X Congreso Latinoamericano de Sociedades de Estadística
Córdoba, Argentina, 16 al 19 de Octubre de 2012

Diagnósticos de influencia en comparación de métodos de medición en presencia de un gold estándar

Manuel Galea[1]

Departamento de Estadística, Pontificia Universidad Católica de Chile, Chile

Resumen

En este trabajo consideramos algunos diagnósticos de influencia en un modelo estadístico para comparar instrumentos de medición en presencia de un gold estándar. Suponemos que las mediciones de los instrumentos siguen una distribución normal multivariada. Implementamos el método de influencia local para analizar la sensibilidad de los estimadores máximo verosímiles a perturbaciones del modelo y/o de los datos. Finalmente ilustramos la metodología con datos reales.

Palabras Claves: Diagnósticos de influencia, Influencia local, Comparación de métodos de medición, Gold estándar.

1. Introduction

En ciencias experimentales un problema común es evaluar el grado de acuerdo entre dos o más instrumentos de medición de alguna cantidad de interés. Existen varios modelos estadísticos propuestos en la literatura para abordar este tópico, ver Dunn (2004) y referencias allí citadas. En este trabajo consideramos el modelo propuesto por St. Laurent (1998), donde se asume la presencia de un gold estándar, y el objetivo es evaluar el grado de acuerdo entre las mediciones hechas por uno o más métodos de medición aproximados y las mediciones

[1]Address for correspondence: Departamento de Estadística, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Macul, Santiago, Chile. E-mail: mgalea@mat.puc.cl

obtenidas por el gold estándar. En efecto, supongamos que una característica x es medida por $p \geq 2$ métodos aproximados y por el gold standard en un grupo común de n unidades experimentales. El modelo propuesto por St. Laurent (1998) está dado por:

$$y_{ij} = x_i + \epsilon_{ij}, \tag{1}$$

donde y_{ij} corresponde a la medición de la cantidad x hecha por el método j en la unidad i, x_i es la medición hecha por el gold estándar en la $i-$ésima unidad y ϵ_{ij} es el error aleatorio de medición sobre la $i-$ésima unidad experimental por el $j-$ésimo método aproximado, con x_i independiente de $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{ip})^T$, $E(x_i) = \mu$, $Var(x_i) = \phi$ y $E(\epsilon_{ij}) = 0$ para $i = 1, ..., n$ y $j = 1, ..., p$. Para permitir la posibilidad de que los errores aleatorios de medición de los métodos aproximados sobre la $i-$ésima unidad experimental sean correlacionados, St. Laurent (1998) asume una estructura de covarianzas general para $\boldsymbol{\epsilon}_i$, $Var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}$, una matriz $p \times p$ definida positiva con elementos σ_{jk}. Una consecuencia de estas suposiciones es que $cov(y_{ij}, x_i) = \phi$ y $cov(y_{ij}, y_{ik}) = \phi + \sigma_{jk}$. St. Laurent (1998) usa $\rho_j = \phi/(\phi + \sigma_{jj})$, el cuadrado del coeficiente de correlación entre y_{ij} y x_i, como medida de acuerdo entre el método aproximado j y el gold estándar. Note que ρ_j, $j = 1, ..., p$, no depende de los elementos fuera de la diagonal de $\boldsymbol{\Sigma}$ y como, en general, $\sigma_{jj} \neq \sigma_{kk}$, este modelo permite diferentes niveles de acuerdo (ρ_j) entre cada uno de los métodos aproximados y el gold estándar. El modelo (1) en notación matricial lo podemos escribir como:

$$\boldsymbol{Y}_i = x_i \mathbf{1}_p + \boldsymbol{\epsilon}_i, \tag{2}$$

donde, $\boldsymbol{Y}_i = (y_{i1}, ..., y_{ip})^T$. Sea $\boldsymbol{Z}_i = (x_i, \boldsymbol{Y}_i^T)^T$, el vector $q \times 1$, con $q = p + 1$, de mediciones hechas por el gold estándar y por los métodos aproximados en la unidad i, $i = 1, ..., n$. Entonces tenemos que los vectores aleatorios, $\boldsymbol{Z}_i$, son iid con $E(\boldsymbol{Z}_i) = \mu \mathbf{1}_q$ y $Var(\boldsymbol{Z}_i) = \boldsymbol{V}$, donde,

$$\boldsymbol{V} = \begin{pmatrix} \phi & \phi \mathbf{1}_p^T \\ \phi \mathbf{1}_p & \phi \mathbf{1}_p \mathbf{1}_p^T + \boldsymbol{\Sigma} \end{pmatrix}. \tag{3}$$

St. Laurent (1998), bajo el supuesto de normalidad de las mediciones, $\boldsymbol{Z}_i$, encuentra los EMV de los parámetros y también de ρ_j.

El principal objetivo de este trabajo es implementamos algunas técnicas de diagnóstico, en el modelo propuesto por St. Laurent (1998), para detectar posibles outliers presentes en los datos, que pueden distorsionar nuestras inferencias.

2. Modelo Gold Estándar Normal

Siguiendo a St. Laurent (1998), suponemos normalidad. En efecto, suponemos que los vectores aleatorios $\boldsymbol{Z}_i$ son iid $N_q(\mu\mathbf{1}_q, \boldsymbol{V})$, una distribución normal multivariada con media vector $\mu\mathbf{1}_q$ y matriz de covarianza $\boldsymbol{V}$, dada en (3) y densidad dada por,

$$f(\boldsymbol{z}_i, \boldsymbol{\theta}) = (2\pi)^{-q/2}|\boldsymbol{V}|^{-1/2}e^{-\delta_i/2}, \tag{4}$$

donde, $\delta_i(\boldsymbol{\theta}) = \delta_i = (\boldsymbol{z}_i - \mu\mathbf{1}_q)^T\boldsymbol{V}^{-1}(\boldsymbol{z}_i - \mu\mathbf{1}_q) = \delta_{ig} + \delta_{ia}$, donde $\delta_{ig} = (x_i - \mu)^2/\phi$ y $\delta_{ia} = \boldsymbol{D}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{D}_i$, $\boldsymbol{\theta} = (\mu, \phi, \boldsymbol{\sigma})^T$, con $\boldsymbol{\sigma} = v(\boldsymbol{\Sigma})$ and $\boldsymbol{D}_i = \boldsymbol{Y}_i - x_i\mathbf{1}_p$, para $i = 1, ..., n$.

