QÜESTIIÓ, vol. 25, 2, p. 263-284, 2001

EL SESGO CONDICIONADO EN EL ANÁLISIS DE INFLUENCIA: UNA REVISIÓN

J. M. MUÑOZ-PICHARDO
J. L. MORENO-REBOLLO
T. GÓMEZ-GÓMEZ
A. ENGUIX-GONZÁLEZ
Universidad de Sevilla*

El sesgo condicionado se ha propuesto como diagnóstico de influencia en distintos modelos y técnicas estadísticas. Tratando de recoger una visión global de la utilidad del concepto, en este trabajo se hace una revisión general del mismo relacionándolo con la curva de sensibilidad y la curva de influencia muestral. Además, se señalan posibles líneas de trabajo que permitirán abordar el análisis de la influencia a través de este enfoque en una gran variedad de técnicas estadísticas.

Conditional bias in influence analysis: a review

Palabras clave: Análisis de influencia, modelos lineales, componentes principales, muestreo en poblaciones finitas, sesgo condicionado

Clasificación AMS (MSC 2000): 62J20, 62H25, 62D05

* Universidad de Sevilla. Facultad de Matemáticas. Departamento de Estadística e Investigación Operativa. Avda. Reina Mercedes s/n. 41012 Sevilla.

– Recibido en octubre de 2000.
– Aceptado en abril de 2001.

1. INTRODUCCIÓN

El objetivo de todo análisis estadístico es obtener conclusiones fiables a partir de los datos resultantes de una experimentación. Por tanto, la fiabilidad de las observaciones del proceso es de especial interés, ya que el análisis se realiza sobre codificaciones del fenómeno natural en estudio y las técnicas estadísticas que se apliquen pueden verse fuertemente afectadas por algunas de las observaciones realizadas. Este problema ha originado un gran número de métodos enfocados, bien al desarrollo de nuevas técnicas que no se vean influenciadas excesivamente por la modelización del fenómeno natural, bien al análisis de la calidad de los datos, o bien al estudio de aquellas observaciones que afectan considerablemente a los resultados del análisis. En este tercer enfoque se han propuesto un conjunto de métodos englobados en lo que genéricamente se conoce como el **Análisis de Influencia**.

La gran mayoría de las técnicas propuestas para el análisis de influencia están basadas en la aproximación genérica realizada por Cook y Weisberg (1982): para medir el efecto que sobre un aspecto de interés del análisis (estadístico, estimación,...) tiene una observación o un conjunto de ellas, se introducen pequeñas perturbaciones, que las afectan de alguna manera, y se cuantifica el cambio producido. Es decir, las técnicas surgen de conjugar adecuadamente *perturbación* del modelo y *comparación* de resultados. Cook (1987) trata de unificar el problema bajo la siguiente formulación general:

> Sea un conjunto de datos D, un modelo M postulado a priori, un resultado $R(D,M)$ seleccionado de una síntesis de los datos y el modelo, y sea $\mathbf{w}$ un vector de perturbaciones, perteneciente a un conjunto Ω de perturbaciones relevantes, siendo $M(\mathbf{w})$ el modelo perturbado, de forma que
>
> $$\exists \mathbf{w}_0 \in \Omega \ / \ M \approx M(\mathbf{w}_0) \ .$$
>
> Así, el Análisis de Influencia consiste en comparar los resultados $R(D,M(\mathbf{w}))$ y $R(D,M)$.

En consecuencia, son cuestiones claves la elección del esquema de perturbación y el método de comparación. De cada forma distinta de conjugar el binomio *perturbaci ón-comparación* resultará una técnica para el análisis de influencia.

En cuanto al primer elemento del binomio, el más utilizado es el esquema de perturbación de la omisión de las observaciones a las que se le pretende evaluar su impacto. No obstante, la anterior formulación general del problema permite otros enfoques. Diversos autores, como Lawrance (1991) y Escobar y Meeker (1992), han propuesto clasificaciones de los distintos tipos de perturbaciones: *perturbaciones en los datos, perturbaciones en las hipótesis del modelo, perturbaciones en ponderaciones de casos.*

Este último esquema es utilizado en el denominado Análisis de Influencia Local (Cook (1986)).

Una vez seleccionado el esquema de perturbación, es necesario elegir el estadístico o resultado $R(D,M)$ del análisis y el método de comparación de resultados. La primera cuestión queda a criterio del investigador, en función del objetivo principal del análisis. Como ilustración de la gran variedad de métodos de comparación propuestos, pueden citarse los siguientes: *Curva de influencia muestral* (Cook y Weisberg (1982), Belsley *y otros* (1980)); *Razón de volúmenes de regiones de confianza*, cuando el resultado $R(D,M)$ seleccionado es una región de confianza (Andrews y Pregibon (1978)); *Raz ón entre determinantes*, para la estimación de matrices de varianzas y covarianzas (Barnett y Lewis (1994), Belsley *y otros* (1980)); *Distancias entre las distribuciones muestrales de los estimadores*, como la distancia de Rao (Muñoz-Pichardo y Fernández-Ponce (1997)); *Desplazamiento de verosimilitud*, como método de comparación respecto a los contornos de la log-verosimilitud del modelo postulado (Cook y Weisberg (1982), Cook *y otros* (1988)), o bien del modelo perturbado (Billor y Loynes (1993)). Brown y Lawrance (2000), a través de la verosimilitud, estudian este tópico en regresión múltiple sobre una amplia gama de esquemas de perturbación, extendiendo el estudio de la influencia a diferentes tests de hipótesis de interés en este modelo.

Muñoz-Pichardo *y otros* (1995) proponen el concepto de sesgo condicionado como enfoque genérico, válido en un gran número de las técnicas y modelos estadísticos. Tras la introducción anterior sobre el problema de la influencia, en este trabajo se pretende realizar una revisión de las distintas aplicaciones del sesgo condicionado (s.c.), así como su relación con otros conceptos estadísticos.

En la Sección 2, se recoge la relación con la curva de influencia muestral y la curva de sensibilidad. En la Sección 3, se recoge la aplicación del mismo a los modelos lineales, aportándose una estimación de la varianza del estimador del sesgo condicionado. A continuación, se aplica al análisis de componentes principales (Sección 4) y en el muestreo en poblaciones finitas (Sección 5). Finalmente, se concluye señalando otros campos en los que se puede abordar el análisis de influencia bajo la perspectiva del concepto de s.c..

2. EL CONCEPTO DE SESGO CONDICIONADO Y EL ANÁLISIS DE INFLUENCIA

Partiendo del Lema de Descomposición de Efron y Stein (1984), que expresa un estadístico T sobre una muestra aleatoria simple como una suma finita, cuyos términos son funciones de las esperanzas condicionadas de T dadas las observaciones muestrales, Muñoz-Pichardo *y otros* (1995) definen el concepto de *sesgo condicionado* y proponen su aplicación en el Análisis de Influencia. A continuación, se recoge el lema citado,

la definición del concepto de sesgo condicionado y la justificación de su aplicación al problema de la influencia.

Lema 1. (Efron y Stein). *Toda variable aleatoria $S(X_1, X_2, \ldots, X_n)$ función de n variables aleatorias independientes $X_1, X_2, \ldots, X_n$ puede expresarse como*

$$S(X_1 \ldots X_n) = E[S] + \sum_{i=1}^{n} A_i[X_i; S] + \sum\sum_{1 \le i < j \le n} B_{ij}[X_i, X_j; S] + \\ + \sum\sum\sum_{1 \le i < j < k \le n} C_{ijk}[X_i, X_j, X_k; S] + \cdots + H[X_1 \ldots X_n; S],$$

donde las $2^n - 1$ variables aleatorias del miembro derecho de la expresión tienen esperanza nula y están mutuamente incorreladas, siendo:

$$\begin{aligned} A_i[x_i; S] &= E[S \,|X_i = x_i] - E[S], \quad \text{«}i\text{-ésimo efecto medio»}, \\ B_{ij}[x_i, x_j; S] &= E[S \,|X_i = x_i,\ X_j = x_j] - E[S \,|X_i = x_i] - \\ &\quad - E[S \,|X_j = x_j] + E[S], \quad \text{«}(i,j)\text{-interacción de } 2^o \text{ orden»}, \end{aligned}$$

y así sucesivamente.

