Metodología de Encuestas
Vol 2, Num 1, 2000, 103-116

ANÁLISIS DE SENSIBILIDAD EN EL MUESTREO EN POBLACIONES FINITAS

José Luís Moreno
Ana María Muñoz
Joaquín Muñoz
Universidad de Sevilla

RESUMEN

Las conclusiones de un análisis estadístico dependen en gran medida de las hipótesis de partida, en general expresadas en términos de un modelo, y de las observaciones experimentales. Esta dependencia motiva la necesidad de estudiar estimadores, tests... que sean robustos ante determinadas perturbaciones del modelo, de identificar las observaciones atípicas o amortiguar el efecto de su presencia, y de evaluar el impacto de cada una de las observaciones sobre las conclusiones del estudio. Naturalmente, la reflexión anterior es válida en el muestreo en poblaciones finitas. El estudio de las referencias a los problemas antes señalados, en el contexto del muestreo en poblaciones finitas, pone de manifiesto un desarrollo desigual de los distintos tópicos, y la importancia del esquema probabilístico desde el que se aborda la inferencia: población fija o modelo de superpoblación.

En este trabajo se presentan algunas de las referencias sobre los problemas descritos. En particular, se recogen algunos resultados referentes a la robustez del estimador de razón y una recopilación exhaustiva de los diagnósticos de influencia en el muestreo en poblaciones finitas.

Palabras clave: robustez, observaciones atípicas, muestreo en poblaciones finitas, diagnóstico de influencia.

Introducción

Las hipótesis de estudio, expresadas en general mediante un modelo matemático-estadístico, y los datos muestrales resultantes de la experimentación desempeñan un papel fundamental en las conclusiones finales de todo análisis estadístico, en particular en el muestreo en poblaciones finitas, por lo que es razonable plantearse las siguientes cuestiones:

- ¿Cómo incide sobre las conclusiones el modelo supuesto?
- ¿Qué incidencia tiene en las conclusiones la presencia en los datos experimentales de "observaciones no representativas" (observaciones atípicas, extremas, outliers)?

Los modelos matemáticos en general, y los estadísticos en particular, son casi siempre una descripción simplificada y aproximada de una realidad más compleja. Neyman y Pearson (1937) afirman: "Los matemáticos tratan con conceptos matemáticos, no con cosas reales y sólo podemos esperar una cierta relación entre ambas".

En ocasiones, pequeñas perturbaciones o modificaciones en el modelo distorsionan o alteran de tal forma las conclusiones que algunos autores, por ejemplo Huber (1975), consideran que proteger las conclusiones de tales distorsiones es a menudo más importante que minimizar la varianza.

Las consideraciones planteadas motivan la necesidad de determinar estimadores, tests... y, en general, procedimientos que no sean sensibles frente a determinadas perturbaciones en el modelo, lo que se ha dado en llamar *procedimientos robustos*. El estudio de este tipo de procedimientos en general se realiza de acuerdo al siguiente esquema:

- Considerar las circunstancias en las que más probablemente el modelo es erróneo.
- Describir estas circunstancias mediante un modelo alternativo.
- Analizar el procedimiento propuesto bajo el modelo alternativo.
- Comparar el procedimiento propuesto con el procedimiento óptimo bajo el modelo alternativo.

En relación con la segunda pregunta antes formulada, cabe afirmar que en ocasiones algunos aspectos de las conclusiones pueden estar dominados por una observación particular, o por un número reducido de ellas. Este problema se puede abordar desde distintas perspectivas, que en ocasiones se superponen: identificar tales observaciones, obtener estimadores, tests... que no sean sensibles o que amortigüen el efecto de la presencia de observaciones extremas (procedimientos robustos) y disponer de medidas capaces de evaluar el impacto que determinadas observaciones tienen sobre las conclusiones del análisis (medidas o diagnósticos de influencia).

Siguiendo la línea de Cook (1987), los problemas expuestos pueden plantearse de forma global de la siguiente forma:

Sea $R(D,M)$ los resultados del análisis debido a los datos D y al modelo postulado M. Generalmente R puede representar una predicción, una estimación de un parámetro, una distribución *a posteriori*... Sea ω un vector de perturbaciones que toma valores en cierto espacio Ω, y sea $M(\omega)$ el modelo perturbado, suponiendo que existe un $\omega_0 \in \Omega$ de forma que $M(\omega_0) = M$. El objetivo es comparar $R(D,M(\omega_0))$ y $R(D,M)$.

