INVERSÃO LINEAR:

ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA (MAXIMUM LIKELIHOOD)

Os estimadores de Máxima Verossimilhança (EMV) usam informação a priori sobre o comportamento estatístico da variável aleatória (v.a.) que presumivelmente contamina as observações geofísicas (dados). Tais estimadores não usam nem explicita nem implicitamente informação a priori sobre os parâmetros como os estimadores que anteriormente estudamos

Conceitos Fundamentais:

Considere uma v.a. ε sobre a qual conhecemos a função de densidade de probabilidade (f.d.p.) $f(\varepsilon, \mu)$ a menos da média μ. Vamos supor que conhecemos N realizações desta v.a., i.e., $\varepsilon_1, \ldots, \varepsilon_N$. Dado uma v.a. ε_i temos associado a ela uma f.d.p. $f(\varepsilon_i, \mu)$.

O método Máxima Verossimilhança (MV) consiste em estimar a média μ. Então, no método MV estima-se o valor $\hat{\mu}$ tal que os valores da f.d.p. $f(\varepsilon_i, \hat{\mu})$, i=1,...,N sejam todos grandes. Em outras palavras o método MV maximiza a seqüência particular das realizações $\varepsilon_1, \ldots, \varepsilon_N$.

A função de densidade de probabilidade conjunta das v.a $\varepsilon_1, \ldots, \varepsilon_N$ é dada por

$$L(\boldsymbol{\varepsilon}, \boldsymbol{\mu}) = f(\varepsilon_1, \mu) \times f(\varepsilon_2, \mu) \times \ldots \times f(\varepsilon_N, \mu) ,$$

ou seja, a função de densidade de probabilidade conjunta é o produtório das N f.d.p. $f(\varepsilon_i, \mu)$, i=1,...,N

$$L(\boldsymbol{\varepsilon}, \boldsymbol{\mu}) = \prod_{i=1}^{N} f(\varepsilon_i, \mu) \qquad (1)$$

O estimador $\hat{\mu}$ da MV da média de $\varepsilon_1, \ldots, \varepsilon_N$ será aquele que satisfizer a condição

$$\max_{\mu} \; L(\boldsymbol{\varepsilon}, \boldsymbol{\mu}) = \max \prod_{i=1}^{N} f(\varepsilon_i, \mu) \quad (2)$$

Na geofísica as nossas v.a. são os dados observados ($\bar{\mathbf{y}}^o$). Comumente fazemos em cada ponto de medida apenas uma observação, ou seja, em cada ponto temos apenas uma realização de uma v.a.

O estimador $\hat{\mu}$ da MV dos dados observados ($\bar{\mathbf{y}}^o$) será aquele que satisfizer a condição

$$\max_{\mu} \; L(\bar{\mathbf{y}}^o, \boldsymbol{\mu}) = \max \; f(\bar{\mathbf{y}}^o, \mu) \quad (3)$$

Mas quem é a média μ dos dados observados $\bar{\mathbf{y}}^o$?

Cálculo da $E\{\bar{\mathbf{y}}^o\}$:

Considerando as premissas estatísticas estabelecidas na Figura 1 e as propriedades da esperança, temos:

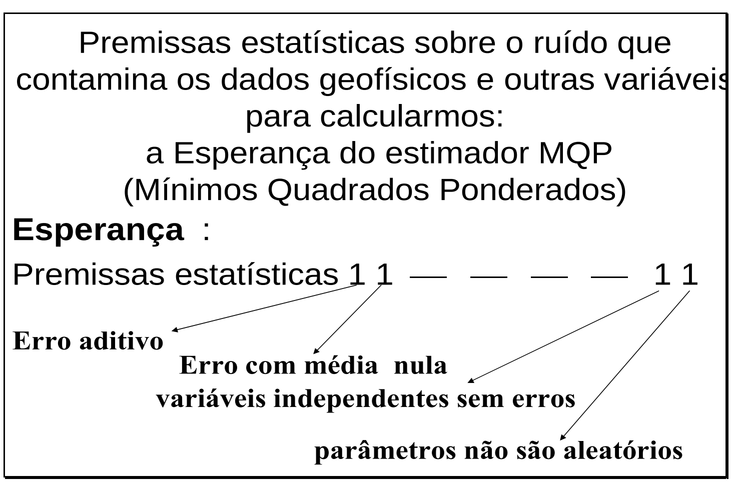

Figura 1

Usando a premissa 1 que os erros são aditivos ($\overline{\mathbf{y}}^{o} = \overline{\mathbf{y}}^{c} + \overline{\boldsymbol{\varepsilon}}$) e usando a informação que em um problema linear a componente determinística (vetor dos dados ajustados ou calculados) é $\overline{\mathbf{y}}^{C} = \overline{\overline{\mathbf{A}}}\,\overline{\mathbf{p}}$ temos que

$$\overline{\mathbf{y}}^{o} = \overline{\overline{\mathbf{A}}}\,\overline{\mathbf{p}} + \overline{\boldsymbol{\varepsilon}}$$