Dado un estadístico $T_n = T_n(Y_1 \ldots Y_n)$ definido sobre una muestra aleatoria, $Y_1 \ldots Y_n$, y su descomposición en términos del resultado anterior, se observa que $A_i[y_i; T_n]$ cuantifica la desviación que la realización muestral $Y_i = y_i$ provoca en el valor esperado de T_n. Por tanto, puede considerarse como una medida de la influencia que dicha realización muestral ejerce sobre T_n. Así, Muñoz-Pichardo *y otros* (1995) proponen la siguiente definición:

Definición 1. *Sea $Y_1 \ldots Y_n$ una m.a. de una v.a. Y, sea $T_n = T_n(Y_1 \ldots Y_n)$ un estadístico y sea $y_1 \ldots y_n$ una realización de la muestra. El sesgo condicionado (s.c.) de T_n dada la i-ésima observación se define como*

$$\mathcal{S}(y_i; T_n) = E[T_n \,|Y_i = y_i] - E[T_n].$$

Sobre esta definición se pueden realizar diversas consideraciones. En primer lugar, el s.c. depende de la distribución del estadístico T_n y del valor observado y_i. Por tanto, al contrario de la curva de influencia muestral, que se define a partir de una realización de toda la muestra, el s.c. mide la influencia del valor observado sobre el estadístico, en términos de la esperanza de su distribución muestral, y por tanto, es independiente de cualquier realización muestral concreta de los restantes elementos de la muestra. Por otra parte, no presupone ningún esquema de perturbación, salvo el condicionamiento

previo impuesto por el conocimiento de la observación bajo estudio. Además, es un parámetro cuya dimensión coincide con la dimensión del estadístico, al igual que ocurre con la curva de influencia muestral, por tanto, para cuantificarlo se ha de proceder a utilizar normas semejantes a las aplicadas sobre ésta. Finalmente, su dependencia de la distribución de T_n, puede provocar que el desconocimiento de algunos parámetros de la misma implique la necesidad de su estimación.

La definición 1 puede generalizarse como sigue:

Definición 2. *En las condiciones de la definición 1, el s.c. de T_n dado el conjunto de observaciones $\{y_{i_1} \dots y_{i_m}\}$ se define como*

$$\mathcal{S}(y_{i_1} \dots y_{i_m}; T_n) = E[T_n | Y_{i_1} = y_{i_1} \dots Y_{i_m} = y_{i_m}] - E[T_n]. \tag{1}$$

En consecuencia, (1) puede considerarse como una medida de la influencia conjunta de las observaciones $\{y_{i_1} \dots y_{i_m}\}$ sobre T_n.

De las dos definiciones anteriores se obtienen las siguientes expresiones:

$$B_{ij}[y_i, y_j; T_n] = \mathcal{S}(y_i, y_j; T_n) - \{\mathcal{S}(y_i; T_n) + \mathcal{S}(y_j; T_n)\}, \tag{2}$$

$$\begin{aligned} C_{ijk}[y_i, y_j, y_k; T_n] = {} & \mathcal{S}(y_i, y_j, y_k; T_n) - \\ & - \{\mathcal{S}(y_i, y_j; T_n) + \mathcal{S}(y_i, y_k; T_n) + \mathcal{S}(y_j, y_k; T_n)\} + \\ & + \{\mathcal{S}(y_i; T_n) + \mathcal{S}(y_j; T_n) + \mathcal{S}(y_k; T_n)\}. \end{aligned} \tag{3}$$

Por tanto, el efecto interacción de segundo orden debido a las observaciones (y_i, y_j), según (2), es la influencia conjunta de ambas observaciones sobre T_n menos la influencia individual de cada una de ellas. Análogo comentario se puede realizar observando la expresión (3).

Otras propiedades del s.c. son las relaciones con los conceptos de curva de sensibilidad y curva de influencia muestral. Hampel (1974), con el objetivo de estudiar la conducta infinitesimal de funcionales estadísticos, propone el concepto de función influencia sobre un funcional definido sobre el espacio de las distribuciones de probabilidad. A partir de dicha definición se han propuesto diversas versiones muestrales. Una de las de mayor interés es la *curva de sensibilidad*, propuesta por Tukey (1970): dada $Y_1 \dots Y_{n-1}$ una muestra aleatoria de una v.a. Y, la curva de sensibilidad (*CS*) de T_n asociada a una realización muestral $y_1 \dots y_{n-1}$ se define como:

$$CS_{n-1}(y; T_n) = n\{T_n(y_1, \dots, y_{n-1}, y) - T_{n-1}(y_1, \dots, y_{n-1})\}.$$

Cuando el objetivo es el estudio de la influencia de observaciones individuales sobre algún estadístico, se propone otra versión muestral, la *curva de influencia muestral*

(*CIM*): dada $Y_1 \dots Y_n$ una muestra aleatoria de una v.a. Y, y T_n un estadístico definido sobre la muestra, la curva de influencia muestral de un estadístico T_n asociada a la *i*-ésima observación de una realización muestral $y_1, \dots, y_n$, viene dada por

$$CIM_i(T_n) = -(n-1)\left\{T_{(i)} - T_n(y_1, \dots, y_n)\right\},$$

siendo $T_{(i)} = T_{n-1}(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$ el estadístico obtenido bajo la omisión de la *i*-ésima observación.

Generalizando la definición anterior, se define la curva de influencia muestral de T_n asociada a un conjunto de *m* observaciones *B* como sigue:

$$CIM_B(T_n) = -(n-m)\left\{T_{(B)} - T_n(y_1, \dots, y_n)\right\},$$

siendo $T_{(B)}$ el estadístico obtenido bajo la omisión de las observaciones contenidas en *B*.

La curva de sensibilidad puede considerarse como función aleatoria, asociada a la muestra $Y_1 \dots Y_{i-1}\, Y_{i+1} \dots Y_n$:

$$CS_{n-1}(y; T_n) = n\left\{T_n(Y_1 \dots Y_{i-1}, y, Y_{i+1} \dots Y_n) - T_{n-1}(Y_1 \dots Y_{i-1}\, Y_{i+1} \dots Y_n)\right\}.$$

Asimismo, la curva de influencia muestral puede considerarse como función aleatoria, asociada a la muestra $Y_1 \dots Y_n$:

$$CIM_i(T_n) = -(n-1)\left\{T_{n-1}(Y_1 \dots Y_{i-1}\, Y_{i+1} \dots Y_n) - T_n(Y_1 \dots Y_n)\right\}.$$

Bajo estas consideraciones, se pueden enunciar los siguientes resultados:

Teorema 1. *Sea $Y_1 \dots Y_n$ una muestra aleatoria, sea $T_n = T_n(Y_1 \dots Y_n)$ un estadístico y $T_{(i)} = T_{n-1}(Y_1 \dots Y_{i-1}\, Y_{i+1} \dots Y_n)$, y supóngase que $E[T_n] = E[T_{(i)}]$.*

1. *La curva de sensibilidad, asociada a la muestra $Y_1 \dots Y_{i-1}\, Y_{i+1} \dots Y_n$, verifica*

$$\frac{1}{n} E\left[CS_{n-1}(y_i; T_n)\right] = \mathcal{S}(y_i; T_n). \tag{4}$$

2. *La curva de influencia muestral, asociada a la muestra $Y_1 \dots Y_n$, verifica*

$$\frac{1}{n-1} E\left[CIM_i(T_n) \,|Y_i = y_i\right] = \mathcal{S}(y_i; T_n). \tag{5}$$

La igualdad (4), además de la relación entre los dos conceptos, viene a profundizar en el interés del s.c. como herramienta útil para el análisis de influencia, y la igualdad (5)

puede para fundamentar teóricamente el concepto de curva de influencia muestral como esquema de comparación válido para el análisis de influencia.

Por ello, Muñoz-Pichardo *y otros* (1995) proponen el siguiente estimador insesgado del s.c.

$$\widehat{\mathcal{S}}(y_i;T) = T_n\,(Y_1 \ldots Y_{i-1}, y_i, Y_{i+1} \ldots Y_n) - T_{n-1}(Y_1 \ldots Y_{i-1}\, Y_{i+1} \ldots Y_n) = T_n - T_{(i)},$$

y en general,

$$\widehat{\mathcal{S}}(y_{i_1} \ldots y_{i_m}; T_n) = T_n - T_{(i_1,\ldots,i_m)},$$

que puede denominarse *sesgo condicionado estimado* (s.c.e.). Conviene hacer notar que en muchas ocasiones para el cálculo del s.c. y su estimador no es necesario conocer la distribución subyacente al modelo, algo que le ocurre a otras técnicas de diagnóstico de influencia como el desplazamiento de verosimilitud o la distancia de Rao.

