Contribuciones a la Estadística
y a la Investigación Operativa
Sicilia et. al. (editores)

ISBN 84-689-8552-X
páginas 533–541

Estadísticos basados en divergencias para el diagnóstico de modelos

M.D. Esteban[1], *T. Hobza*[2], *Y. Marhuenda*[1], *D. Morales*[1],

[1]md.esteban@umh.es, y.marhuenda@umh.es, d.morales@umh.es, Centro de Investigación Operativa, Universidad Miguel Hernández de Elche
[2]hobza@km1.fjfi.cvut.cz, Departamento de Matemáticas, Universidad Politécnica Checa

Abstract

En este trabajo se introduce un procedimiento basado en divergencias para el contraste de bondad de ajuste cuando las observaciones no son idénticamente distribuidas, se estudian sus propiedades asintóticas y se consideran sus versiones bootstrap. La metodología es aplicable a la diagnosis de modelos lineales generalizados.

Palabras Clave: Divergencias, modelos lineales generalizados, bootstrap.

AMS: 62F03, 62B10, 65C05, 65C60.

1. Introducción

El problema de bondad de ajuste a una distribución en la recta real, $H_0 : F = F_0$, se trata frecuentemente particionando el rango de los datos en intervalos disjuntos y contrastando la hipótesis $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$ en una distribución multinomial.

Sean $Y_1, \ldots, Y_n$ variables aleatorias i.i.d. con función de distribución F. Sea $E_1, \ldots, E_m$ una partición de $R = (-\infty, \infty)$ en m intervalos. Sea $\boldsymbol{p} = (p_1, \ldots, p_m)$ y $\boldsymbol{p}_0 = (p_{01}, \ldots, p_{0m})$ los vectores de probabilidades verdaderas e hipotéticas de los intervalos E_k; es decir,

$$p_{0k} = \int_{E_k} dF_0, \quad p_k = \int_{E_k} dF, \quad k = 1, \ldots, m.$$

Definimos los conteos observados

$$N_k = \sum_{j=1}^{n} 1_{(Y_j \in E_k)} = \#(1 \leq j \leq n : \ Y_j \in E_k), \quad k = 1, \ldots, m,$$

y las probabilidades estimadas de las celdas $\widehat{\boldsymbol{p}} = (\widehat{p}_1, \ldots, \widehat{p}_m)$ con $\widehat{p}_k = N_k/n$, $k = 1, \ldots, m$. Para contrastar $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$ se usa habitualmente el estadístico χ^2 de Pearson

$$\chi^2_P(\widehat{\boldsymbol{p}}, \boldsymbol{p}_0) = n \sum_{k=1}^{m} \frac{(\widehat{p}_k - p_{0k})^2}{p_{0k}}, \tag{138}$$

que es un caso particular de la familia de divergencias introducidas por Cressie y Read [1]

$$T_n^r(\widehat{\boldsymbol{p}}, \boldsymbol{p}_0) = \frac{2n}{r(r+1)} \sum_{k=1}^{m} \widehat{p}_k \left[\left(\frac{\widehat{p}_k}{p_{0k}} \right)^r - 1 \right], \quad -\infty < r < \infty. \tag{139}$$

Los estadísticos $T_n^0(\widehat{\boldsymbol{p}}, \boldsymbol{p}_0)$ y $T_n^{-1}(\widehat{\boldsymbol{p}}, \boldsymbol{p}_0)$ se definen por continuidad. Algunos estadísticos conocidos se obtienen especificando valores de r en (139). Algunos ejemplos son $r = 1$ para el test de Pearson, $r = 0$ para el test de la razón de verosimilitudes, $r = -1/2$ para el test de Freeman-Tukey, $r = -2$ para el test modificado de Neyman y $r = 2/3$ para el test de Cressie-Read.

Se verifica además que $T_n^r(\widehat{\boldsymbol{p}}, \boldsymbol{p}_0)$ es un caso particular del estadístico ϕ-divergencia

$$T_n^\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}_0) = \frac{2n}{\phi''(1)} D_\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}_0) = \frac{2n}{\phi''(1)} \sum_{k=1}^{m} p_{0k} \phi\left(\frac{\widehat{p}_k}{p_{0k}} \right), \tag{140}$$