Calculando a esperança da v.a. que são os dados ($\overline{\mathbf{y}}^{\mathbf{o}}$) temos

$$E\{\overline{\mathbf{y}}^{o}\} = E\{\overline{\overline{\mathbf{A}}}\,\overline{\mathbf{p}}\} + E\{\overline{\boldsymbol{\varepsilon}}\}$$

Usando a premissa 7 que as variáveis independentes não são v.a. e a premissa 8 que os parâmetros não são aleatórios temos

$$E\{\overline{\mathbf{y}}^{o}\} = \overline{\overline{\mathbf{A}}}\,\overline{\mathbf{p}} + E\{\overline{\boldsymbol{\varepsilon}}\}$$

Pela premissa 2 os erros tem média nula o que implica $E[\overline{\boldsymbol{\varepsilon}}] = \overline{0}$ então

$$E\{\overline{\mathbf{y}}^{o}\} = \overline{\overline{\mathbf{A}}}\,\overline{\mathbf{p}}. \tag{4}$$

O ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA

Substituindo a equação (4) na equação (3) temos o estimador da MV dos dados observados ($\overline{\mathbf{y}}^{o}$)

$$\max_{\mu} \; L(\overline{\mathbf{y}}^{o}, E\{\overline{\mathbf{y}}^{o}\}) = \max \; f(\overline{\mathbf{y}}^{o}, E\{\overline{\mathbf{y}}^{o}\})$$

$$\max_{\mu} \; L(\overline{\mathbf{y}}^{o}, \overline{\overline{\mathbf{A}}}\,\overline{\mathbf{p}}) = \max \; f(\overline{\mathbf{y}}^{o}, \overline{\overline{\mathbf{A}}}\,\overline{\mathbf{p}}) \tag{5}$$

Caso particular: Distribuição Gaussiana

Considerando apenas uma observação (y_i^o) e considerando o caso particular da distribuição Gaussiana temos que a f.d.p para y_i^o é dada por

$$f(y_i^O, y_i^C) = \frac{1}{(2\pi)^{1/2}\sigma_i} e^{-(y_i^O - y_i^C)^2/2\sigma_i^2}$$

em que σ_i é o desvio padrão da i-ésima observação y_i^o

A função de máxima verossimilhança é dada pelo produtório das N f.d.p. $f(y_i^O, y_i^C)$, i=1,...,N

$$L(\bar{\mathbf{y}}^o, \bar{\bar{\mathbf{A}}}\bar{\mathbf{p}}) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{1/2}\sigma_i} e^{-(y_i^O - y_i^C)^2/2\sigma_i^2}$$

$$L(\bar{\mathbf{y}}^o, \bar{\bar{\mathbf{A}}}\bar{\mathbf{p}}) = \frac{1}{(2\pi)^{N/2}\prod_{i=1}^{N}\sigma_i} \prod_{i=1}^{N} e^{-(y_i^O - y_i^C)^2/2\sigma_i^2} \quad (6)$$

A função acima é a função de máxima verossimilhança que notação matricial pode ser reescrita como

$$L(\bar{\mathbf{y}}^o, \bar{\bar{\mathbf{A}}}\bar{\mathbf{p}}) = \frac{1}{(2\pi)^{N/2}\,\bar{\bar{\mathbf{W}}}^{1/2}} e^{-\frac{1}{2}\left\{\left(\bar{\mathbf{y}}^{\mathbf{o}} - \bar{\bar{\mathbf{A}}}\bar{\mathbf{p}}\right)^{\mathsf{T}} \bar{\bar{\mathbf{W}}}^{-1}\left(\bar{\mathbf{y}}^{\mathbf{o}} - \bar{\bar{\mathbf{A}}}\bar{\mathbf{p}}\right)\right\}} \quad (7)$$

em que $\bar{\bar{\mathbf{W}}}$ (N x N) é uma matriz diagonal cujo i-ésimo elemento da diagonal é a variância σ_i^2 do i-ésimo elemento do vetor dos resíduos ($\bar{\boldsymbol{\varepsilon}}$) então temos

$$\bar{\bar{\mathbf{W}}} = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2{}^2 & & \\ & & \ddots & \\ & & & \sigma^2{}_N \end{bmatrix}_{N\times N}$$