Muñoz-Pichardo *y otros* (1998) aplican este concepto al Análisis de Influencia Local, en particular en el Modelo Lineal General. Un planteamiento genérico de dicha aplicación, siguiendo las pautas marcadas por Cook en su descripción del problema de la influencia, puede ser el siguiente:

> Sea un modelo M postulado a priori, sea $Y_1 \ldots Y_n$ una muestra aleatoria de la(s) variable(s) que intervienen en el mismo, sea un estadístico $T_n(Y_1 \ldots Y_n; M)$, $\mathbf{w}$ un vector perteneciente a un conjunto Ω de perturbaciones relevantes, siendo $M(\mathbf{w})$ el modelo perturbado, de forma que
>
> $$\exists \mathbf{w}_0 \in \Omega \;/\; M \approx M(\mathbf{w}_0) \;\text{ y }\; T_n(Y_1 \ldots Y_n; M) \approx T_n(Y_1 \ldots Y_n; M(\mathbf{w}_0)),$$
>
> y sea y_i una realización muestral de la i-ésima componente de la muestra aleatoria. El estudio de la función de $\mathbf{w} \in \Omega$:
>
> $$\mathcal{S}_w\,(y_i; T_n) = E_{M(\mathbf{w})}\,[T_n \;|Y_i = y_i\,] - E_{M(\mathbf{w})}\,[T_n\,],$$
>
> en un entorno de $\mathbf{w}_0$ permitirá realizar el análisis de influencia local que la realización muestral $Y_i = y_i$ ejerce sobre T_n, siendo $E_{M(\mathbf{w})}\,[-]$ la esperanza en el modelo perturbado $M(\mathbf{w})$.

Para este planteamiento general, son válidas las consideraciones realizadas anteriormente sobre el concepto de s.c.. En particular, en ocasiones se tendrá que buscar un estimador, pudiéndose considerar el estimador insesgado:

$$\widehat{\mathcal{S}}_w\,(y_i; T_n) = T_n\,(Y_1 \ldots Y_{i-1}, y_i, Y_{i+1} \ldots Y_n; M(\mathbf{w})) - T_{n-1}(Y_1 \ldots Y_{i-1}\, Y_{i+1} \ldots Y_n; M(\mathbf{w})).$$

Para ilustrar la utilidad de lo expresado hasta ahora, a continuación, se recoge una aplicación simple: el s.c. y diagnósticos de influencia sobre el estadístico media muestral.

2.1. Una aplicación ilustrativa: estadístico media muestral

Sea $Y_1 \dots Y_n$ una muestra aleatoria de una v.a. Y, con $E[Y] = \mu$ y $Var[Y] = \sigma^2$, y consideremos el estimador BLUE de μ, el estadístico media muestral $T_n = \overline{Y}$. Puede obtenerse fácilmente el s.c. de dicho estadístico dada la i-ésima realización muestral y_i:

$$\mathcal{S}\left(y_i;\overline{Y}\right) = \frac{1}{n}\left(y_i - \mu\right).$$

Por tanto, como lógicamente se podía intuir, la influencia de una observación será mayor cuanto más diste de la media poblacional. Análogamente,

$$\mathcal{S}\left(y_{i_1} \dots y_{i_m};\overline{Y}\right) = \frac{1}{n}\sum_{j=1}^{m}\left(y_{i_j} - \mu\right).$$

En ambas expresiones se observa su dependencia del parámetro μ, por lo que es necesario obtener las correspondientes estimaciones:

$$\widehat{\mathcal{S}}\left(y_i;\overline{Y}\right) = \frac{1}{n-1}\left(y_i - \overline{Y}\right) \tag{6}$$

y

$$\widehat{\mathcal{S}}\left(y_{i_1} \dots y_{i_m};\overline{Y}\right) = \frac{1}{n-m}\sum_{j=1}^{m}\left(y_{i_j} - \overline{Y}\right), \tag{7}$$

respectivamente. Para evitar el problema del signo, a efectos prácticos, como medidas de influencia pueden considerarse los cuadrados de las expresiones (6) y (7):

$$S_i^2 = \left(\frac{y_i - \overline{Y}}{n-1}\right)^2 \quad , \quad S_{i_1 \dots i_m}^2 = \left\{\frac{1}{n-m}\sum_{j=1}^{m}\left(y_{i_j} - \overline{Y}\right)\right\}^2.$$

Con el objetivo de estudiar la influencia local sobre el BLUE de μ, se puede considerar el modelo perturbado determinado por:

$$\begin{aligned} E[Y_k] &= \mu, \quad \forall k \\ Var[Y_k] &= \sigma^2, \quad \forall k \neq i \,;\, Var[Y_i] = \sigma^2/w, \quad w > 0. \end{aligned} \tag{8}$$

El BLUE de μ en el modelo especificado en (8) viene dado por:

$$\overline{Y}_w = \frac{1}{n-1+w}\sum_{k \neq i} Y_k + \frac{w}{n-1+w} Y_i \,,$$

y por tanto,

$$\mathcal{S}_w^{(i)}\left(y_i;\overline{Y}_w\right) = \frac{w}{n-1+w}\left(y_i - \mu\right) = \frac{n\,w}{n-1+w}\mathcal{S}\left(y_i;\overline{Y}\right).$$

(El superíndice (i) se utiliza para indicar que sólo la *i*-ésima observación es perturbada).

Se puede observar que el s.c. en el modelo perturbado es proporcional al s.c. en el modelo postulado, siendo la razón de proporcionalidad:

$$\alpha(w) = \frac{n\,w}{n-1+w}\,.$$

En cuanto al estimador del s.c., se obtiene:

$$\widehat{\mathcal{S}}_w^{(i)}\left(y_i;\overline{Y}_w\right) = \frac{1}{n-1}\frac{n\,w}{n-1+w}\left(y_i-\overline{Y}\right) = \frac{n\,w}{n-1+w}\widehat{\mathcal{S}}\left(y_i;\overline{Y}\right)\,. \tag{9}$$

Considerando nuevamente su cuadrado para evitar el signo, se puede utilizar con medida de influencia local la función de *w*:

$$S_i^2(w) = S_i^2\,\left\{\alpha\left(w\right)\right\}^2.$$

La siguiente igualdad redunda en la relevancia de la razón de proporcionalidad $\alpha(w)$:

$$\widehat{\mathcal{S}}_w^{(i)}\left(y_j;\overline{Y}_w\right) = \alpha(w)\widehat{\mathcal{S}}\left(y_j;\overline{Y}\right) + (1-\alpha(w))\widehat{\mathcal{S}}\left(y_j;\overline{Y}_{(i)}\right), \quad j = 1\ldots n. \tag{10}$$

Para $w \in (0,1)$, dicha razón verifica $0 < \alpha(w) < 1$, y en consecuencia, $\widehat{\mathcal{S}}_w^{(i)}\left(y_j;\overline{Y}_w\right)$ es combinación lineal convexa entre $\widehat{\mathcal{S}}\left(y_j;\overline{Y}\right)$ y$\widehat{\mathcal{S}}\left(y_j;\overline{Y}_{(i)}\right)$, donde los coeficientes no dependen de las observaciones *i*-ésima ni *j*-ésima, tan sólo del tamaño muestral y del factor de perturbación considerado.

Dado el papel relevante que juega esta razón de proporcionalidad, Muñoz-Pichardo *y otros* (1998) la denominan **potencial de influencia local de la *i*-observación**, interpretándola como la proporción de influencia de cualquier observación sobre el BLUE de μ en el modelo perturbado (8), explicada por la influencia que ejerce dicha observación sobre el BLUE de μ en el modelo postulado.

3. SESGO CONDICIONADO EN EL MODELO LINEAL GENERAL MULTIVARIANTE

En esta sección estudiamos el s.c. sobre el Modelo Lineal General Multivariante (MLGM), con objeto de tratar de forma unificada el análisis de influencia en los diversos modelos que se obtienen como casos particulares de él: Modelos de Regresión (múltiple y multivariante), Modelos de Análisis de la Varianza (univariante y multivariante), Modelos de Análisis de la Covarianza (univariante y multivariante), Análisis de Perfiles, Análisis de Medidas Repetidas, etc.