donde $D_\phi(\cdot, \cdot)$ denota la ϕ-divergencia entre dos distribuciones de probabilidad introducida por Csiszár [2] y Ali y Silvey [3] para cualquier ϕ del conjunto Φ de funciones reales, convexas, definidas en $[0, \infty)$, continuamente diferenciables en un entorno de 1 y verificando $\phi(1) = \phi'(1) = 0$, $\phi''(1) > 0$. En la fórmula (140) si p_{0k} o p_{0k} y $\widehat{p}_k$ son cero, las expresiones $0\phi(x/0)$ y $0\phi(0/0)$ se definen como $x \cdot \lim_{u\to\infty} \phi(u)/u$ y 0 respectivamente. Las propiedades de las ϕ-divergencias han sido estudiadas por Liese y Vajda [4] y Vajda [5]. Zografos y otros [6] demostraron que $T_n^\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}_0) \xrightarrow{L} \chi^2_{m-1}$ cuando $n \to \infty$ bajo $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$, donde $\xrightarrow{L}$ denota convergencia en ley.

Con frecuencia interesa contrastar la hipótesis compuesta de que una función de distribución F pertenece a una cierta familia paramétrica $\{F_\theta\}_{\theta\in\Theta}$, donde $\Theta \subset R^d$ es un subconjunto abierto. En tales casos, las probabilidades de las celdas dependen de un parámetro desconocido θ; es decir,

$$p_k(\theta) = \int_{E_k} dF_\theta, \quad k = 1, \ldots, m,$$

de modo que pueden ser estimadas usando el estimador de mínima ϕ-divergencia

$$\widehat{\theta}_\phi = \arg \inf_{\theta\in\Theta} D_\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\theta)),$$

el cual contiene como caso particular al estimador de máxima verosimilitud (EMV) basado en los datos categorizados. Morales y otros [7] demostraron que si se verifican las condiciones de regularidad de Birch [8], entonces

$$T_n^{\phi_1}(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\widehat{\theta}_{\phi_2})) \xrightarrow{L} \chi^2_{m-d-1}, \quad \text{cuando } n \to \infty,$$

bajo $H_0 : F = F_\theta$ para cualesquiera $\phi_1, \phi_2 \in \Phi$. Sin embargo, si el EMV $\widehat{\theta}$ está basado en los datos originales, entonces la distribución asintótica de $T_n^\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\widehat{\theta}))$ bajo $H_0 : F = F_\theta$ es una combinación lineal de variables χ^2_1 independientes. Este resultado fue originalmente demostrado por Chernoff y Lehmann [9] y posteriormente extendido a cualquier $\phi \in \Phi$ por Morales y otros [7].

Si las variables originales son independientes con distribuciones $F_1, \ldots, F_n$, dependientes de un parámetro desconocido $\theta \in \Theta \subset R^d$ abierto, la hipótesis de interés es

$$H_0 : Y_1 \sim F_1, \ldots, Y_n \sim F_n. \tag{141}$$

Definimos $p_k(\theta) = \mathsf{E}_\theta[N_k]/n$, con $\mathsf{E}_\theta[N_k] = \sum_{j=1}^n P_\theta(Y_j \in E_k)$. Jiang [9] propuso contrastar H_0 con

$$\chi^2_J(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\widehat{\theta})) = n \sum_{k=1}^m (\widehat{p}_k - p_k(\widehat{\theta}))^2, \tag{142}$$

donde $\widehat{\theta}$ es un estimador consistente de θ, y dio condiciones de regularidad bajo las cuales la distribución asintótica de $\chi^2_J(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\widehat{\theta}))$ es una combinación lineal de variables χ^2_1 independientes.

El objetivo de este trabajo es extender el resultado de Jiang a la clase de estadísticos $T_n^\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\widehat{\theta}))$ e introducir sus versiones bootstrap. En la sección 2 se deduce la distribución asintótica de $T_n^\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\widehat{\theta}))$ y en la sección 3 se introducen los tests bootstrap asociados.

2. Distribución asintótica del estadístico T_n^ϕ

En esta sección se obtiene la distribución asintótica del estadístico

$$T_n^\phi = T_n^\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\widehat{\theta})) = \frac{2n}{\phi''(1)} \sum_{k=1}^m p_k(\widehat{\theta})\, \phi\left(\frac{\widehat{p}_k}{p_k(\widehat{\theta})}\right) \tag{143}$$

para la clase de funciones $\phi \in \Phi$ bajo la hipótesis nula (141). Ello permite construir un test de bondad de ajuste para el caso de observaciones independientes pero no idénticamente distribuidas. El articulo de Jiang [9] dando la distribución asintótica de $\chi^2_J(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\widehat{\theta}))$ es esencial. Comenzaremos introduciendo algo de notación y las condiciones de regularidad dadas por Jiang.