Como no podía ser de otra forma, los problemas señalados, genéricos de cualquier estudio estadístico, pueden plantearse en el muestreo en poblaciones finitas. El análisis de los trabajos en este campo ha de tener presente el enfoque desde el que se realiza la inferencia.

Inferencia y muestreo en poblaciones finitas

Como planteamiento general supondremos que se tiene una población finita, $U = \{1,..., N\}$, formada por N individuos. A cada unidad poblacional, k, $k = 1, ..., N$, se le asocia un vector $(y_k, x_{1k}, ..., x_{pk})$, siendo y_k la variable objeto de estudio desconocida y $(x_{1k}, ..., x_{pk})$ que se puede considerar como una información adicional disponible, conocida, siendo el parámetro de interés el total poblacional $T(y) = \Sigma_U y_i$.

Existen, básicamente, dos aproximaciones distintas a la teoría del muestreo en poblaciones finitas. La diferencia esencial entre ambas radica en la estructura probabilística que subyace a la hora de realizar la inferencia.

La aproximación clásica utiliza como base de la inferencia la distribución de probabilidad generada por un diseño muestral, $D = (M, P(\cdot))$, siendo M es espacio muestral y $P(\cdot)$ la distribución de probabilidad definida sobre M. Cuando la inferencia se realiza a partir del diseño, la validez de los resultados depende sólo del proceso de selección aleatoria, un proceso creado y controlado por el experimentador.

Sin embargo, algunos problemas de muestreo pueden ser analizados de forma útil y realista como problemas de predicción bajo un modelo adecuado de superpoblación ξ, originando lo que se ha dado en llamar la aproximación predictiva, en la que ξ desempeña un papel esencial en la inferencia.

Los distintos trabajos sobre robustez, observaciones extremas (outliers) e influencia que se han desarrollado en la literatura dependen de la perspectiva desde la que se realiza la inferencia.

Robustez respecto del modelo

Bajo el término genérico de robustez se ha englobado a una serie de procedimientos en los que subyacen las ideas anteriormente expresadas, aunque admite distintas matizaciones. Por ejemplo, cuando la inferencia se basa en el diseño el término robusta, generalmente, se utiliza para referirse a estimadores asintóticamente insesgados, respecto del diseño.

A continuación, se concretan algunas de las ideas expuestas anteriormente, cuando la inferencia se realiza desde el enfoque predictivo.

En el enfoque predictivo se supone que $(y_1, ..., y_N)$ es una realización de un vector aleatorio $(Y_1, ..., Y_N)$, sobre cuya distribución se realizan ciertas hipótesis, que es lo que se conoce como *modelo de superpoblación*, y que en términos generales se denotará por ξ.

Dada una muestra s, se puede representar

$$T(y) = \sum_{s} y_i + \sum_{\bar{s}} y_i.$$

y el problema de estimar $T(y)$ se puede considerar como el de predecir la suma de las variables no observadas

$$\sum_{\bar{s}} Y_i.$$

El nexo de unión entre lo observado y lo no observado lo proporciona el modelo ξ. A partir de la información muestral, $\{(i, y_i), i \in s\}$, se estiman los parámetros del modelo que se utilizan para predecir los valores no observados.

Ha de señalarse que la estimación de los parámetros del modelo ξ debe considerarse en todo caso como un elemento accesorio, pues el objetivo es estimar $T(y)$.

En este contexto se plantea el problema de determinar un predictor de $T(Y)$, $T(X, \xi)$ que dependa de la información auxiliar, X, y del modelo considerado, ξ, de tal forma que sea óptimo en algún sentido.

Obviamente, la validez de las conclusiones se sustenta en la validez del modelo de superpoblación considerado, y ya que en la práctica rara vez se tiene certeza absoluta sobre éste, puede plantearse qué efectos tienen, sobre la teoría generada, algunas modificaciones en el modelo.