El modelo MLGM está definido por la igualdad matricial $\mathbf{Y} = \mathbf{XB} + \mathcal{E}$, donde $\mathbf{Y}$ es la matriz $(n \times q)$ de respuestas, $\mathbf{X}$ es una matriz $(n \times p)$ conocida, con rango $r (r \leq p \leq n)$, $\mathbf{B}$ es una matriz $(p \times q)$ de parámetros desconocidos y $\mathcal{E} = (\varepsilon_{ik})$ es la matriz de perturbaciones aleatorias que verifica:

$$\begin{aligned} E(\varepsilon_{ik}) &= 0, & 1 \leq i \leq n, 1 \leq k \leq q, \\ Cov(\varepsilon_{ik}, \varepsilon_{js}) &= \sigma_{ks}\, \delta_{ij}, & 1 \leq i, j \leq n, 1 \leq k, s \leq q, \end{aligned}$$

siendo δ_{ij} la delta de Kronecker.

Si $\mathbf{\Lambda B}$ es una función linealmente estimable (f.l.e.), siendo $\mathbf{\Lambda}$ una matriz de dimensiones $d \times q$, de rango d, y $\widehat{\mathbf{B}}$ cualquier estimador de mínimos cuadrados, Muñoz-Pichardo *y otros* (2000) obtienen el s.c. del BLUE, $\mathbf{\Lambda}\widehat{\mathbf{B}}$, asociado a la i-ésima observación de las respuestas $\mathbf{y}_i$,

$$\mathcal{S}(\mathbf{y}_i; \mathbf{\Lambda}\widehat{\mathbf{B}}) = \mathbf{\Lambda S}^{-}\mathbf{x}_i \left(\mathbf{y}_i' - \mathbf{x}_i'\mathbf{B}\right),$$

donde $\mathbf{x}_i'$ e $\mathbf{y}_i'$ son las i-ésimas filas de $\mathbf{X}$ e $\mathbf{Y}$, respectivamente, $\mathbf{S} = \mathbf{X}'\mathbf{X}$ y $\mathbf{S}^{-}$ es una inversa generalizada de $\mathbf{S}$. El s.c.e. puede expresarse en los siguientes términos:

$$\widehat{\mathcal{S}}(\mathbf{y}_i; \mathbf{\Lambda}\widehat{\mathbf{B}}) = \frac{1}{1 - v_{ii}} \mathbf{\Lambda S}^{-}\mathbf{x}_i \mathbf{e}_i' , \tag{11}$$

donde $\mathbf{e}_i'$ es la i-ésima fila de la matriz de residuos $\mathbf{E} = \mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}} = (\mathbf{I} - \mathbf{V})\mathbf{Y}$, siendo v_{ii} el elemento i-ésimo diagonal de la matriz de predicción $\mathbf{V} = \mathbf{X}'\mathbf{S}^{-}\mathbf{X}$.

En el modelo univariante ($q = 1$), $\underline{Y} = \mathbf{X}\beta + \varepsilon$, (MLG), donde $E[\varepsilon] = \mathbf{0}$, $Var[\varepsilon] = \sigma^2 \mathbf{I}_n$, si $\boldsymbol{\lambda}'\beta$ es una f.l.e., el s.c.e. de $\boldsymbol{\lambda}'\widehat{\beta}$, dada la i-ésima observación, viene dado por (Muñoz-Pichardo *y otros* (1995))

$$\widehat{\mathcal{S}}(y_i; \boldsymbol{\lambda}'\widehat{\beta}) = \frac{1}{1 - v_{ii}} \boldsymbol{\lambda}'\mathbf{S}^{-}\mathbf{x}_i (y_i - \mathbf{x}_i'\widehat{\beta}).$$

Con objeto de analizar la precisión de dicho estimador, a continuación se obtiene su varianza y una estimación de la misma.

Teorema 2. *En el MLG, se verifica*

$$Var\left[\widehat{\mathcal{S}}(y_i; \boldsymbol{\lambda}'\widehat{\beta}) \,|Y_i = y_i\right] = \frac{v_{ii}}{1 - v_{ii}} (\boldsymbol{\lambda}'\mathbf{S}^{-}\mathbf{x}_i)^2 \sigma^2 .$$

La igualdad recogida en el teorema se obtiene de forma directa, dado que

$$\begin{aligned} Var\left[\widehat{\mathcal{S}}(y_i; \boldsymbol{\lambda}'\widehat{\beta}) \,|Y_i = y_i\right] &= \frac{\left(\boldsymbol{\lambda}'\mathbf{S}^{-}\mathbf{x}_i\right)^2}{(1 - v_{ii})^2} Var\left[\mathbf{x}_i'\widehat{\beta} \,|Y_i = y_i\right] \\ &= \frac{\left(\boldsymbol{\lambda}'\mathbf{S}^{-}\mathbf{x}_i\right)^2}{(1 - v_{ii})^2} Var\left[\mathbf{x}_i'\mathbf{S}^{-}\mathbf{X}_{(i)}'\underline{Y}_{(i)}\right] . \end{aligned}$$

En consecuencia, en el MLG un estimador de la varianza del s.c.e. viene dado por

$$\widehat{Var}\left[\widehat{\mathcal{S}}(y_i;\boldsymbol{\lambda}'\widehat{\beta})\,|Y_i=y_i\right]=\frac{v_{ii}}{(1-v_{ii})^2}(\boldsymbol{\lambda}'\mathbf{S}^{-}\mathbf{x}_i)^2\widehat{\sigma}^2_{(i)},$$

donde $\widehat{\sigma}^2_{(i)}$ es el estimador insesgado de σ^2 en el modelo bajo la omisión del *i*-ésimo caso.

3.1. Diagnósticos de influencia en el MLGM

En el modelo multivariante, la dimensión de la expresión (11) es $d\times q$. Por tanto, para cuantificar la influencia que la *i*-ésima observación ejerce sobre el BLUE de la f.l.e., $\boldsymbol{\Lambda}\widehat{\mathbf{B}}$, se ha de considerar una norma matricial.

Muñoz *y otros* (2000), como generalización de la norma propuesta por Cook y Weisberg (1982) y Belsley *y otros* (1980), proponen la siguiente: dada una matriz $\mathbf{A}$, de dimensiones $(d\times q)$,

$$\|\mathbf{A}\|_{(\mathbf{Q},\mathbf{C})}=\left[tr(\mathbf{A}'\mathbf{Q}\mathbf{A}\mathbf{C}^{-1})\right]^{1/2}, \tag{12}$$

donde $\mathbf{Q}$ y $\mathbf{C}$ son matrices simétricas d.p. de dimensiones $(q\times q)$ y $(d\times d)$, respectivamente.

Como casos particulares, se proponen las siguientes distancias como diagnósticos de influencia:

- D_i-distancia asociada a la *i*-ésima observación: $\mathbf{Q}=(\boldsymbol{\Lambda}\mathbf{S}^{-}\boldsymbol{\Lambda}')^{-1}$ y $\mathbf{C}=d\widehat{\boldsymbol{\Sigma}}$ (donde $\widehat{\boldsymbol{\Sigma}}=\frac{1}{n-r}\mathbf{E}'\mathbf{E}$ es un estimador insesgado de $\boldsymbol{\Sigma}=(\sigma_{ks})$),

  $$D_i(\boldsymbol{\Lambda}\widehat{\mathbf{B}})=\left\|\widehat{\mathcal{S}}(\mathbf{y}_i;\boldsymbol{\Lambda}\widehat{\mathbf{B}})\right\|^2_{((\boldsymbol{\Lambda}\mathbf{S}^{-}\boldsymbol{\Lambda}')^{-1},d\widehat{\boldsymbol{\Sigma}})}.$$

  Esta distancia debe considerarse como una generalización de la distancia de Cook (Cook y Weisberg (1982)), propuesta para el Modelo de Regresión Múltiple, posteriormente extendida al Modelo de Regresión Multivariante por Hossain y Naik (1989) y por Barret y Ling (1992).