Se sabe que la elección de $\widehat{\theta}$ tiene un gran impacto en la distribución asintótica de T_n^ϕ. En este artículo se supone que $\widehat{\theta}$ es un estimador consistente de θ que

admite la forma asintótica

$$\sqrt{n}(\widehat{\theta}-\theta) = A_n\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\psi_j(Y_j,\theta)\right) + o_P(1). \tag{144}$$

Por ejemplo, bajo condiciones de regularidad, el EMV de θ tiene la forma (144), donde ψ_j es la función de puntuaciones correspondiente a la observación j y A_n es igual a n veces la matriz de información de Fisher (la matriz basada en todos los datos).

Sea $\xi_n = (\xi_{nk})_{1\le k\le m}$, donde $\xi_{nk} = N_k - \mathsf{E}_{\widehat{\theta}}N_k$; $\boldsymbol{p}_j(\theta) = (p_{jk}(\theta))_{1\le k\le m}$ y $p_{jk}(\theta) = P_\theta(Y_j \in E_k)$. Definimos

$$h_{nj} = (1_{(Y_j\in E_k)} - p_{jk}(\theta))_{1\le k\le m} - \left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial\theta}\boldsymbol{p}_j(\theta)\right)A_n\psi_j(Y_j,\theta)$$

y $\Sigma_n = \Sigma_n(\theta) = n^{-1}\sum_{j=1}^{n}\mathsf{Var}(h_{nj})$. Sea T_n una matriz ortogonal tal que

$$T_n^t\Sigma_n T_n = D_n = \mathrm{diag}(\lambda_{n1},\ldots,\lambda_{nm}),$$

donde $\lambda_{n1}\ge\ldots\ge\lambda_{nm}$ son los autovalores de Σ_n.

Se admiten las siguientes hipótesis:

(i) $Y_1,\ldots,Y_n$ son independientes,

(ii) $\Sigma_n \longrightarrow \Sigma$ cuando $n\to\infty$,

(iii) Se verifica (144) con $\mathsf{E}\psi_j(Y_j,\theta)=0$, $1\le j\le n$,

(iv) Se verifica que

$$\frac{1}{n}\max_{1\le j\le n}\mathsf{E}|A_n\psi_j(Y_j,\theta)|^4 \longrightarrow 0, \quad \max_{1\le j\le n}\left|\frac{\partial}{\partial\theta}\boldsymbol{p}_j(\theta)\right| = O(1),$$

y existe un $\delta>0$ tal que

$$\frac{1}{n}\sum_{j=1}^{n}\sup_{|\widetilde{\theta}-\theta|\le\delta}\left\|\frac{\partial^2}{\partial\theta^2}p_{jk}(\widetilde{\theta})\right\| = O(1), \quad 1\le k\le m.$$

Bajo las hipótesis (i)-(iv) Jiang demostró que la distribución asintótica de χ^2_J es la misma que la de $\sum_{k=1}^{m}\lambda_k Z_k^2$ donde $Z_1,\ldots,Z_m$ son variables i.i.d. $N(0,1)$ y $\lambda_1,\ldots,\lambda_m$ son los autovalores de Σ. En primer lugar presentamos el resultado

equivalente para el estadístico de Pearson $\chi^2_P(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\widehat{\theta}))$ definido en (138). Para alcanzar este objetivo se necesita la hipótesis adicional

$$\boldsymbol{p}(\theta) = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{p}_j(\theta) \longrightarrow \boldsymbol{q}, \quad \text{donde} \quad q_k > 0 \text{ para todo } k \in \{1, \ldots, m\}, \quad (145)$$

cuando $n \to \infty$. Los resultados de esta sección se presentan en el Lema 1 y en el Teorema 1.

Lema 1. Si se verifican las hipótesis (i)-(iv) y (145) entonces el estadístico

$$T_n^1 = \chi^2_P(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\widehat{\theta}))$$

tiene, bajo la hipótesis nula (141), la misma distribución asintótica que $\sum_{k=1}^{m}(\lambda_k/q_k)Z_k^2$, donde $Z_1, \ldots, Z_m$ son variables i.i.d. $N(0,1)$ y $\lambda_1 \geq \ldots \geq \lambda_m$ son los autovalores de Σ.