En general $T(X, \xi)$ dependerá del modelo ξ propuesto, por lo que cabe preguntarse por su validez en el caso en que se considerara un modelo alternativo ξ^*. En este caso se dirá que $T(X, \xi)$ es robusto respecto del modelo ξ^* si

$$T(X, \xi) = T(X, \xi^*)$$

Modelo y predicción

Entre los trabajos pioneros en esta línea podemos citar el de Royal y Herson (1973), ya que muchos de los trabajos posteriores abordan generalizaciones o extensiones del problema aquí planteado.

Royal y Herson consideran el modelo de superpoblación ξ $[\delta_0, ..., \delta_J; \nu(x)]$, caracterizado por las siguientes especificaciones:

- $Y_1, ..., Y_N$ son variables aleatorias incorreladas.
- $E(Y_k) = h(x_k)$, siendo

$$h(x) = \delta_0\beta_0 + \delta_1\beta_1 x + \delta_2\beta_2 x^2 + \ldots + \delta_J\beta_J x^3$$

donde $\delta_j \in \{0,1\}$, $j = 1, \ldots, J$, indica si se incluye o no el término de orden j.

- $var(Y_k) = \sigma^2 v(x_k)$.

Supuesto el modelo anteriormente especificado, y dada una muestra s, se plantea determinar el predictor lineal

$$\hat{T} = \sum_s \omega_i Y_i$$

ξ es insesgado de $T = T(Y)$, es decir, verificando

$$E_\xi\left[\hat{T} - T\right] = 0$$

y que sea óptimo en el sentido mínimo cuadrático, es decir, solución de

$$\min E_\xi\left(\hat{T} - T\right)^2$$

Si denotamos por $\hat{T}[\delta_o,\ldots,\delta_J;\upsilon(x)]$ a la solución del problema planteado, se obtiene que

$$\hat{T}[\delta_o,\ldots,\delta_J;\upsilon(x)] = \sum_s y_k + \sum_{j=0}^{J}\left(\sum_{\bar{s}} x_k^j\right)\delta_j\hat{\beta}_j$$

donde los x son los estimadores de mínimos cuadrados ponderados de los coeficientes de regresión del modelo especificado.

Como caso particular, se obtiene que en el modelo puramente aleatorio, $\xi[1:1]$, el estimador óptimo, en el sentido definido, es el estimador de expansión

$$\hat{T}[1:1] = N\bar{y}(s)$$

donde por $\bar{y}(s)$ se denota a la media muestral, o que en un modelo de regresión, a través del origen, supuesto que la varianza de Y es proporcional a x, $\xi\ [0, 1 : x]$, el estimador óptimo es el estimador de razón

$$\hat{T}[0,1:x] = T(x)\frac{\sum_s Y_k}{\sum_s x_k}$$

Estrategia óptima

En el caso particular de $\hat{T}[0,1:x]$ o de cualquier otro estimador, $\hat{T}[\delta_o,\ldots,\delta_J;\upsilon(x)]$, cabría preguntarse, además, por aquellas muestras, $s[\delta_o,\ldots,\delta_J;\upsilon(x)]$, para las que se minimiza el error cuadrático medio

$$\min_s E_\xi\left[\hat{T}[\delta_o,\ldots,\delta_J;\upsilon(x)] - T\right]^2$$

En estas condiciones el par $\left(s[\delta_o,...,\delta_J;\upsilon(x)];\hat{T}[\delta_o,...,\delta_J;\upsilon(x)]\right)$ es la estrategia óptima bajo el modelo $\xi[\delta_o,...,\delta_J;\upsilon(x)]$.

En particular, para el estimador de razón se verifica que

$$E_\xi\left[\hat{T}[0,1:x]-T\right] = \sigma^2 T(x)\frac{\sum_{\bar{s}} x_k}{\sum_s x_k}$$

por lo que la muestra óptima es aquélla formada por los n mayores valores de la variable x, supuesto que x es positiva.