O estimador de máxima verossimilhança dos dados observados (equação 5) consiste em maximizar a função de máxima verossimilhança (equação 7) para estimarmos o vetor de parâmetros $\overline{\mathbf{p}}$, i.e.:

$$\max_{\overline{\mathbf{p}}} L(\overline{\mathbf{y}}^o, \overline{\overline{\mathbf{A}}}\overline{\mathbf{p}}) = \max \frac{1}{(2\pi)^{N/2} \overline{\overline{\mathbf{W}}}^{1/2}} e^{-\frac{1}{2}\left\{\left(\overline{\mathbf{y}}^{\mathbf{o}} - \overline{\overline{\mathbf{A}}}\overline{\mathbf{p}}\right)^T \overline{\overline{\mathbf{W}}}^{-1}\left(\overline{\mathbf{y}}^{\mathbf{o}} - \overline{\overline{\mathbf{A}}}\overline{\mathbf{p}}\right)\right\}} \quad (8)$$

Para maximizarmos a função acima basta minimizar o expoente da exponencial. Portanto a condição necessária para que $L(\overline{\mathbf{y}}^o, \overline{\overline{\mathbf{A}}}\overline{\mathbf{p}})$ tenha um máximo é

$$\overline{\nabla}_{\overline{\mathbf{p}}} \left\{\left(\overline{\mathbf{y}}^{\mathbf{o}} - \overline{\overline{\mathbf{A}}}\overline{\mathbf{p}}\right)^T \overline{\overline{\mathbf{W}}}^{-1}\left(\overline{\mathbf{y}}^{\mathbf{o}} - \overline{\overline{\mathbf{A}}}\overline{\mathbf{p}}\right)\right\} = \overline{\mathbf{0}} \quad (9)$$

$$-2\, \overline{\overline{\mathbf{A}}}^T \overline{\overline{\mathbf{W}}}^{-1}\left(\overline{\mathbf{y}}^{\mathbf{o}} - \overline{\overline{\mathbf{A}}}\hat{\overline{\mathbf{p}}}_L\right) = \overline{\mathbf{0}}$$

que resulta no Estimador de Máxima Verrossimilhança (EMV):

$$\hat{\overline{\mathbf{p}}}_L = \left(\overline{\overline{\mathbf{A}}}^T \overline{\overline{\mathbf{W}}}^{-1} \overline{\overline{\mathbf{A}}}\right)^{-1} \overline{\overline{\mathbf{A}}}^T \overline{\overline{\mathbf{W}}}^{-1} \overline{\mathbf{y}}^{\mathbf{o}} \quad \textbf{(EMV)} \quad (10)$$

Note que o estimador de Máxima Verrossimilhança é a soma dos resíduos ponderados pelo o inverso da variâncias dos erros. Portanto o EMV é igual ao estimador MQP (mínimos quadrados ponderados que apresentamos no tópico 8 deste curso).

$$\min_{\overline{\mathbf{p}} \in F} \left\{\overline{\boldsymbol{\varepsilon}}^T \overline{\overline{\mathbf{W}}}^{-1} \overline{\boldsymbol{\varepsilon}}\right\} = \sum_{i=1}^{N} \frac{\varepsilon_i^2}{\sigma_i^2}$$

Assim, valem as mesmas observações que fizemos para o MQP, como por exemplo, observações mais confiáveis (variância pequena) produzirão resíduo pequeno.

Equivalência entre estimadores que usam o critério de minimização de uma norma e o EMV

Veremos a Equivalência entre estimadores que usam como critério a minimização de uma norma dos resíduos e certos estimadores de Máxima Verossimilhança. Por simplicidade limitaremos ao caso de um único parâmetro.

