Cuadrados mínimos: errores correlacionados y ajuste lineal

Tenemos n mediciones $\{x_i, y_i\}$, sin errores en x y errores gaussianos correlacionados en y. O sea, los y_i son variables aleatorias con distribución multinormal, $\mathcal{N}(\boldsymbol{y} \,|\, \boldsymbol{\mu}, \boldsymbol{V})$, con $\boldsymbol{\mu}$ la esperanza de $\boldsymbol{y}$, y $\boldsymbol{V}$ la matriz de covarianza de $n{\times}n$. La esperanza $\mu_i = \mathrm{E}(y_i) = f(x_i; \boldsymbol{\theta})$ depende de x_i y de los k parámetros $\boldsymbol{\theta}$ que queremos estimar.

Maximizar la verosimilitud $\mathcal{L}\,(\boldsymbol{\theta}|\boldsymbol{y})$ es lo mismo que minimizar la suma de cuadrados $S(\boldsymbol{\theta})$:

$$\mathcal{L}\,(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{V}|}}\ \exp\left\{ -\tfrac{1}{2}\big(\boldsymbol{y}-\boldsymbol{\mu}\big)^{\mathbf{T}}\,\boldsymbol{V}^{-1}\,\big(\boldsymbol{y}-\boldsymbol{\mu}\big)\right\}$$

$$\ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}) = -\tfrac{1}{2}\ln(2\pi)^n|\boldsymbol{V}| - \tfrac{1}{2}\,S(\boldsymbol{\theta})$$

$$S(\boldsymbol{\theta}) = \big(\boldsymbol{y}-\boldsymbol{\mu}\big)^{\mathbf{T}}\,\boldsymbol{V}^{-1}\,\big(\boldsymbol{y}-\boldsymbol{\mu}\big) \qquad\qquad \text{con} \quad \mu_i = \mathrm{E}(y_i) = f(x_i, \boldsymbol{\theta})$$

Un caso particularmente importante es cuando la función a fitear es lineal en los parámetros,

$$f(x;\boldsymbol{\theta}) = \sum_{j=1}^{k} \theta_j\, f_j(x) \qquad\qquad \text{siendo} \quad f_j(x) \quad \text{funciones arbitrarias}$$

Evaluando $f(x;\boldsymbol{\theta})$ en cada x_i,

$$\mu_i = \mathrm{E}(y_i) = f(x_i;\boldsymbol{\theta}) = \sum_{j=1}^{k}\theta_j\, f_j(x_i) = \sum_{j=1}^{k} A_{ij}\,\theta_j \qquad \text{donde} \qquad \boxed{A_{ij} = f_j(x_i)}$$

Para el caso que la dependencia en los parámetros es lineal, queda $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{y}) = \boldsymbol{A}\,\boldsymbol{\theta}$,

$$\begin{aligned}
S(\boldsymbol{\theta}) &= \big(\boldsymbol{y}-\boldsymbol{\mu}\big)^{\mathbf{T}}\,\boldsymbol{V}^{-1}\,\big(\boldsymbol{y}-\boldsymbol{\mu}\big) \\
&= \big(\boldsymbol{y}-\boldsymbol{A}\,\boldsymbol{\theta}\big)^{\mathbf{T}}\,\boldsymbol{V}^{-1}\,\big(\boldsymbol{y}-\boldsymbol{A}\,\boldsymbol{\theta}\big) \\
&= \big(\boldsymbol{y}^{\mathbf{T}}-\boldsymbol{\theta}^{\mathbf{T}}\boldsymbol{A}^{\mathbf{T}}\big)\,\boldsymbol{V}^{-1}\,\big(\boldsymbol{y}-\boldsymbol{A}\,\boldsymbol{\theta}\big) \\
&= \boldsymbol{y}^{\mathbf{T}}\,\boldsymbol{V}^{-1}\boldsymbol{y} - \boldsymbol{y}^{\mathbf{T}}\,\boldsymbol{V}^{-1}\boldsymbol{A}\,\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathbf{T}}\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{V}^{-1}\,\boldsymbol{y} + \boldsymbol{\theta}^{\mathbf{T}}\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{V}^{-1}\,\boldsymbol{A}\,\boldsymbol{\theta}
\end{aligned}$$

Para minimizar $S(\boldsymbol{\theta})$ tenemos que encontrar los $\hat{\boldsymbol{\theta}}$ que hacen $\boldsymbol{\nabla} S(\hat{\boldsymbol{\theta}}) = 0$:

$$\boldsymbol{y}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A}\, \boldsymbol{\theta} \quad \longrightarrow \quad y_i\, V^{-1}_{ij}\, A_{jk}\, \theta_k \quad \longrightarrow \quad \frac{\partial(\ )}{\partial \theta_m} = y_i\, V^{-1}_{ij}\, A_{jm} = A^{\mathrm{T}}_{mj}\, V^{-1}_{ji}\, y_i = \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1}\, \boldsymbol{y}\,\big|_m$$