Sesgo del estimador de razón

Naturalmente, todos los resultados expuestos dependen del modelo supuesto. Por ejemplo, bajo el modelo de regresión simple, no necesariamente a través del origen, y varianza proporcional a x, ξ [1, 1 : x], el estimador de razón ya no es óptimo, ni siquiera es, en general, insesgado, y su sesgo viene dado por

$$E_\xi\left[\hat{T}[0,1:x]-T\right] = \beta_0 N\frac{\bar{x}-\bar{x}(s)}{\bar{x}(s)}$$

En general, bajo el modelo $\xi[\delta_o,...,\delta_J;\upsilon(x)]$, se tiene que

$$E_\xi\left[\hat{T}[0,1:x]-T\right] = \sum_{j=0}^{J}\delta_j\beta_j N\bar{x}\left\{\frac{\overline{x^j}(s)}{\bar{x}(s)}-\frac{\overline{x^j}}{\bar{x}}\right\} \qquad (1)$$

es decir, en general, el estimador de razón no conserva la propiedad de insesgadez.

Muestras balanceadas y robustez

Sin embargo, a partir de (1) se observa que el estimador de razón sigue siendo insesgado bajo el modelo $\xi[\delta_o,...,\delta_J;\upsilon(x)]$ si la muestra, s, verifica que

$$\frac{\overline{x^j}(s)}{\bar{x}(s)} = \frac{\overline{x^j}}{\bar{x}}$$

para aquellos valores j en los que $\delta_j = 1$, lo que origina el concepto de muestra balanceada.

Definición: dado J, entero positivo, una muestra se dirá balanceada si

$$\overline{x^j}(s) = \overline{x^j};\ j=1,...,J$$

denotándose por s (J) al conjunto de muestras balanceadas.

Es decir, una muestra es balanceada si los momentos de orden j, j=1,...J, coinciden sobre la muestra y la población. En este sentido, se podría decir que la muestra es una *fidedigna* representación de la población.

Es de destacar que para cualquier muestra balanceada

$$\hat{T}[0,1:x] = T(x)\frac{\bar{y}(s)}{\bar{x}(s)} = N\bar{y}(s) = \hat{T}[1:1]$$

es decir, el estimador de razón, óptimo en el modelo ξ [0, 1 : x], coincide con el estimador de expansión, que es óptimo en el modelo ξ [1 : 1].

Así pues, si se utiliza una muestra balanceada, el estimador de razón es óptimo bajo los dos modelos, es decir, es robusto.

Este resultado se puede generalizar, como se recoge en el siguiente teorema.

Teorema: Si s $\in$ s (J) entonces

$$\hat{T}[1,\delta_1,...,d_J:1] = \hat{T}[\delta_0,1,\delta_2,...,\delta_J:x] = \hat{T}[\delta_0,\delta_1,1,...,\delta_J:x^2] =$$
$$... = \hat{T}[\delta_0,\delta_1,\delta_2,...,1:x^J] = N\bar{y}(s)$$

para cualquier secuencia $\delta_0,\delta_0,\delta_0,\delta_0$, de ceros y unos.

El teorema nos afirma que, cuando la muestra es balanceada, el estimador de expansión, $N\bar{y}(s)$, que coincide con el de razón, es óptimo en cualquier modelo de regresión de grado J, en el que la varianza sea proporcional a x^j, para algún $j = 1, ..., J$, siempre y cuando la función de regresión contenga al término de grado j, $\beta_j x^j$.

Hay que tener presente que considerar una muestra balanceada, junto con el estimador de razón, presenta una protección frente a un determinado tipo de error en el modelo, pero pierde eficiencia, ya que las muestras balanceadas en general no son las muestras óptimas y, por tanto, la estrategia formada por una muestra balanceada y el estimador de razón no es óptima en todos los modelos.

Generalizaciones

A partir del trabajo de Royal y Herson surgen algunas generalizaciones del mismo como, por ejemplo, el de Scott, Brewer and Ho (1978), en el que se considera una función de varianza genérica $V(x)$; o el de Bragança, Pereira y Rodrigues (1983), que considera un modelo lineal dependiente de k variables. También cabe destacar los trabajos de Royal y Pfefferman (1982) y Bolfarine, Bragança y Rodrigues (1987), en los que se analiza el problema de la robustez, pero desde una perspectiva bayesiana.

Análisis de influencia

El estudio desarrollado hasta el momento se ha centrado en uno de los elementos básicos de la inferencia: el modelo. A continuación, nos centramos en el otro elemento: las observaciones muestrales.