- W_i-distancia asociada a la *i*-ésima observación: $\mathbf{Q}=(\boldsymbol{\Lambda}\mathbf{S}^{-}\boldsymbol{\Lambda}')^{-1}$ y $\mathbf{C}=\widehat{\boldsymbol{\Sigma}}_{(i)}$ (estimador insesgado de $\boldsymbol{\Sigma}$ en el modelo bajo la omisión de la *i*-ésima observación),

  $$W_i(\boldsymbol{\Lambda}\widehat{\mathbf{B}})=\left\|\widehat{\mathcal{S}}(\mathbf{y}_i;\boldsymbol{\Lambda}\widehat{\mathbf{B}})\right\|^2_{((\boldsymbol{\Lambda}\mathbf{S}^{-}\boldsymbol{\Lambda}')^{-1},\widehat{\boldsymbol{\Sigma}}_{(i)})}.$$

  Análogamente, esta distancia es una generalización de distancia de Welsch-Kuh (Belsley *y otros* (1980)), propuesta también en el modelo de regresión múltiple y posteriormente extendida al modelo de regresión multivariante por Hossain y Naik (1989).

- C_i-distancia asociada a la i-ésima observación: $\mathbf{Q} = (\mathbf{\Lambda S}^{-}\mathbf{\Lambda}')^{-1}$ y $\mathbf{C} = \frac{r}{n-r}\widehat{\mathbf{\Sigma}}_{(i)}$,

$$C_i(\mathbf{\Lambda}\widehat{\mathbf{B}}) = \left\| \widehat{\mathcal{S}}(\mathbf{y}_i; \mathbf{\Lambda}\widehat{\mathbf{B}}) \right\|^2_{((\mathbf{\Lambda S}^{-}\mathbf{\Lambda}')^{-1}, \frac{r}{n-r}\widehat{\mathbf{\Sigma}}_{(i)})}.$$

  Esta medida de influencia es una generalización de la distancia modificada de Cook, propuesta por Atkinson (1981) para el modelo de regresión múltiple.

La generalización de estas medidas para el estudio del análisis de influencia conjunta de una colección de observaciones es simple y directa.

3.2. Influencia Local en el MLGM

Sobre este modelo también se puede utilizar el s.c. para el análisis de influencia local. Muñoz-Pichardo *y otros* (1998) proponen medidas de influencia local en el modelo univariante ($q = 1$). Posteriormente Moreno-Rebollo *y otros* (2000) extienden dichos resultados al modelo multivariante.

Se considera el modelo perturbado, MLGM(i, w), bajo el esquema de ponderación de casos, en el que sólo el caso bajo estudio es ponderado con un peso $w > 0$, es decir,

$$Cov(\underline{\varepsilon}_j) = \begin{cases} \mathbf{\Sigma} & j = 1, \ldots, n;\ j \neq i \\ w\mathbf{\Sigma} & j = i, \end{cases}$$

donde $\underline{\varepsilon}'_j$ es la j-ésima fila de $\mathcal{E}$.

En MLGM(i, w), el s.c. del BLUE $\mathbf{\Lambda}\widehat{\mathbf{B}}_w$ de una f.l.e. viene dado por:

$$(13)\quad \mathcal{S}_w^{(i)}\left(\mathbf{y}_i; \mathbf{\Lambda}\widehat{\mathbf{B}}_w\right) = \frac{w}{1+(w-1)v_{ii}}\Lambda\mathbf{S}^{-}\mathbf{x}_i\left(\mathbf{y}_i - \mathbf{x}_i'\mathbf{B}\right) = \frac{w}{1+(w-1)v_{ii}}\mathcal{S}\left(\mathbf{y}_i; \mathbf{\Lambda}\widehat{\mathbf{B}}\right).$$

En (13) puede observarse la proporcionalidad entre el s.c. del BLUE de una f.l.e. en MLGM y en MLGM(i, w), de forma semejante a (9). En este caso, la razón de proporcionalidad viene dada por

$$\alpha_i^*(w) = \frac{w}{1+(w-1)v_{ii}},$$

que no depende de la f.l.e., sólo del elemento i-ésimo diagonal de la matriz de predicción.

Análogamente, el s.c.e. de $\mathbf{\Lambda}\widehat{\mathbf{B}}_w$ dada la i-ésima observacion y_i viene dado por:

$$\widehat{\mathcal{S}}_w^{(i)}\left(\mathbf{y}_i; \mathbf{\Lambda}\widehat{\mathbf{B}}_w\right) = \alpha_i^*(w)\,\widehat{\mathcal{S}}\left(\mathbf{y}_i; \mathbf{\Lambda}\widehat{\beta}\right).$$

Dado que estas funciones de w son vectoriales de dimensión d, Moreno-Rebollo *y otros* (2000) proponen las normas ya utilizadas en el modelo MLGM, y recogidas anteriormente, obteniéndose las siguientes igualdades:

$$\begin{aligned} D_i(w,\boldsymbol{\Lambda}\widehat{\mathbf{B}}) &= [\alpha_i^*(w)]^2 D_i(\boldsymbol{\Lambda}\widehat{\mathbf{B}}), \\ W_i(w,\boldsymbol{\Lambda}\widehat{\mathbf{B}}) &= [\alpha_i^*(w)]^2 W_i(\boldsymbol{\Lambda}\widehat{\mathbf{B}}), \\ C_i(w,\boldsymbol{\Lambda}\widehat{\mathbf{B}}) &= [\alpha_i^*(w)]^2 C_i(\boldsymbol{\Lambda}\widehat{\mathbf{B}}). \end{aligned}$$

En las tres expresiones anteriores se observa la proporcionalidad entre la medida de influencia local y la medida de influencia, con razón de proporcionalidad, función de w, idéntica para las tres, el cuadrado de la razón de proporcionalidad que relaciona el s.c. (y su estimación) en MLGM y MLGM(i,w). El análisis de estas funciones en un entorno de $w=1$ permitirá realizar el análisis de influencia local de la observación bajo estudio.

Finalmente, puede obtenerse una expresión semejante a (10),

$$\mathcal{S}_w^{(i)}\left(\mathbf{y}_j;\boldsymbol{\Lambda}\widehat{\mathbf{B}}_w\right) = \alpha_i^*(w)\mathcal{S}\left(\mathbf{y}_j;\boldsymbol{\Lambda}\widehat{\mathbf{B}}\right) + (1-\alpha_i^*(w))\mathcal{S}\left(\mathbf{y}_j;\boldsymbol{\Lambda}\widehat{\mathbf{B}}_{(i)}\right), \forall w>0 \text{ y } \quad j=1,\ldots,n.$$

Para $w\in(0,1)$, $\mathcal{S}_w^{(i)}\left(\mathbf{y}_j;\boldsymbol{\Lambda}\widehat{\mathbf{B}}_w\right)$ es una combinación lineal convexa entre $\mathcal{S}\left(\mathbf{y}_j;\boldsymbol{\Lambda}\widehat{\mathbf{B}}\right)$ y $\mathcal{S}\left(\mathbf{y}_j;\boldsymbol{\Lambda}\widehat{\mathbf{B}}_{(i)}\right)$. Sus coeficientes no dependen de la j-ésima observacion ni de la f.l.e. $\boldsymbol{\Lambda\beta}$. Así, $\alpha_i^*(w)$ es la proporción de $\mathcal{S}_w^{(i)}\left(\mathbf{y}_j;\boldsymbol{\Lambda}\widehat{\mathbf{B}}_w\right)$ en la dirección de $\mathcal{S}\left(\mathbf{y}_j;\boldsymbol{\Lambda}\widehat{\mathbf{B}}\right)$ para cualquier j. Es decir, $\alpha_i^*(w)$ es la proporción de $\mathcal{S}_w^{(i)}\left(\mathbf{y}_j;\boldsymbol{\Lambda}\widehat{\mathbf{B}}_w\right)$ explicada por $\mathcal{S}\left(\mathbf{y}_j;\boldsymbol{\Lambda}\widehat{\mathbf{B}}\right)$. Análogamente, $[1-\alpha_i^*(w)]$ es la proporción en la dirección de $\mathcal{S}\left(\mathbf{y}_j;\boldsymbol{\Lambda}\widehat{\mathbf{B}}_{(i)}\right)$. En consecuencia, podemos interpretar $\alpha_i^*(w)$ como la proporción de influencia de cualquier observación sobre el BLUE de cualquier f.l.e. en el modelo MLGM(w,i) explicada por la influencia que dicha observación ejerce sobre el BLUE de la f.l.e. en el modelo MLGM. De forma semejante puede interpretarse la razón $[1-\alpha_i^*(w)]$. Como la razón de proporcionalidad entre las distancias son los cuadrados de $\alpha_i^*(w)$ y $[1-\alpha_i^*(w)]$, Muñoz *y otros* (1998) definen $[\alpha_i^*(w)]^2$ como el Potencial de Influencia Local (*LIP*) de la i-ésima observación:

$$LIP_i(w) = \left(\frac{w}{1+(w-1)v_{ii}}\right)^2.$$

4. SESGO CONDICIONADO EN COMPONENTES PRINCIPALES

Dada la importancia de esta técnica estadística en distintas áreas de investigación, se han propuesto métodos para abordar el problema de la influencia en el Análisis de

Componentes Principales (ACP), basados en el esquema de perturbación de la omisión de observaciones y la curva de influencia muestral (Critchley (1985)).

