Método de reponderación aplicado en la EPA

La Encuesta de Población Activa (EPA), como cualquier otra encuesta a hogares, puede tener distorsiones en las estimaciones que produce, debido a una serie de causas ligadas al trabajo de campo y al diseño muestral. A continuación se enumeran dichas causas, y sus principales consecuencias en los resultados:

a) Falta de respuesta, que produce sesgos en los estimadores finales al afectar de forma relativamente distinta a hogares o grupos de población de determinada composición.

b) Posible tendencia de la muestra a sobrerrepresentar a hogares de una determinada estructura en detrimento de otros. En este sentido, la actualización del marco de la encuesta (y a pesar de que el proceso de actualización del seccionado se ha introducido de forma habitual en el diseño de la muestra) por la propia duración del proceso, introduce inevitablemente un retraso en la captura de los cambios de población que puede tener influencia en las estimaciones. En definitiva, es más fácil detectar un cambio en los niveles de población (en cifras agregadas), que la localización concreta de los nuevos contingentes de población.

Por otra parte, a menudo se cuenta con fuentes estadísticas externas fiables que pueden mejorar la calidad de las estimaciones de la encuesta Un procedimiento para llevar a la práctica dicha mejora es la reponderación.

La reponderación consiste en corregir los pesos o factores de elevación originales deducidos del diseño de la encuesta de forma que se llegue a unos factores finales tales que, al aplicarlos, la estimación de las variables para las que se dispone de información de la fuente externa fiable (datos de referencia para la reponderación) coincidan con los datos de dicha fuente.

La reponderación de los factores de elevación deducidos del diseño de la muestra, es una práctica habitual en la Unión Europea. En el caso de la Encuesta de Fuerza de Trabajo (la EPA europea), se prevé como mejora técnica deseable a implantar una vez que se disponga de la información necesaria (Reglamento del Consejo 577/1998, relativo a la organización de una encuesta muestral sobre la población activa en la comunidad, en su artículo 3, apartado 5).

Para llevar a cabo cualquier procedimiento de reponderación es necesario elegir unas variables auxiliares o explicativas X que existan tanto en la encuesta como en una fuente estadística ajena a esta, bien sea un Censo, el Padrón o un Registro Administrativo.

Las variables X a elegir, aparte de existir tanto en la encuesta como en fuente estadística alternativa, es deseable que presenten una correlación lo más fuerte posible con las variables de interés Y.

Por otro lado, las variables de interés Y son aquellas cuyas estimaciones revisten la mayor importancia en la encuesta, en el caso de la EPA serían las relacionadas directamente con la actividad, como son la condición de ocupados, parados, etc.

Es indudable que estas variables están fuertemente condicionadas por el sexo y la edad de la persona.

También teniendo en cuenta el carácter trimestral de la encuesta, será necesario disponer de una serie homogénea de las variables X en la fuente alternativa con la misma periodicidad que la EPA.

Generalmente las variables explicativas favoritas son los efectivos de población clasificados por grupos de edad y sexo. En el caso de la EPA se ha optado por utilizar las proyecciones de población relativas a cada Comunidad Autónoma y referidas a cada trimestre.

Se han tomado once grupos de edad cruzados con sexo, lo cual representa para cada comunidad autónoma un vector de veintidós efectivos poblacionales proporcionados por las proyecciones en cada trimestre.

Si en la encuesta se parte de una muestra de tamaño n, llamando ω al vector de pesos originales de dimensión $n \times 1$ y $\dot{\omega}$ al vector homólogo de pesos transformados, cualquier procedimiento de reponderación que se aplique dará lugar a una relación funcional del tipo $\dot{\omega} = \dot{\omega}(\omega, X)$, es decir, los nuevos pesos van a ser función de los originales y de las variables auxiliares elegidas.

La información auxiliar proporcionada por la encuesta va a estar contenida en una matriz $X_{n \times p}$ donde en cada fila aparecen los valores de las variables auxiliares o de las modalidades de estas para cada uno de los individuos de la muestra. En este caso p tomaría el valor 22.