1) EMV – Distribuição Gaussiana

Considerando apenas uma observação (y_i^o) e considerando o caso particular da distribuição Gaussiana temos que a f.d.p para y_i^o é dada por

$$f(y_i^O,\mu)=\frac{1}{(2\pi)^{1/2}\sigma}e^{-(y_i^O-\mu)^2/2\sigma^2}$$

em que σ é o desvio padrão da observação y_i^o

Considerando um único parâmetro a ser estimado (μ) e um único desvio padrão para todas as observações (σ) a função de máxima verossimilhança é dada pelo produtório das N f.d.p. $f(y_i^O,\mu)$, i=1,...,N

$$L(\overline{\mathbf{y}}^o,\overline{\overline{\mathbf{A}}}\,\overline{\mathbf{p}})=\prod_{i=1}^{N}\frac{1}{(2\pi)^{1/2}\sigma}e^{-(y_i^O-\mu)^2/2\sigma^2}$$

$$L(\overline{\mathbf{y}}^o,\overline{\overline{\mathbf{A}}}\,\overline{\mathbf{p}})=\frac{1}{(2\pi)^{N/2}\sigma^N}e^{-\frac{1}{2}\sum\limits_{i=1}^{N}\left(\frac{y_i^O-\mu}{\sigma}\right)^2}$$

Para maximizarmos a função acima basta minimizar o expoente da exponencial. Portanto o máximo ocorre quando minimiza-se: $\sum_{i=1}^{N}\left(\dfrac{y_i^O-\mu}{\sigma}\right)^2$.

Para minimizarmos esta expressão em relação ao parâmetro desconhecido μ fazemos

$$\frac{\partial}{\partial\mu}\sum_{i=1}^{N}\left(\frac{y_i^O-\mu}{\sigma}\right)^2=0$$

$$\frac{2}{\sigma^2}\sum_{i=1}^{N}\left(y_i^O-\hat{\mu}\right)(-1)=0\,.$$

e estimamos

$$\hat{\mu}=\frac{1}{N}\sum_{i=1}^{N}y_i^O$$

Assim concluímos que o EMV sob a hipótese que os erros tem distribuição Gaussiana:

1.1) é equivalente a minimização da norma L2 dos resíduos, i.e.,

$\left\|\overline{\boldsymbol{\varepsilon}}\right\|_2=\left(\sum_{i=1}^{N}\left|\varepsilon_i\right|^2\right)^{1/2}$ com $\varepsilon_i=y_i^O-\mu$. Note que a função com raiz quadrada é monotônica de modo que se $\hat{\mu}$ minimiza $\left\|\overline{\boldsymbol{\varepsilon}}\right\|_2$ minimiza também $\left\|\overline{\boldsymbol{\varepsilon}}\right\|_2^{1/2}$ e $\left\|\overline{\boldsymbol{\varepsilon}}\right\|_2^2$.

1.2) baseia-se na **média amostral**

2) **EMV – Distribuição de Laplace**

Considerando apenas uma observação (y_i^o) e considerando o caso particular da distribuição de Laplace temos que a f.d.p para y_i^o é dada por

$$f(y_i^O,\mu)=\frac{1}{2\sigma}e^{-\frac{1}{\sigma}\left|y_i^O-\mu\right|}$$

em que σ é o desvio padrão da observação y_i^o

Considerando um único parâmetro a ser estimado (μ) e um único desvio padrão para todas as observações (σ) a função de máxima verossimilhança é dada pelo produtório das N f.d.p. $f(y_i^O,\mu)$, i=1,...,N

$$L(\overline{\mathbf{y}}^o,\overline{\overline{\mathbf{A}}}\,\overline{\mathbf{p}})=\prod_{i=1}^{N}\frac{1}{2\sigma}e^{-\frac{1}{\sigma}\left|y_i^O-\mu\right|}$$

$$L(\overline{\mathbf{y}}^o,\overline{\overline{\mathbf{A}}}\,\overline{\mathbf{p}})=\frac{1}{2\sigma}e^{-\frac{1}{\sigma}\sum\limits_{i=1}^{N}\left|y_i^O-\mu\right|}$$

Para maximizarmos a função acima basta minimizar o expoente da exponencial. Portanto o máximo ocorre quando

minimiza-se: $\sum\limits_{i=1}^{N}\left|y_i^O-\mu\right|$.

Para minimizarmos esta expressão em relação ao parâmetro desconhecido μ fazemos

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{N} \left| y_i^O - \mu \right| = 0$$

$$\sum_{i=1}^{N} \mathrm{sgn}\left(y_i^O - \hat{\mu} \right) = 0 \,.$$

em que $\mathrm{sgn}(y_i^O - \hat{\mu})$ é o sinal do i-ésimo resíduo. Assim concluímos que o EMV sob a hipótese que os erros têm distribuição de Laplace :

2.1) é equivalente a minimização da norma L1 dos resíduos, i.e.,

$$\left\| \bar{\boldsymbol{\varepsilon}} \right\|_1 = \left(\sum_{i=1}^{N} \left| \varepsilon_i \right| \right) \text{ com } \varepsilon_i = y_i^O - \mu \,.$$

2.2) baseia-se na **mediana amostral**