También este problema puede abordarse a través del s.c.. Enguix-González *y otros* (2000) proponen medidas de influencia para las estimaciones de los autovalores y autovectores de la matriz de varianzas y covarianzas muestrales, resultados que se recogen a continuación.

Dado un vector aleatorio p-dimensional $X \sim N_p(\mu, \mathbf{\Sigma})$, y una muestra aleatoria $X_1 \dots X_n$, sea

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X})(X_j - \overline{X})',$$

la matriz de varianzas y covarianzas muestrales, y denotemos por $\widehat{\lambda}_k$, $\widehat{\alpha}_k$ ($k = 1 \dots p$), a los autovalores y autovectores asociados, respectivamente.

Para una realización muestral $\mathbf{x}_1 \dots \mathbf{x}_n$, se obtienen los siguientes resultados

$$\begin{aligned} S(\mathbf{x}_i, \widehat{\lambda}_k) &= \frac{1}{n-1} I(\mathbf{x}_i, \lambda_k) + O(n^{-2}), \\ S(\mathbf{x}_i, \widehat{\alpha}_k) &= \frac{1}{n-1} I(\mathbf{x}_i, \alpha_k) + O(n^{-2}), \end{aligned}$$

donde λ_k y α_k son los autovalores y autovectores asociados de la matriz $\mathbf{\Sigma}$, siendo $I(\mathbf{x}, \lambda_k)$ e $I(\mathbf{x}, \alpha_k)$ las funciones de influencia de dichos parámetros,

$$\begin{aligned} I(\mathbf{x}, \lambda_k) &= \left\{ \alpha_k'(\mathbf{x} - \mu) \right\}^2 - \lambda_k, \\ I(\mathbf{x}, \alpha_k) &= -\alpha_k'(\mathbf{x} - \mu) \sum_{j \neq k} \frac{\alpha_j'(\mathbf{x} - \mu)}{\lambda_j - \lambda_k} \alpha_j. \end{aligned}$$

En base a (5), se propone los siguientes estimadores de $S(\mathbf{x}_i, \widehat{\lambda}_k)$ y $S(\mathbf{x}_i, \widehat{\alpha}_k)$,

$$\widehat{S}(\mathbf{x}_i, \widehat{\lambda}_k) = \widehat{\lambda}_k - \widehat{\lambda}_k^{(i)},$$

y

$$\widehat{S}(\mathbf{x}_i, \widehat{\alpha}_k) = \widehat{\alpha}_k - \widehat{\alpha}_k^{(i)},$$

donde $\widehat{\lambda}_k^{(i)}$ y $\widehat{\alpha}_k^{(i)}$ son los autovalores y autovectores asociados a $\widehat{\mathbf{\Sigma}}_{(i)}$. Ambas estimaciones pueden considerarse como estadísticos de diagnóstico de influencia en el ACP. Para evitar el problema del signo del s.c.e. en el caso de los autovalores, y el problema de la dimensión del s.c.e. en el caso de los autovectores, se pueden aplicar distancias, de forma análoga a lo realizado anteriormente en el MLGM. Así, se proponen:

- Para un autovalor, el cuadrado del s.c.e., normalizado con la estimación de $Var\left[\widehat{\lambda}_k\right]$,

$$(n-1)\frac{\left\{\widehat{S}(\mathbf{x}_i,\widehat{\lambda}_k)\right\}^2}{2\widehat{\lambda}_k^2}.$$

- Para un autovector, se pueden aplicar las normas anteriormente definidas de acuerdo a (12). En particular, dado que en este caso $d=1$, para $\mathbf{Q}=\mathbf{I}_p$ y $\mathbf{C}$ la unidad, se obtendría la norma euclídea del vector $\widehat{S}(\mathbf{x}_i,\widehat{\alpha}_k)$.

Este enfoque del estudio de la influencia permite generalizar los estadísticos de diagnóstico anteriores para un conjunto de autovalores, o un conjunto de autovectores, seleccionando adecuadamente las matrices que determinan la norma.

5. SESGO CONDICIONADO EN EL MUESTREO EN POBLACIONES FINITAS

Moreno-Rebollo *y otros* (1999) adaptan el concepto de s.c. como diagnóstico de influencia en el muestreo en poblaciones finitas. En particular, lo desarrollan para el estimador de Horvitz-Thompson del total poblacional.

Sea $U=\{u_1,\ldots,u_N\}$ una población finita y $\{\mathcal{M},p(.)\}$ un diseño muestral definido sobre U, donde $\mathcal{M}$ es el espacio muestral y $p(.)$ una distribución de probabilidad sobre $\mathcal{M}$. Sea Y una característica de la población, $Y=\{Y_1,\ldots,Y_N\}$, $\theta=\theta(Y)$ el parámetro de interés y $\widehat{\theta}=\widehat{\theta}(s)$ un estimador de θ basado en $s\in\mathcal{M}$. Se propone la siguiente definición.

Definición 3. *El sesgo condicionado sobre* $\widehat{\theta}$*, causado por la presencia de* u_i *en la muestra, se define por*

$$S(I_i=1;\widehat{\theta})=E[\widehat{\theta}\,|I_i=1]-E[\widehat{\theta}]\,,$$

donde $I_i(s)$ *(*$i=1\ldots N$*) son las variables aleatorias*

$$I_i(s)=\begin{cases}1 & si\ u_i\in s\\ 0 & en\ otro\ caso.\end{cases}$$

Es decir, el s.c. mide la desviación en el valor esperado del estimador cuando el diseño muestral se perturba, restringiéndolo sobre las muestras que contienen a u_i.

En particular, si se considera el estimador de Horvitz-Thompson, $\widehat{T}_{HT} = \sum_s \frac{Y_i}{\pi_i}$, del total poblacional de la característica Y, $T(Y) = \sum_{i=1}^{N} Y_i$, se obtiene que

$$S(I_i = 1; \widehat{T}_{HT}) = Y_i \frac{1-\pi_i}{\pi_i} + \sum_{j \neq i} Y_j \frac{\Delta_{ij}}{\pi_i \pi_j} ,$$

donde $\pi_j = \Pr[I_j = 1]$, $j = 1, \ldots, N$, representan las probabilidades de inclusión de primer orden asociadas al diseño muestral $\{\mathcal{M}, p(.)\}$ y $\Delta_{ij} = Cov(I_i, I_j)$, $i, j = 1 \ldots N$.

Dado que el s.c. es un parámetro poblacional desconocido, se propone como estimador del mismo el estimador de Horvitz-Thompson sobre el diseño muestral restringido sobre las muestras que contienen a u_i:

$$\widehat{S}_{HT}(I_i = 1; \widehat{T}_{HT}) = Y_i \frac{1-\pi_i}{\pi_i} + \sum_{j \neq i} Y_j \frac{\Delta_{ij}}{\pi_{ij} \pi_j} I_i ,$$

donde $\pi_{ij} = \Pr[I_i = 1, I_j = 1]$, son las probabilidades de inclusión de segundo orden en el diseño muestral.

Un estimador insesgado de la varianza de $\widehat{S}_{HT}(I_i = 1; \widehat{T}_{HT})$ en el diseño restringido viene dado por:

$$\widehat{Var}\left[\widehat{S}_{HT}(I_i = 1; \widehat{T}_{HT})\right] = \sum_{j \neq i} \sum_{h \neq i} Y_j Y_h \frac{\Delta_{ij}\, \Delta_{ih}}{\pi_i \pi_j \pi_h} \left[\frac{\pi_i}{\pi_{ij}\, \pi_{ih}} - \frac{1}{\pi_{ijh}}\right] I_j I_h ,$$

donde $\pi_{ijh} = \Pr[I_i = 1, I_j = 1, I_h = 1]$, son las probabilidades de inclusión de tercer orden en $\{\mathcal{M}, p(.)\}$.