El problema de la calidad, representatividad o influencia de los datos muestrales es de gran interés ya que las conclusiones del análisis estadístico se basa en ellos, al

menos en gran medida. Este problema se puede abordar desde diversas perspectivas que, en ocasiones, se superponen, dependiendo del objetivo final perseguido.

El objetivo del estudio puede ser el de identificar aquellas observaciones que son extremas (*outliers*) en algún sentido, que se desvían marcadamente del comportamiento del resto. Las consecuencias que se derivan de identificar una o varias observaciones como outliers puede ser diversa. Una opción es excluirlas del análisis por no considerarse representativas, pero en ocasiones ponen de manifiesto algunos aspectos de la población objetivo que no habían sido considerados a priori, lo que origina que el estudio de este tipo de observaciones pueda ser un objetivo por si mismo.

Algunos autores (Barnett, 1993) señalan que el problema de la identificación de outliers en poblaciones finitas sólo puede realizarse en el caso en que la inferencia se realice desde el enfoque predictivo, ya que de alguna forma se necesita un modelo para poder cuantificar el hecho de "desviarse marcadamente del comportamiento del resto". En este caso serían adaptables algunos de los procedimientos propuestos en poblaciones infinitas.

En ocasiones, el objetivo no es precisamente el de identificar las observaciones extremas, sino el de obtener estimadores, tests,... sobre los que la presencia de outliers tenga unos efectos limitados. Es decir, el objetivo es obtener procedimientos robustos, en este caso frente a la presencia de outliers. Esta forma de abordar el problema se puede denominar de acomodación, en contraposición a la identificación antes expuesta.

Los textos clásicos de muestreo no contienen ninguna referencia explícita a este problema. Sin embargo, una lectura detenida muestra algunas ideas muy próximas a lo expuesto bajo el epígrafe de poblaciones sesgadas. Por ejemplo, Kish (1965): "La muestra de una población muy sesgada puede estar distorsionada si en ella aparecen unas cuantas unidades con valores muy grandes. Si éstas aparecen con una probabilidad de selección pequeña, reciben grandes pesos y tienen un gran efecto sobre la media muestral y su varianza".

A la hora de construir un procedimiento robusto frente a los outliers deben considerarse los elementos que controla el experimentador, fundamentalmente el diseño muestral y el estimador a utilizar. Si existe información a priori sobre la existencia de valores grandes en la población, una estrategia obvia es la de estratificar la población y colocar todos los valores grandes en un estrato separado, aunque desde un punto de vista práctico se origina un problema adicional al ser, en general, desconocido el tamaño del estrato así definido. Entre los trabajos desarrollados en esta línea podemos citar los de Hidiroglu y Srinath (1981) y el de Chambers (1986).

Un tercer enfoque posible es el de evaluar la influencia que cada observación o grupo de observaciones tienen sobre las conclusiones del estudio estadístico, identificando aquéllas que tienen un efecto considerable sobre los resultados del análisis, lo que ha originado el denominado *Análisis de Influencia*. Obviamente, el planteamiento de este problema no es tampoco exclusivo del muestreo en poblaciones finitas. Todo

lo contrario, son escasísimas las referencias a este tópico dentro de este campo, siendo por el contrario en el Modelo Lineal en el que más se ha desarrollado.

Los métodos propuestos en la literatura para medir la influencia son muy diversos, originando lo que se denominan diagnósticos o medidas de influencia. Según Cook y Weisberg (1982), "La idea básica en el análisis de influencia es muy simple. Introducimos pequeñas perturbaciones en la formulación del problema y entonces calculamos cuánto cambian los resultados del análisis por la perturbación".

De acuerdo a la idea expresada por Cook y Weisberg, queda por definir el esquema de perturbación que se va a considerar y la forma de cuantificar los cambios que se producen en las conclusiones del estudio, es decir, en los estimadores, tests,...

Aunque la forma de introducir una perturbación y de cuantificar los cambios puede ser muy diversa e, incluso, en algunos casos específicos del problema que se considere, en la literatura se han propuesto algunos procedimientos que tienen cierta aplicación general, destacando el denominado esquema de omisión y la función de influencia, en alguna de sus versiones.

