilhança não-linear

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{X}^t \boldsymbol{y} - \boldsymbol{X}^t \boldsymbol{p} = \mathbf{0}$$

  onde $\boldsymbol{p} = (p_1, \ldots p_n)$ e $p_i = 1/(1 + \exp(-\boldsymbol{x}_i^t \boldsymbol{\theta}))$.
- O MLE é o resultado da da convergência do algoritmo de Newton-Raphson:

$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta}^{\text{old}} - \left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}^t} \right)^{-1} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

  onde as derivadas são avaliadas usando-se o valor corrente $\boldsymbol{\theta}^{\text{old}}$.
- Como valor inicial, podemos usar $\boldsymbol{\theta}^{(0)} = \mathbf{0}$.

MLE de regressão logística

- Temos

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{X}^t \boldsymbol{y} - \boldsymbol{X}^t \boldsymbol{p}$$

- e

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} = -\boldsymbol{X}^t \boldsymbol{W} \boldsymbol{X}$$

onde

$$\boldsymbol{W} = \begin{bmatrix} p_1(1-p_1) & 0 & 0 & \dots & 0 \\ 0 & p_2(1-p_2) & 0 & \dots & 0 \\ 0 & 0 & p_3(1-p_3) & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & p_n(1-p_n) \end{bmatrix}$$

- Temos então

$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta}^{\text{old}} + \left(\boldsymbol{X}^t \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^t (\boldsymbol{y} - \boldsymbol{p})$$

$I(\boldsymbol{\theta})$

- Para obter a matriz de informação de Fisher, calculamos a matriz Hessiana (a matriz de derivadas parciais de segunda ordem da log-verossimilhança $\ell(\boldsymbol{\theta})$):

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}^t}$$

  e depois calculamos $I(\boldsymbol{\theta})$ como o negativo do valor esperado de cada elemento desta matriz hessiana.
- A matriz Hessiana é igual a

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}^t} = -\boldsymbol{X}^t \, \boldsymbol{W} \, \boldsymbol{X}$$

- Que sorte!! Esta é uma matriz não-aleatória, somente com constantes.
- Assim, sua esperança é igual a ela mesma!
- Portanto, a matriz de informação de Fisher é

$$I(\boldsymbol{\theta}) = \boldsymbol{X}^t \, \boldsymbol{W} \, \boldsymbol{X}$$

Exemplo - EMV de logística

- Conclusão:

$$\widehat{\boldsymbol{\theta}} \sim N_{k+1}\left(\boldsymbol{\theta}, (\boldsymbol{X}^t\, \boldsymbol{W}\, \boldsymbol{X})^{-1}\right)$$

com

$$\boldsymbol{W} = \text{diag}\left(p_1(1-p_1), p_2(1-p_2), \ldots, p_n(1-p_n)\right)$$

- Cada p_i é função das covariáveis e de $\boldsymbol{\theta}$:

$$p_i = 1/(1+\exp(-\boldsymbol{x}_i^t\, \boldsymbol{\theta}))$$

- Isto é, o EMV de $\boldsymbol{\theta}$ tem uma distribuição aproximadamente normal multivariada de dimensão $k+1$ centrada no verdadeiro valor do parâmetro $\boldsymbol{\theta}$ e com matriz de covariância $I^{-1}(\boldsymbol{\theta}) = \left(\boldsymbol{X}^t\, \boldsymbol{W}\, \boldsymbol{X}\right)^{-1}$
- A matrix $\boldsymbol{W}$ depende do valor desconhecido do parâmetro $\boldsymbol{\theta}$.
- Nós podemos usar nosso (melhor) estimador disponível $\widehat{\boldsymbol{\theta}}$ em $\boldsymbol{W}$ para obter uma aproximação para $I(\boldsymbol{\theta})$.

Intervalos de Confiança com o MLE

- O MLE $\widehat{\boldsymbol{\theta}} = (\widehat{\theta_1}, \ldots, \widehat{\theta_k})$ é um vetor ALEATÓRIO.
- A sua distribuição conjunta é dada por

$$\widehat{\boldsymbol{\theta}} \sim N_k\left(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})\right)$$

- Considerando apenas a i-ésima coordenada, temos

$$\hat{\theta}_i \approx N_1(\theta_i, \left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii})$$

onde $\left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii}$ é o elemento (i, i) na diagonal da INVERSA da matriz de informação de Fisher.

- Podemos fornecer intervalos de confiança para cada θ_i.

Intervalos de Confiança com o MLE

- Considerando apenas a i-ésima coordenada, temos

$$\hat{\theta}_i \approx N_1(\theta_i, \left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii})$$

- Numa gaussiana qualquer, a probabilidade da variável afastar-se de sua esperança por menos que dois desvios-padrão é aproximadamente 0.95.
- Portanto, aproximadamente,

$$\mathbb{P}\left(|\hat{\theta}_i - \theta_i| \leq 2\sqrt{\left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii}}\right) \approx 0.95$$

- Desigualdade básica: $|a - b| \leq 2$ se, e somente se, $a - 2 \leq b \leq a + 2$
- Assim,

$$\mathbb{P}\left(\hat{\theta}_i - 2\sqrt{\left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii}} \leq \theta_i \leq \hat{\theta}_i + 2\sqrt{\left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii}}\right) \approx 0.95$$

Intervalos de Confiança com o MLE

- O intervalo

$$\left[\hat{\theta}_i - 2\sqrt{[\mathbf{I}^{-1}(\boldsymbol{\theta})]_{ii}}, \leq \hat{\theta}_i + 2\sqrt{[\mathbf{I}^{-1}(\boldsymbol{\theta})]_{ii}}\right]$$

  é chamado intervalo de confiança de 95% para o parâmetro θ_i.
- A confiança é no método. 95% dos intervalos que você fizer usando o MLE vão cobrir o verdadeiro valor do parâmetro.
- Em geral, $\mathbf{I}^{-1}(\boldsymbol{\theta})$ depende do parâmetro θ, que é desconhecido.
- Nestes casos, substitua θ por $\hat{\theta}$.