Es evidente que tanto el s.c., como el s.c.e. y el estimador de su varianza dependen del diseño muestral $\{\mathcal{M}, p(.)\}$, por lo que la influencia de cada unidad muestral seleccionada es función del comportamiento de dicha unidad respecto a la característica bajo estudio y del diseño muestral elegido por el investigador. La aplicación a cada tipo de diseño es directa, sin más que determinar los parámetros dependientes del mismo (probabilidades y covarianzas de las variables aleatorias indicadores), aunque en ocasiones pueda resultar compleja. No obstante, frente a la posible complejidad, el concepto de s.c. posee la característica de su generalidad. El planteamiento arriba recogido puede utilizarse en un amplio espectro del muestreo en poblaciones finitas, quedando abiertas un número considerable de líneas de investigación y desarrollo.

6. CONCLUSIÓN

Este trabajo recoge una revisión del concepto de s.c. y su aplicación al análisis de influencia en diversas técnicas estadísticas. Su definición genérica, su fácil interpretación y su cálculo no excesivamente complejo le permite tener un extenso campo de

aplicación y un amplio abanico de posibilidades en el análisis de influencia y análisis de influencia local de cualquier estadístico en cualquier modelo. Muestra de tales afirmaciones son los resultados anteriormente recogidos y los trabajos de Jiménez (1994) y Jiménez *y otros* (1995) en el área de las técnicas de remuestreo, en particular en el bootstrap, para detectar muestras bootstrap con un efecto considerable sobre los resultados de las estimaciones.

Las medidas de diagnóstico de influencia propuestas pueden generalizarse fácilmente al análisis de influencia conjunto de dos o más observaciones. Algunos resultados sobre tal aspecto están recogidos en los trabajos citados anteriormente.

Profundizando por la línea de los modelos lineales se puede abordar la influencia en los modelos lineales generalizados (univariantes y multivariantes), análisis de supervivencia, etc. En la línea del ACP, pueden obtenerse resultados de interés en Análisis Discriminante, Análisis Factorial, Análisis Canónico, etc. Finalmente, en el área del muestreo en poblaciones finitas, como se recoge en el apartado anterior, queda un amplio campo abierto para seguir desarrollando técnicas de diagnóstico de influencia, más aún cuando en este área la problemática de la influencia no ha sido, hasta ahora, tratada en profundidad.

REFERENCIAS

Andrews, D. F. y Pregibon, D. (1978). «Finding outliers that matter». *J. Royal Statistics Soc., Ser. B*, 40, 85-93.

Atkinson, A. C. (1981). «Two graphical displays for outliying and influential observations in regression». *Biometrika*, 68, 13-20.

Barnett, V. y Lewis, T. (1994). *Outliers in Statistical Data.* John Wiley and Sons.

Barret, B. E. y Ling, R. F. (1992). «General classes of influence measures for multivariate regression». *J.A.S.A.*, (7), 184-191.

Belsley, D. A., Kuh, E. y Welsch, R. E. (1980). *Regresion diagnostics: Identifying influential data and sources of collinearity.* John Wiley and Sons.

Billor, N. y Loynes, R. M. (1993). «Local influence: a new approach». *Commun. Statist. Theory and Methods*, 22(6), 1595-1611.

Brown, G. C. y Lawrance, A. J. (2000). «Theory and illustration of regression influence diagnostics». *Commun. Statist. Theory and Methods*, 29 (9 & 10), 2079-2107.

Cook, R. D. y Weisberg, S. (1982). *Residuals and influence in regresion*, Chapman and Hall.

—— (1986). «Assessment of Local Influence (with discussion)». *J. Royal Statistics Soc., Ser. B*, 48, 133-169.

—— (1987). «Influence assessment». *Journal of Applied Statistics*, 14, 2, 117-132.

Cook, D., Peña, D. y Weisberg, S. (1988). «The likelihood displacement: a unifying principle for influence measures». *Commun. Statist.-Theory and Methods*, 17 (3), 623-640.

Critchley, F. (1985). «Influence in Principals Components Analysis». *Biometrika*, 43, 128-136.

Efron, B. y Stein, C. (1984). «The jackknife estimate of variance». *Ann. Statist.*, 9 (3), 586-596.

Enguix-González, A., Moreno-Rebollo, J. L., Jiménez-Gamero, M. D. y Muñoz-Pichardo, J. M. (2000). «Estudio de influencia en componentes principales a través del sesgo condicionado». *Actas XXV Congreso Nac. de Estadística e Investigación Operativa*, (Vigo, España).

Escobar, L. A. y Meeker, W. Q. (1992). «Assessing influence in regression analysis with censored data». *Biometrics*, 48, 507-528.

Hampel, F. R. (1974). «The influence curve and its role in robust estimation». *J. Am. Statist. Assoc.*, 69, 383-393.

Hossain, A. y Naik, D. N. (1989). «Detection of influential observations in multivariate regression». *Journal of Applied Statistics*, 16, 25-37.

Jiménez Gamero, M. D. (1994). *Análisis de las muestras generadas en el proceso de simulación bootstrap.* Tesis Doctoral. Universidad de Sevilla.

Jiménez Gamero, M. D., Muñoz Pichardo, J. M. y Muñoz Reyes, A. (1995). «Medida de influencia en las estimaciones bootstrap». *Actas del XXII Congreso Nac. de Estadística e I.O.*, (Sevilla).

Lawrance, A. J. (1991). «Local and deletion influence». En *Directions in robust statistics and diagnostics*, W. Stahel y S. Weisberg (eds.). Springer-Verlag.

Moreno-Rebollo, J. L., Enguix-González, A., Muñoz-Pichardo, J. M. y Alba, M. V. (2000). «Influencia Local en el Modelo Lineal General Multivariante». *Actas XXV Congreso Nac. de Estadística e Investigación Operativa*, (Vigo, España).

Moreno Rebollo, J. L., Muñoz Reyes, A. M. y Muñoz Pichardo, J. M. (1999). «Influence diagnostic in survey sampling: conditional bias». *Biometrika*, 86, (4), 923-928.

Muñoz Pichardo, J. M. y Fernández Ponce, J. M. (1997). «Distancias de Mahalanobis y Rao: Influencia en el Modelo Lineal General». *Actas IV International Meeting of Multidimensional Data analysis*, (Bilbao, Spain), 259-262.

Muñoz Pichardo, J. M., Muñoz García, J., Fernández Ponce, J. M. y Jiménez Gamero, M. D. (2000). «Influence analysis in multivariate linear general models». *Commun. Statist.-Theory and Methods*, 29 (aceptado para publicación).

Muñoz Pichardo, J. M., Muñoz García, J., Fernández Ponce, J. M. y López Blázquez, F. (1998). «Local Influence on the General Linear Model». *Sankhyã*, Ser. B, 60 (3).

Muñoz Pichardo, J. M., Muñoz García, J., Moreno Rebollo, J. y Pino Mejías, R. (1995). «A new approach to influence analysis in linear models». *Sankhyã*, Ser. A, 57 (3). 393-409.

Tukey, J. W. (1970). *Exploratory Data Analysis*, (1970/71: edición preliminar). Reading Mass. Addison-Wesley.

ENGLISH SUMMARY

CONDITIONAL BIAS IN INFLUENCE ANALYSIS: A REVIEW

J. M. MUÑOZ-PICHARDO
J. L. MORENO-REBOLLO
T. GÓMEZ-GÓMEZ
A. ENGUIX-GONZÁLEZ
Universidad de Sevilla*

Conditional bias has been proposed as an influence diagnostic on different models and statistical techniques. In this paper, we sum up these applications and we relate it to the sensitivity curve and the sample influential curve. Moreover, we point out some areas in which the Influence Analysis could be studied through this approach.

Keywords: Influence analysis, linear models, principal components, survey sampling, conditional bias

AMS Classification (MSC 2000): 62J20, 62H25, 62D05

* Universidad de Sevilla. Facultad de Matemáticas. Departamento de Estadística e Investigación Operativa. Avda. Reina Mercedes s/n. 41012 Sevilla.