$$\boldsymbol{\theta}^{\mathbf{T}} \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1}\, \boldsymbol{y} \quad \longrightarrow \quad \theta_i\, A^{\mathrm{T}}_{ij}\, V^{-1}_{jk}\, y_k \quad \longrightarrow \quad \frac{\partial(\ )}{\partial \theta_m} = A^{\mathrm{T}}_{mj}\, V^{-1}_{jk}\, y_k = \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1}\, \boldsymbol{y}\,\big|_m$$

$$\begin{aligned}\boldsymbol{\theta}^{\mathbf{T}} \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A}\, \boldsymbol{\theta} \quad \longrightarrow \quad \theta_i\, A^{\mathrm{T}}_{ij}\, V^{-1}_{jk}\, A_{kl}\, \theta_l \quad \longrightarrow \quad \frac{\partial(\ )}{\partial \theta_m} &= A^{\mathrm{T}}_{mj}\, V^{-1}_{jk}\, A_{kl}\, \theta_l + \theta_i\, A^{\mathrm{T}}_{ij}\, V^{-1}_{jk}\, A_{km} \\ &= A^{\mathrm{T}}_{mj}\, V^{-1}_{jk}\, A_{kl}\, \theta_l + \theta_i\, A_{ji}\, V^{-1}_{kj}\, A^{\mathrm{T}}_{mk} \\ &= A^{\mathrm{T}}_{mj}\, V^{-1}_{jk}\, A_{kl}\, \theta_l + A^{\mathrm{T}}_{mk}\, V^{-1}_{kj}\, A_{ji}\, \theta_i = 2\, \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A}\, \boldsymbol{\theta}\,\big|_m\end{aligned}$$

$$\frac{\partial S}{\partial \theta_m} = -2\, \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{y}\,\big|_m + 2\, \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A}\, \boldsymbol{\theta}\,\big|_m \qquad \Rightarrow \qquad \boldsymbol{\nabla} S(\boldsymbol{\theta}) = -2\, \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{y} + 2\, \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A}\, \boldsymbol{\theta}$$

$$\boldsymbol{\nabla} S(\hat{\boldsymbol{\theta}}) = 0 \qquad \Rightarrow \qquad -2\, \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{y} + 2\, \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A}\, \hat{\boldsymbol{\theta}} = 0 \qquad \Rightarrow \qquad \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{y} = \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A}\, \hat{\boldsymbol{\theta}}$$

$$\boxed{\hat{\boldsymbol{\theta}} \ = \ (\boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A})^{-1}\, \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1}\, \boldsymbol{y}}$$

La matriz de covarianza $k \times k$ de los estimadores, $\mathrm{Cov}(\hat{\theta}_i, \hat{\theta}_j)$, se obtiene a partir de la matriz de covarianza $n \times n$ de los datos, $\mathrm{Cov}(y_r, y_s)$. Es inmediato dado la relación lineal entre $\boldsymbol{\theta}$ e $\boldsymbol{y}$:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{C} \boldsymbol{y} \quad \text{con} \quad \boldsymbol{C} = (\boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A})^{-1}\, \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1}$$

$$\begin{aligned} V(\hat{\boldsymbol{\theta}})_{ij} = \mathrm{Cov}(\hat{\theta}_i, \hat{\theta}_j) &= \sum_{r,s} C_{ir}\, C_{js}\, \mathrm{Cov}(y_r, y_s) \\ &= \sum_{r,s} C_{ir}\, \mathrm{Cov}(y_r, y_s)\, C^{\mathrm{T}}_{sj} \\ &= (\boldsymbol{C}\, \boldsymbol{V} \boldsymbol{C}^{\mathrm{T}})_{ij} \qquad \Longrightarrow \qquad \boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{C}\, \boldsymbol{V} \boldsymbol{C}^{\mathrm{T}} \end{aligned}$$

$$\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{C}\, \boldsymbol{V} \boldsymbol{C}^{\mathrm{T}} = \underbrace{(\boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A})^{-1}\, \boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1}}_{\boldsymbol{C}}\ \ \boldsymbol{V}\ \ \underbrace{\boldsymbol{V}^{-1} \boldsymbol{A}\, (\boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A})^{-1}}_{\boldsymbol{C}^{\mathbf{T}}} \ = \ (\boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A})^{-1}$$

$$\boxed{\boldsymbol{V}(\hat{\boldsymbol{\theta}}) \ = \ (\boldsymbol{A}^{\mathbf{T}}\, \boldsymbol{V}^{-1} \boldsymbol{A})^{-1}}$$

Los estimadores son no-sesgados :

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\theta}}) &= \mathrm{E}\left[(\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{V}^{-1}\boldsymbol{A})^{-1}\,\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{V}^{-1}\mathbf{y}\right] \\ &= (\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{V}^{-1}\boldsymbol{A})^{-1}\,\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{V}^{-1}\,\mathrm{E}(\mathbf{y}) \\ &= (\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{V}^{-1}\boldsymbol{A})^{-1}\,\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{V}^{-1}\boldsymbol{A}\,\boldsymbol{\theta} \;=\; \boldsymbol{\theta} \end{aligned}$$

Teorema de Gauss-Markov: Entre todos los estimadores no sesgados, que son lineales en las observaciones, el estimador de cuadrados mínimos es el de menor varianza (pero es 100% eficiente sólo si los errores son gaussianos).