Los nuevos pesos han de cumplir la condición de equilibrado de la muestra es decir $X'\dot{\omega} = x$, siendo x el vector de efectivos poblacionales proporcionados por las proyecciones.

Con los pesos $\dot{\omega}$, se pasaría a obtener nuevas estimaciones para cualquier variable de interés Y.

Procedimiento empleado

Como procedimiento para llevar a cabo la reponderación, el INE ha optado por el CALMAR que es una macro en SAS desarrollada por el INSEE de Francia.

En este método se define previamente una función de distancia $G(\omega, \dot{\omega})$ y se exige que $\sum_{k=1}^{n} \omega_k G(\omega_k, \dot{\omega}_k)$ sea mínimo para el conjunto de la muestra con la ligadura $\sum_{k=1}^{n} \dot{\omega}_k = N$, es decir, la suma de pesos transformados debe recuperar un determinado total de población.

Llamando h al cociente $\frac{\dot{\omega}_k}{\omega_k}$, se definen las dos distancias más usualmente utilizadas:

cuadrática $G(h) = \left(\frac{h-1}{2}\right)^2$

y logarítmica $G(h) = h \log(h) - h + 1 \qquad h > 0$

Asociadas a estas funciones de distancia existen las siguientes funciones de transformación de los pesos nuevos respecto a los originales

$\dot{\omega} = \omega(1+u)$ lineal

$\dot{\omega} = \omega e^u$ exponencial

Con la lineal existe el riesgo de obtener pesos negativos mientras que con la exponencial puede haber mayor distorsión de pesos nuevos respecto a los originales.

La macro del CALMAR también ofrece la posibilidad de poner cotas a la transformación de los pesos originales, es decir, se buscan dos valores L y U tal que

$L < h_k < U \qquad k = 1,2....n \qquad$ donde $\qquad h_k = \frac{\dot{\omega}_k}{\omega_k}$

El INE ha optado por utilizar el método lineal truncado tomando como cota inferior $L = 0.1$ y cota superior $U = 10.0$

La macro puede aplicarse desde 1 a 7 dimensiones; en este caso se ha aplicado en 2 dimensiones a una tabla de contingencia cuyas frecuencias marginales de fila son las proyecciones de población de 16 y más años correspondientes a cada una de las provincias que constituyen una Comunidad Autónoma dada. Las frecuencias marginales de columna son la población de 16 y más años clasificada por sexo y grupos de edad quinquenales para el conjunto de la Comunidad Autónoma.

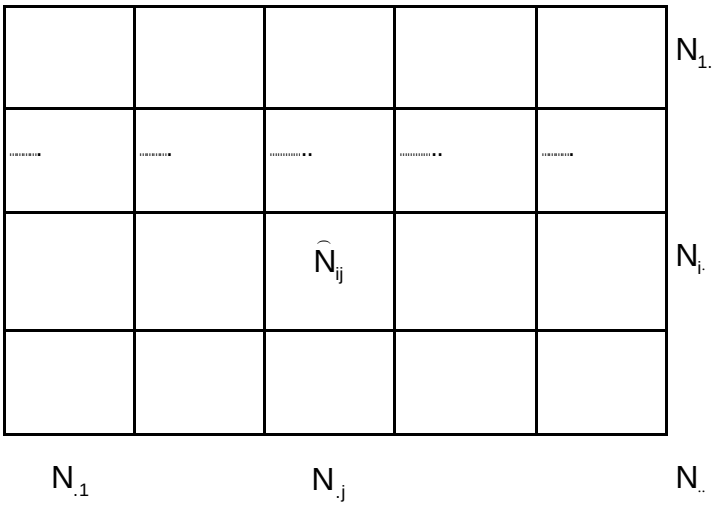
Según el diagrama adjunto (DIAGRAMA 1), $\hat{N}_{ij}$ es la estimación que proporciona la encuesta con los pesos originales para el efectivo poblacional de la provincia i y el grupo de edad j en la Comunidad autónoma considerada.