Cuando, en poblaciones finitas, se adopta el enfoque predictivo, en general se considera un modelo de tipo lineal, que es precisamente el modelo para el que más diagnósticos de influencia se han desarrollado. Los diagnósticos propuestos para el Modelo Lineal podrían adaptarse para poblaciones finitas, por lo que nos centramos a continuación en los que se han propuesto de forma específica para el caso de poblaciones finitas, cuando la inferencia se basa en un diseño muestral.

Diagnósticos basados en la omisión

El esquema de la omisión, quizás el más utilizado, es tal vez también el más simple. Se comparan los resultados del análisis, considerando todas las observaciones experimentales y omitiendo una de ellas. Obviamente, este tipo de diagnóstico puede ser construido en el caso del muestreo en poblaciones finitas, aunque presenta algunas peculiaridades, como destaca Smith (1987).

En el caso de que se considere un estimador de tipo lineal, $\hat{T} = \sum_s \omega_i Y_i$, un diagnóstico del tipo omisión para evaluar la influencia de la i-ésima unidad muestral es

$$I(i) = 100 \frac{\hat{T} - \hat{T}_{(i)}}{\hat{T}}$$

donde $\hat{T}_{(i)}$ representa el estimador basado en la muestra, excluyendo la unidad i. En muestras de gran tamaño, el diagnóstico es aproximadamente

$$\frac{100\,\omega_i Y_i}{\sum_s \omega_i Y_i}$$

donde se observa claramente que tanto Y_i como ω_i contribuyen a la influencia del i-ésimo caso.

Smith detaca que, en general, el i-ésimo caso muestral puede influir en el estimador o sobre su varianza, debido a valores extremos de Y_i, o de ω_i, o a una combinación de ambos, resaltando que éste es un hecho distintivo del muestreo en poblaciones finitas.

Diagnósticos basados en la función influencia

Hampel (1974) introduce el concepto de función influencia, con el objeto de estudiar el comportamiento de los estimadores ante pequeñas perturbaciones de la función de distribución.

Las conclusiones de un análisis estadístico vienen determinadas en general por un estadístico $T(X_1, ..., X_n)$, basado en una muestra, $X_1, ..., X_n$, seleccionada de una población con función de distribución F. En ocasiones, $T(X_1, ..., X_n)$, puede expresarse como un funcional de la función de distribución empírica, F_n, es decir, $T = T(F_n)$, pudiéndose expresar además el parámetro de interés, θ, de la forma $\theta = T(F)$.

En este contexto, la función influencia (*IF*) de T en F se define como

$$IF(x,T,F) = \lim_{t\to 0} \frac{T[(1-t)F + t\Delta_x] - T(F)}{t}$$

para aquello puntos x en los que exista el límite, siendo Δ_x la función de distribución asociada a una variable aleatoria degenerada en el punto x.

La importancia de la función influencia radica en su interpretación heurística: describe el efecgto de una perturbación infinitesimal en el punto x sobre θ, y que bajo ciertas condiciones de regularidad, y desde el punto de vista asintótico

$$T(X_1,...,X_n) - \theta = T(F_n) - T(F) \approx \sum_{i=1}^{n} IF(X_i,T,F)$$

Es decir, la diferencia entre el estimador y el parámetro a estimar se puede expresar, de forma aproximada, en términos de la función influencia evaluada sobre cada una de las observaciones muestrales.

Esta idea genérica puede adaptarse al caso de poblaciones finitas, aunque se ha llevado a cabo en muy pocas situaciones, recogidas en los trabajos de Gwet y Rivest (1992) y Hulliger (1995), y tratando problemas muy particulares. Gwet y Rivest estudian el estimador de Horvitz-Thompson en un diseño II *PS*.

En ambos casos, aunque la inferencia se basa en el diseño, el proceso de construcción de la función influencia se basa en el modelo lineal, $\xi\,[0, 1 : x]$, que subyace cuando se aplica el estimador de razón y el de Horvitz-Thompson.

Anteriormente, hemos señalado que los distintos enfoques planteados se sobreponen en algunas ocasiones. El objetivo básico de los trabajos de Gwet y Rivest y de Hulliger es el de obtener versiones robustas, en concreto *M* o *GM* estimadores, frente a la presencia de observaciones extremas del estimador de razón y de Horwitz-

Thompson, respectivamente. La obtención de un diagnóstico de influencia en este caso es un elemento colateral del estudio.