La densidad (4) puede ser escrita como

$$f(\boldsymbol{z}_i, \boldsymbol{\theta}) = f_{N1}(x_i, \mu, \phi) * f_{Np}(\boldsymbol{d}_i, \boldsymbol{\sigma}) = \frac{1}{\sqrt{2\pi\phi}}e^{-\delta_{ig}/2} * \frac{1}{(2\pi)^{p/2}\sqrt{|\boldsymbol{\Sigma}|}}e^{-\delta_{ia}/2}. \tag{5}$$

O sea las variables aleatorias x_i y $\boldsymbol{D}_i$ son independientes, con distribuciones $N(\mu, \phi)$ y $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ respectivamente, para $i = 1, ..., n$.

Luego la densidad conjunta para las n observaciones está dada por

$$f(\boldsymbol{z}_1, ..., \boldsymbol{z}_n) = \prod_{t=1}^{n} f(\boldsymbol{z}_i) = \prod_{t=1}^{n}(2\pi)^{-p/2}|\boldsymbol{V}|^{-1/2}e^{-\delta_i/2}, \tag{6}$$

de donde sigue que la función de log-verosimilitud correspondiente al modelo (6) está dada por,

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}), \tag{7}$$

donde, $l_i(\boldsymbol{\theta}) = log f(\boldsymbol{z_i})$ para $i = 1, ..., n$.

3. Influencia local

Sean $\boldsymbol{\omega}$ un vector de perturbación $r \times 1 \in \boldsymbol{\Omega}$ subconjunto de $\mathbb{R}^r$, y el siguiente modelo estadístico perturbado $\mathcal{M} = \{f(Z, \boldsymbol{\theta}, \boldsymbol{\omega}) : \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, donde $f(Z, \boldsymbol{\theta}, \boldsymbol{\omega})$ es la función de densidad $Z = (\boldsymbol{z}_1, ..., \boldsymbol{z}_n)^T$ perturbada por $\boldsymbol{\omega}$ y $\ell(\boldsymbol{\theta}, \boldsymbol{\omega}) = \log f(Z, \boldsymbol{\theta}, \boldsymbol{\omega})$ su correspondiente función de log-verosimilitud. Denotando el vector de no perturbación, vector nulo, por $\boldsymbol{\omega}_0$, suponemos que $\ell(\boldsymbol{\theta}, \boldsymbol{\omega}_0) = \ell(\boldsymbol{\theta})$. Para evaluar la influencia de la perturbación sobre el estimador de máxima verosimilitud, EMV, de $\boldsymbol{\theta}$, podemos utilizar el desplazamiento de verosimilitudes

$DV(\boldsymbol{\omega}) = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_\omega)\}$, donde $\widehat{\boldsymbol{\theta}}_\omega$ y $\widehat{\boldsymbol{\theta}}$ denota el EMV bajo $\mathcal{M}$ y bajo el modelo postulado, respectivamente. La idea de influencia local, Cook (1986), es caracterizar el comportamiento de $DV(\boldsymbol{\omega})$ en $\boldsymbol{\omega}_0$. Por ejemplo, Cook (1986) demuestra que la curvatura normal en la dirección $\boldsymbol{h}$, con $||\boldsymbol{h}|| = 1$, toma la forma

$$C_h(\boldsymbol{\theta}) = 2|\boldsymbol{h}^T \boldsymbol{\Delta}^T \boldsymbol{L}^{-1} \boldsymbol{\Delta} \boldsymbol{h}|, \tag{8}$$

donde $-\boldsymbol{L}$ es la matriz de información observada para el modelo postulado ($\boldsymbol{\omega} = \boldsymbol{\omega}_0$) y $\boldsymbol{\Delta}$ es una matriz $p^* \times r$ con elementos $\Delta_{ij} = \partial^2 \ell(\boldsymbol{\theta}, \boldsymbol{\omega}) / \partial\theta_i \partial\omega_j$, evaluada en $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ y $\boldsymbol{\omega} = \boldsymbol{\omega}_0, i = 1, ..., p^*$ y $j = 1, ..., r$. Aquí p^* denota el número de parámetros del modelo postulado. En este trabajo usamos a $\boldsymbol{h}_{\max}$ y a C_i como diagnósticos de influencia local, bajo esquemas de perturbación frecuentemente utilizados en la literatura, ver Cook (1986), Verbeke y Molenberghs (2000) y Galea, Bolfarine y de Castro (2002).

Acknowledgements Este trabajo es financiado parcialmente por el Proyecto Fondecyt 1110318, Conicyt, Chile.

Referencias

Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society*, B, 48, 133–169.

Dunn, G. (2004). *Statistical Evaluation of Measurement Errors: Design and Analysis of Reliability Studies*. Edward Arnold. New York.

Galea, M., Bolfarine, H. y de Castro, M. (2002). Local influence in comparative calibrations models. *Biometrical Journal*, 44, 59–81.

St. Laurent, R. T. (1998). Evaluating agreement with a gold standard in method comparison studies. *Biometrics,* 54, 2, 537-545.

Verbeke, G. y Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer. New York.