Teorema 1. Si se verifica (i)-(iv) y (145) entonces para cualquier $\phi \in \Phi$ el estadístico

$$T_n^{\phi} = T_n^{\phi}(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\widehat{\theta}))$$

definido en (143) tiene, bajo la hipótesis nula (141), la misma distribución asintótica que $\sum_{k=1}^{m}(\lambda_k/q_k)Z_k^2$, donde $Z_1, \ldots, Z_m$ son variables i.i.d. $N(0,1)$ y $\lambda_1 \geq \ldots \geq \lambda_m$ son los autovalores de Σ.

Obsérvese que para usar la clase de estadísticos de contraste T_n^{ϕ} hay que reemplazar los autovalores $\lambda_1, \ldots, \lambda_m$ y el vector de probabilidades $\boldsymbol{q} = (q_1, \ldots, q_m)$ por sus respectivos estimadores. Bajo las hipótesis del Lema 1, se comprueba que $\boldsymbol{p}(\widehat{\theta}) = (1/n)\sum_{j=1}^{n} \boldsymbol{p}_j(\widehat{\theta})$ es un estimador consistente del vector $\boldsymbol{q}$. Si denotamos a los autovalores de $\widehat{\Sigma}_n = \Sigma_n(\widehat{\theta})$ mediante $\widehat{\lambda}_{n1}, \ldots, \widehat{\lambda}_{nm}$ entonces, aplicando el teorema de perturbación de Weyl (véase Bhatia [11]), se obtiene que $|\widehat{\lambda}_{nk} - \lambda_{nk}| \leq \|\Sigma_n(\widehat{\theta}) - \Sigma_n(\theta)\|$ tiende a cero cuando $n \to \infty$ al ser $\widehat{\theta}$ consistente. Aplicando el mismo teorema se obtiene que $\lambda_{nk} \to \lambda_k$ y en consecuencia $\widehat{\lambda}_{nk}$ es un estimador consistente de λ_k, $k = 1, \ldots, m$. Así pues, se puede proponer el siguiente procedimiento de contraste: Rechazar H_0 si T_n^{ϕ} excede el valor crítico de $\sum_{k=1}^{m}(\widehat{\lambda}_{nk}/p_k(\widehat{\theta}))\, Z_k^2$.

Como el cálculo de los cuantiles de la distribución de $\sum_{k=1}^{m}(\widehat{\lambda}_{nk}/p_k(\widehat{\theta}))\, Z_k^2$ no está implementado en los paquetes de estadística, Rao y Scott [2] sugirieron considerar la distribución aproximada $\delta\chi^2_m$, donde $\delta = \frac{1}{m}\sum_{k=1}^{m}(\widehat{\lambda}_{nk}/p_k(\widehat{\theta})) = \frac{1}{m}\text{traza}(\text{diag}(p_1^{-1}(\widehat{\theta}), \ldots, p_m^{-1}(\widehat{\theta}))\Sigma_n(\widehat{\theta}))$.

3. Tests bootstrap de bondad de ajuste

La aplicación del estadístico de Jiang y de T_n^{ϕ}, definidos en (142) y (143) respectivamente, para contrastar la hipótesis (141) requiere el uso de sus distribucio-

nes asintóticas. Los estadísticos aplicados encontrarán las siguientes dificultades al aplicar este enfoque.

a. En casos no triviales, la deducción de Σ_n no es directa y el cálculo numérico es necesario.

b. Σ_n se estima con $\widehat{\Sigma}_n = \Sigma_n(\widehat{\theta})$ y $\Sigma_n(\widehat{\theta})$ se supone próxima a $\Sigma_n(\theta)$. En consecuencia el tamaño muestral debería ser suficientemente grande para aproximar el tamaño del test.

Los tests bootstrap evitan las dificultades mencionadas ya que solamente requieren el cálculo del estadístico en muestras bootstrap independientes. De esa forma se aproxima su distribución bajo H_0.