En el caso particular en que el diseño muestral sea un Muestreo Aleatorio Simple y el estimador considerado sea el de razón, del trabajo de Gwet y Rivest se obtiene el siguiente diagnóstico de influencia para la i-ésima unidad muestral

$$\widehat{IC}_i = T(X)\frac{y_i - \frac{\bar{y}(s)}{\bar{x}(s)}x_i}{\bar{x}(s)}, \quad i = 1,...,n \tag{2}$$

es decir, la influencia se expresa en términos del residuo respecto del modelo lineal, que subyace en el estudio.

En el trabajo de Hulliger (1995), se considera que el diseño muestral es un diseño $\text{II}PS(N, n, X)$, es decir, las probabilidades de inclusión de primer orden $\pi_i = P(i \in s)$, $i = 1, ..., N$, son proporcionales al tamaño X

$$\pi_i = \frac{nx_i}{T(x)}, \; i = 1,...,N$$

y, supuesto que $\pi_i > 0$, $i = 1, ..., N$, que el estimador que se utiliza es el de Horvitz-Thompson

$$\hat{T}_{HT} = \sum_s \frac{Y_i}{\pi_i}$$

En este caso particular, obtiene la siguiente medida de influencia

$$SC\left(\hat{T}_{HT}\right)_i = N\frac{y_i - \beta_{ls}x_i}{\pi_i\left(\sum_s \frac{1}{\pi_j}\right)} \tag{3}$$

siendo

$$\beta_{ls} = \frac{1}{n}\sum_s \frac{Y_i}{x_i}$$

Nuevamente se observa cómo la medida de influencia, al igual que en el caso anterior, se expresa en función del residuo del modelo lineal que subyace, aunque la inferencia no se realice desde el enfoque predictivo.

Otro tipo de diagnósticos

Los diagnósticos que se han considerado en el epígrafe anterior son muy específicos, aunque quizás siguiendo la misma línea de los trabajos descritos pudieran obtenerse medidas de influencia en otros casos particulares.

Moreno, Muñoz y Muñoz (1999) proponen un diagnóstico de influencia, adaptando el propuesto por Muñoz, Muñoz y Moreno (1995), basado en el sesgo condicionado y que, a diferencia de los referenciados anteriormente, se puede aplicar en un contexto muy general.

Sea $\theta = \theta(Y)$ el parámetro de interés, no necesariamente el total poblacional; $\hat{\theta} = \hat{\theta}(s)$ un estimador de θ, basado en una muestra s, e $I_i(s)$, $i = 1, ..., N$, las variables aleatorias, $I_i = 1$ si $u_i \in s$; $I_i(s) = 0$ en c.c., que especifican si el individuo u_i pertenece o no a la muestra.

Si $0 < \pi_i < 1$, el sesgo condicionado de $\hat{\theta}$, causado por la presencia de u_i en la muestra, $S(I_i = 1; \hat{\theta})$, se define como

$$S(I_i = 1; \hat{\theta}) = E(\hat{\theta} | I_i = 1) - E(\hat{\theta})$$

Es decir, el sesgo condicionado como diagnóstico de influencia cuantifica la desviación que se produce en el valor esperado del estimador cuando el diseño muestral se perturba, restringiéndolo sobre las muestras que contienen a u_i.

Una expresión alternativa de $S(I_i = 1; \hat{\theta})$ viene dada por

$$S(I_i = 1; \hat{\theta}) = (1 - \pi_i)\left\{E(\hat{\theta} | i_i = 1) - E(\hat{\theta} | i_i = 0)\right\}$$

y, por tanto, el sesgo condicionado es 0 si y sólo si el valor esperado sobre las muestras que contienen a u_i es el mismo que el valor esperado sobre las muestras que no lo contienen.

Obviamente, el sesgo condicionado depende tanto del diseño muestral como del estimador que se considere. Además, ha de tenerse presente que, en general, es un parámetro poblacional desconocido, por lo que desde un punto de vista práctico será necesario estimarlo.

A continuación se recogen algunos ejemplos de aplicación del sesgo condicionado, en comparación con algunos de los diagnósticos anteriormente expuestos.