Minimización χ^2

Al método de cuadrados mínimos se los denomina también minimización χ^2 pues $S(\boldsymbol{y}, \boldsymbol{\theta})$ es una variable aleatoria con distribución χ^2_n cuando $\boldsymbol{y}$ es multinormal de dimensión n.

$$\chi^2 : \quad S(\boldsymbol{y}; \boldsymbol{\theta}) = \big(\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{\theta}\big)^{\mathbf{T}}\,\boldsymbol{V}^{\text{-}1}\,\big(\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{\theta}\big)$$

¿Qué ocurre si se remplaza θ por sus estimadores $\hat{\boldsymbol{\theta}}$?

$$\chi^2_{\min} : \quad S(\boldsymbol{y}; \hat{\boldsymbol{\theta}}) = \big(\boldsymbol{y} - \boldsymbol{A}\,\hat{\boldsymbol{\theta}}\big)^{\mathbf{T}}\,\boldsymbol{V}^{\text{-}1}\,\big(\boldsymbol{y} - \boldsymbol{A}\,\hat{\boldsymbol{\theta}}\big)$$

Ahora $S(\boldsymbol{y}; \hat{\boldsymbol{\theta}})$ no es más función de $\boldsymbol{\theta}$, es un estadístico función exclusiva de $\boldsymbol{y}$, aunque una función complicada ya que a su vez $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{y})$.

Un resultado importante es que $S(\boldsymbol{y}; \hat{\boldsymbol{\theta}})$ tiene distribución $\chi^2_{n\text{-}k}$, donde k es el número de parámetros ajustados a partir de los n datos $\boldsymbol{y}$. O sea que por cada parámetro estimado de los datos, se pierde un grado de libertad (de ahí el nombre “grados de libertad”).

La demostración se basa en probar que

$$\chi^2 = \chi^2_{\min} + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathbf{T}}\,\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{V}^{-1}\boldsymbol{A}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

El primer término tiene distribución χ^2_n y el tercero χ^2_k pues es $\big(\hat{\boldsymbol{\theta}} - \mathrm{E}(\hat{\boldsymbol{\theta}})\big)^{\mathbf{T}}\,\boldsymbol{V}^{\text{-}1}(\hat{\boldsymbol{\theta}})\,\big(\hat{\boldsymbol{\theta}} - \mathrm{E}(\hat{\boldsymbol{\theta}})\big)$. Por lo tanto el segundo término tiene que ser $\chi^2_{n\text{-}k}$.

$$\begin{aligned}
\chi^2 &= \big(\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{\theta}\big)^{\mathbf{T}}\,\boldsymbol{V}^{\text{-}1}\,\big(\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{\theta}\big) \\
&= \big(\boldsymbol{y} - \boldsymbol{A}\,\hat{\boldsymbol{\theta}} + \boldsymbol{A}\,\hat{\boldsymbol{\theta}} - \boldsymbol{A}\,\boldsymbol{\theta}\big)^{\mathbf{T}}\,\boldsymbol{V}^{\text{-}1}\,\big(\boldsymbol{y} - \boldsymbol{A}\,\hat{\boldsymbol{\theta}} + \boldsymbol{A}\,\hat{\boldsymbol{\theta}} - \boldsymbol{A}\,\boldsymbol{\theta}\big) \\
&= \big(\boldsymbol{y} - \boldsymbol{A}\,\hat{\boldsymbol{\theta}}\big)^{\mathbf{T}}\,\boldsymbol{V}^{\text{-}1}\,\big(\boldsymbol{y} - \boldsymbol{A}\,\hat{\boldsymbol{\theta}}\big) + (\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{A}\boldsymbol{\theta})^{\mathbf{T}}\,\boldsymbol{V}^{\text{-}1}\,(\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{A}\boldsymbol{\theta}) + \\
&+\ \big(\boldsymbol{y} - \boldsymbol{A}\,\hat{\boldsymbol{\theta}}\big)^{\mathbf{T}}\,\boldsymbol{V}^{\text{-}1}\,(\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{A}\boldsymbol{\theta}) + (\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{A}\boldsymbol{\theta})^{\mathbf{T}}\,\boldsymbol{V}^{\text{-}1}\,\big(\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{\theta}\big) \\
&= \chi^2_{\min} + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathbf{T}}\,\boldsymbol{A}^{\mathbf{T}}\,\boldsymbol{V}^{-1}\boldsymbol{A}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})
\end{aligned}$$

Queda como ejercicio mostrar que los dos últimos términos (iguales pues uno es el traspuesto del otro) valen cero. Esto se obtiene remplazando $\hat{\boldsymbol{\theta}}$ por su expresión en términos de $\boldsymbol{y}$.