Sean las variables aleatorias $Y_1, \ldots, Y_n$ y las funciones de distribución $F_{1\theta}, \ldots, F_{n\theta}$ dependientes de un parámetro común $\theta \in \Theta \subset R^d$ abierto. La hipótesis (141) de interés es

$$H_0 : Y_1 \sim F_{1\theta}, \ldots, Y_n \sim F_{n\theta} \text{ independientes, } \theta \in \Theta.$$

Sea $T_n = T_n(Y_1, \ldots, Y_n)$ un estadístico para este problema y supongamos que se rechaza H_0 si $T_n > c_n$ para un valor crítico $c_n > 0$. Sea $F_{T_n\theta}(x) = P_\theta^n(T_n \leq x)$ la distribución de T_n bajo H_0, donde P_θ^n es la probabilidad correspondiente a la distribución conjunta $\prod_{j=1}^n F_{j\theta}$. Supongamos que se tiene un estimador $\widehat{\theta}$ de θ tal que $\widehat{\theta}$ es consistente bajo H_0 en el sentido de que

$$P_\theta^n \left(\|\widehat{\theta} - \theta\| > \varepsilon \right) \longrightarrow 0, \quad \text{para todo } \varepsilon > 0,$$

cuando $n \to \infty$. Suponiendo que $F_{T_n\theta}$ es continua, un estimador bootstrap de c_n es

$$\widehat{c}_n = F^{-1}_{T_n\widehat{\theta}}(1-\alpha),$$

donde $\alpha \in (0,1)$ es el tamaño del test. El cálculo de $\widehat{c}_n$ puede hacerse por simulación Monte Carlo de la siguiente forma. Generar B muestras bootstrap independientes $\{Y_{1b}^*, \ldots, Y_{nb}^*\}$ de la distribución conjunta $\prod_{j=1}^n F_{j\widehat{\theta}}$. Entonces $\widehat{c}_n$ se puede aproximar por el estadístico de orden $\{[(1-\alpha)B]+1\}$ en el conjunto de valores $\{T_n(Y_{1b}^*, \ldots, Y_{nb}^*) : b = 1, \ldots, B\}$.

Alternativamente el p-valor bootstrap estimado puede usarse para decidir si H_0 se rechaza o no. Sean $Y_1 = y_1, \ldots, Y_n = y_n$ los valores observados. El p-valor de un test con región de rechazo de la forma $T_n > c$ es

$$p_n = P_\theta^n(T_n(Y_1, \ldots, Y_n) > T_n(y_1, \ldots, y_n)),$$

y se rechaza la hipótesis nula si $p_n < \alpha$. Un estimador bootstrap de p_n es

$$\widehat{p}_n = P_*(T_n(Y_1^*, \ldots, Y_n^*) > T_n(y_1, \ldots, y_n)),$$

donde $Y_1^* \sim F_{1\widehat{\theta}}, \ldots, Y_n^* \sim F_{n\widehat{\theta}}$ son los datos bootstrap independientes. El cálculo de $\widehat{p}_n$ puede hacerse por simulación Monte Carlo de la siguiente forma. Generar B muestras bootstrap independientes $\{Y_{1b}^*, \ldots, Y_{nb}^*\}$ de la distribución conjunta $\prod_{j=1}^n F_{j\widehat{\theta}}$. Entonces $\widehat{p}_n$ se aproxima mediante

$$\widehat{p}_n = \frac{\#\left(T_n(y_1^*, \ldots, y_n^*) > T_n(y_1, \ldots, y_n)\right)}{B}.$$

Este enfoque se puede usar para calcular el p-valor

$$p_n = P\left(\sum_{k=1}^{m} (\widehat{\lambda}_{nk}/p_k(\widehat{\theta}))\, Z_k^2 > t\right)$$

cuando $T_n^\phi = t$ ha sido observado.

4. Agradecimientos

Este trabajo ha sido financiado por los proyectos BMF2003-04820, GV04B-670 y MSMTV 1M0572. Uno de los autores agradece la ayuda económica recibida del Cento de Investigación Operativa durante su estancia en Elche.

5. Bibliografía

[1] Cressie N.A.C. and Read T.R.C. (1984). Multinomial goodness of fit tests. *Journal of the Royal Statistical Society, Series B*, 46, 440-464.

[2] Csiszár I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of Hungarian Academy of Sciences, Series A*, 8, 85-108.

[3] Ali S.M. and Silvey S.D. (1966). A general class of coefficient of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 286, 131-142.

[4] Liese F. and Vajda I. (1987). *Convex Statistical Distances*. Teubner, Leipzig.

[5] Vajda I. (1989). *Theory of Statistical Inference and Information*, Kluwer. Dordrecht.

[6] Zografos K., Ferentinos K. and Papaioannou T. (1990). ϕ-divergence statistics: sampling properties, multinomial goodness of fit and divergence tests. *Communications in Statistics - Theory and Methods*, 19, 1785-1802.

[7] Morales D., Pardo L. and Vajda I. (1996). Divergence between various estimates of quantized information sources. *Kybernetika*, 32, 4, 395-407.