– Received October 2000.
– Accepted April 2001.

From the Decomposition Lemma of Efron and Stein (1984), Muñoz - Pichardo *et al.* (1995) proposed the conditional bias (c.b.) as a general approach in Influence Analysis. In this paper, we gather together some applications of this concept in several models and statistical techniques.

Definition 1. *Let $Y_1 \dots Y_n$ be a random sample of a random variable Y, let $T_n = T_n(Y_1 \dots Y_n)$ be a statistic and let $y_1 \dots y_n$ be a sample realization. The c.b. of T_n given y_i is defined as*

$$\mathcal{S}(y_i; T_n) = E[T_n | Y_i = y_i] - E[T_n].$$

From Definition 1, we note that the c.b. depends on the distribution of T_n and on the observed value y_i, and it assesses the influence of y_i on T_n in terms of its expected value. Therefore, it not depends on $y_2, \dots, y_n$. The perturbation considered it is due to the knowledge of the observation under study, y_i. Moreover, if T_n is q-dimensional ($q > 1$), then a norm must be used in order to define an influence measure from $\mathcal{S}(y_i; T_n)$. Finally, we note that, in general, $\mathcal{S}(y_i; T_n)$ depends on unknown parameters, so it must be estimated.

In Section 2 we relate the c.b. to the expected value of the sensitivity curve (Tukey, 1970), and the sample influence curve. From these relations, Muñoz-Pichardo *et al.* (1995) proposed

$$\widehat{\mathcal{S}}(y_i; T) = T_n(Y_1 \dots Y_{i-1}, y_i, Y_{i+1} \dots Y_n) - T_{n-1}(Y_1 \dots Y_{i-1}\, Y_{i+1} \dots Y_n) = T_n - T_{(i)},$$

as estimator of $\mathcal{S}(y_i; T)$.

Following the guidelines laid down by Cook (1987), Muñoz-Pichardo *et al.* (1998) apply the c.b. in Local Influence Analysis through the study of the function

$$\mathcal{S}_w(y_i; T_n) = E_{M(\mathbf{w})}[T_n \,|Y_i = y_i] - E_{M(\mathbf{w})}[T_n],$$

being $M(\mathbf{w})$ a perturbation of the postulated model M.

As an illustration, in section 2.1 we obtain the c.b. and some influence measures on the sample mean. Also, we study the local influence on the sample mean of a random sample.

In section 3, we study the c.b. in the Multivariate General Linear Model (MGLM), in order to handle in an unified way the influence analysis on the models that are obtained as particular cases of it.

Given the MGLM, that is defined by the matrix identity $\mathbf{Y} = \mathbf{XB} + \mathcal{E}$, with the usual hypotheses, and a linear estimable function (l.e.) $\mathbb{A}\mathbf{B}$, it is obtained that (Muñoz-Pichardo *et al.*, 2000)

$$\mathcal{S}(\mathbf{y}_i;\boldsymbol{\Lambda}\widehat{\mathbf{B}}) = \boldsymbol{\Lambda}\mathbf{S}^{-}\mathbf{x}_i\left(\mathbf{y}_i' - \mathbf{x}_i'\mathbf{B}\right), \quad \widehat{\mathcal{S}}(\mathbf{y}_i;\boldsymbol{\Lambda}\widehat{\mathbf{B}}) = \frac{1}{1-v_{ii}}\boldsymbol{\Lambda}\mathbf{S}^{-}\mathbf{x}_i\mathbf{e}_i',$$

being $\widehat{\mathbf{B}}$ a least squares estimator of $\mathbf{B}$, $\mathbf{x}_i'$ and $\mathbf{y}_i'$ the i-th rows of $\mathbf{X}$ and $\mathbf{Y}$, respectively, $\mathbf{S} = \mathbf{X}'\mathbf{X}$, $\mathbf{S}^{-}$ an inverse generalized of $\mathbf{S}$, $\mathbf{e}_i'$ the i-th row of the matrix of ordinary residuals, and v_{ii} the i-diagonal element of the hat matrix $\mathbf{V} = \mathbf{X}'\mathbf{S}^{-}\mathbf{X}$.

In order to assess the influence of y_i on the BLUE, $\boldsymbol{\Lambda}\widehat{\mathbf{B}}$, we apply matrix norms, obtaining various influence diagnostics. The application of these diagnostics in the Multiple Regression Model, make possible to obtain, as particular cases, some of the influence measures defined in the literature. Moreover, we study the Local Influence Analysis in this model.

In section 4, we study the influence analysis in Principal Components, under hypothesis of normality. The c.b. of $\widehat{\lambda}_k$, $\widehat{\alpha}_k$, $k = 1\ldots p$, the eigenvalues and eigenvectors of the sample covariance matrix are given by (Enguix-González *et al.*, 2000) ,

$$S(\mathbf{x}_i,\widehat{\lambda}_k) = \frac{1}{n-1}\,I(\mathbf{x}_i,\lambda_k) + O(n^{-2}) \;\text{ and }\; S(\mathbf{x}_i,\widehat{\alpha}_k) = \frac{1}{n-1}\,I(\mathbf{x}_i,\alpha_k) + O(n^{-2}),$$

being $\mathbf{x}_1\ldots\mathbf{x}_n$ the sample realization and $I(\mathbf{x},\lambda_k)$, $I(\mathbf{x},\alpha_k)$ the influence functions of λ_k and α_k, the population eigenvalues and eigenvectors,

$$I(\mathbf{x},\lambda_k) = \left\{\alpha_k'(\mathbf{x}-\mu)\right\}^2 - \lambda_k \;\text{ and }\; I(\mathbf{x},\alpha_k) = -\alpha_k'(\mathbf{x}-\mu)\sum_{j\neq k}\frac{\alpha_j'(\mathbf{x}-\mu)}{\lambda_j-\lambda_k}\,\alpha_j.$$

From $\widehat{S}(\mathbf{x}_i,\widehat{\lambda}_k)$ and $\widehat{S}(\mathbf{x}_i,\widehat{\alpha}_k)$ several influence measure on $\widehat{\lambda}_k$ and $\widehat{\alpha}_k$ are proposed.

Finally, in section 5, we show that the c.b. can be applied in sampling from a finite population, when the inference is design-based. Moreno-Rebollo *et al. (1999)* adjust the definition of c.b. in order to obtain an influence measure in survey sampling. Let $U = \{u_1,\ldots,u_N\}$ be a finite population and $\{\mathcal{M}, p(.)\}$ a sampling design defined on U, let π_i, $i = 1,\ldots,N$, be the first order inclusion probabilities. Let Y be a characteristic of the population, $Y = \{Y_1,\ldots,Y_N\}$, $\theta = \theta(Y)$ the parameter of interest and $\widehat{\theta} = \widehat{\theta}(s)$ an estimator of θ, for $s \in \mathcal{M}$.

Definition 3. *The conditional bias of* $\widehat{\theta}$, *caused by the presence of* u_i *in the sample* s, *is defined by* $S(I_i = 1;\widehat{\theta}) = E[\widehat{\theta}\,|I_i = 1] - E[\widehat{\theta}]$, *being* $I_i(s) = 1$ *if* $u_i \in s$, $I_i(s) = 0$ *otherwise.*

Particularly, if the Horvitz-Thompson (HT) estimator, $\widehat{T}_{HT} = \sum_s \frac{Y_i}{\pi_i}$, is considered, it is obtained that

$$S(I_i = 1;\widehat{T}_{HT}) = Y_i\frac{1-\pi_i}{\pi_i} + \sum_{j\neq i} Y_j\frac{\Delta_{ij}}{\pi_i\pi_j},$$

being $\Delta_{ij} = Cov(I_i, I_j)$, $i, j = 1 \ldots N$, on $\{\mathcal{M}, p(.)\}$.

Moreno-Rebollo *et al. (1999)* proposed to estimate $S(I_i = 1; \widehat{T}_{HT})$ by the HT-estimator based on the restricted sampling design, on the samples containing u_i,

$$\widehat{S}_{HT}(I_i = 1; \widehat{T}_{HT}) = Y_i \frac{1 - \pi_i}{\pi_i} + \sum_{j \neq i} Y_j \frac{\Delta_{ij}}{\pi_{ij}\pi_j} I_j.$$

We note that both the c.b. and its estimation depend on the sampling design, that is a distinctive feature of the influence measures in sample survey.