En el caso del Muestreo Aleatorio Simple (MAS), y supuesto que el estimador que se utiliza es el estimador de expasión, $\hat{T} = N\bar{y}(s)$, se obtiene que

$$\hat{S}(I_i = 1; \hat{T}_R) = T(x)(1 - f)\frac{y_i - \dfrac{\bar{y}(s - u_i)}{\bar{x}(s - u_i)}x_i}{n\bar{x}(s)}$$

siendo $f = \dfrac{n}{N}$, la fracción de muestreo.

Puede compararse este diagnóstico con el propuesto por Gwet y Rivest en el mismo contexto (2), MAS y estimador de razón, observándose que ambos evalúan la influencia a través de un residuo, aunque de forma ligeramente distinta.

En el caso del estimador de Horvitz-Thompson, bajo un diseño II*PS*, se obtiene la siguiente estimación del sesgo condicionado

$$\hat{S}(I_i = 1; \hat{T}_{HT}) = \sum_{j=1}^{N} Y_j \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_j \pi_{ij}} I_j \qquad (4)$$

siendo π_{ij} las probabilidades de inclusión de segundo orden, $\pi_{ij} = P(u_i, u_j \in s)$.

Una simple comparación formal de (4) con el diagnóstico que se obtiene, en el mismo contexto, a partir del trabajo de Hulliger (3), pone de manifiesto algunas diferencias significativas. El diagnóstico obtenido a partir del sesgo condicionado no está expresado en términos de un residuo, y además depende de las probabilidades de inclusión de segundo orden, que no son las mismas en todos los diseños II*PS*, sino que dependen del procedimiento empleado para seleccionar la muestra.

Referencias

Bolfarine, H.; Bragança, C.A. y Rodrigues, J. (1987) Robust linear prediction in finite populations: A bayesian perspective. *Sankhyā*, Series B, 49, 23-55.

Bragança, C.A. y Rodrigues (1983) Robust linear prediction in finite population. *International Statistical Review*, 51, 293-300.

Chambers, R.L. (1986) Outlier robust finite population estimation. *Journal of American Statistical Association*, 81, 1063-1069.

Cook, R.D. (1987) Influence Assesment. *Journal of Applied Statistics*, 14, 117-132.

Kish, L. (1965) *Survey Sampling*, New York: John Wiley and Sons.

Cook, R.D. y Weisberg, S. (1982) *Residuals and Influence in Regression*. London: Chapman and Hall.

Gwet, J.P. y Rivest, L.P. (1992) Outlier resistant alternatives to the ratio estimator. *Journal of American Statistical Association*, 87, 1174-1182.

Hidiroglou, M.A. y Srinath, K.P. (1981) Some estimators of a population total from simple random samples containing large units. *Journal of American Statistical Association*, 76, 690-695.

Hampel, F.R. (1974) The influence curve and its role in robust estimation. *Journal of American Statistical Association*, 69, 383-393.

Huber, P.J. (1975) *Robustness and designs. In a survey of statistical designs and linear models*. Amsterdam: North-Holland.

Hulliger, B. (1995) Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21, 79-87.

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

Muñoz Pichardo, J.M.; Muñoz García, J.; Moreno, J.L. y Pino, R. (1995) A new approach to influence analysis in linear models. *Sankhyā*, 57, 393-409.

Neyman, J. y Pearson, E. (1937) A note on some points in 'Student's' paper on 'Comparison between balanced and random arrangements in field plots'. *Biometrika*, 29, 380-388.

Royal, R.M. y Herson, J. (1973) Robust finite estimation in finite population I. *Journal of American Statistical Association*, 68, 880-889.

Royal, R.M. y Herson, J. (1973) Robust finite estimation in finite population II. *Journal of American Statistical Association*, 68, 890-893.

Royal, R.M. y Pfefferman, D. (1982) Balanced samples and robust bayesian inference in finite sampling. *Biometrika*, 69, 401-409.

Scott, A.J.; Brewer, K.R.W. y Ho, E.W.H. (1978) Finite population sampling and robust estimation. *Journal of American Statistical Association*, 73, 359-361.

Smith, T.M.F. (1987) Influential observations in survey sampling. *Journal of Applied Statistics*, 14, 143-152.