Al considerar la transformación lineal aunque con truncamiento, la relación entre los pesos nuevos y originales relativos a todos los registros k que pertenecen a la casilla i,j sería

$\dot{\omega}_k = \omega_k(1 + u_i + v_j)$

Las cantidades u_i y v_j son las incógnitas que resuelve la MACRO. Pueden tener cualquier signo aunque normalmente al sumarse dan una magnitud cercana a cero para así satisfacer mejor la condición de variación mínima en la transformación de los pesos.

DIAGRAMA 1



$N_{i.}$ = Total de población de 16 y más en la provincia i

$N_{.j}$ = Efectivo de población para el grupo de edad y sexo j en la Comunidad Autónoma considerada.

Las nuevas frecuencias de casillas transformadas $\hat{N}_{ij}$ habrán de recuperar las marginales $N_{i.}$ y $N_{.i}$ proporcionadas por la fuente estadística de referencia ajena a la encuesta, en nuestro caso, se trata de las proyecciones de población elaboradas por el INE.

El problema de la reponderación única

Las variables auxiliares empleadas en la reponderación son cualitativas (más concretamente binarias, toman los valores 0 ó 1) como la práctica totalidad de las variables de la EPA. Al aplicar el CALMAR en 2 dimensiones, cada fila de la matriz $X_{n\,x\,p}$ tendrá p=22 columnas; en una de ellas habrá un 1 y en las restantes un 0 ya que cada individuo de la muestra pertenece a una sola modalidad de grupo de edad y sexo. Por otro lado cada individuo tendrá k columnas adicionales que indican su pertenencia o no a alguna de las k provincias que constituyen la Comunidad Autónoma a la cual se está aplicando el CALMAR.

Los individuos de un mismo hogar que tengan coordenadas diferentes en la matriz $X_{n\,x\,p}$ verán también afectados sus pesos originales de forma diferente ya que el

factor de transformación $(1+u_i+v_j)$, v_j variará según el grupo de edad y sexo al que pertenece el individuo; esto hace que se pierda una de las cualidades originales del diseño como es la ponderación única para todos los individuos del mismo hogar. El parámetro u_i del factor de transformación no ocasiona ningún problema ya que sólo depende de la provincia a la que pertenece el hogar.

Para obviar el problema planteado, se construye una matriz $Z_{s\,x\,p}$ donde s es el número de hogares, que se obtiene muy fácilmente a partir de la $X_{n\,x\,p}$, sin más que sumar por columnas las coordenadas de los individuos pertenecientes a un mismo hogar.

Así, por ejemplo, si tenemos un hogar con 3 personas y las siguientes coordenadas de edad y sexo:

individuo 1	|0|	|0|	.....	|1|	.....	|0|
individuo 2	|0|	|1|	.....	|0|	.....	|0|
individuo 3	|0|	|0|	.....	|1|	.....	|0|

Las coordenadas de edad y sexo para el hogar serán

\|0\| \|1\| ..... \|2\| ..... \|0\|

Se aplicará ahora el CALMAR a la matriz Z (matriz que toma valores discretos, 0, 1, 2, etc., dependiendo del número de miembros del hogar de 16 y más años) que además tiene la ventaja de ser de dimensiones más reducidas que la X; el nuevo peso transformado que se obtiene para el hogar se imputará a todos sus miembros de forma que se recupera la cualidad inicial del diseño.

Al aplicar la macro del CALMAR pueden aparecer fácilmente problemas de colinealidad entre las variables, cuestión que se puede resolver suprimiendo una de las columnas en la matriz explicativa Z.

Para las Comunidades Autónomas uniprovinciales también se ha aplicado el CALMAR recurriendo a suprimir una de las coordenadas de las variables auxiliares.

Para conocer con detalle el fundamento matemático del CALMAR, se sugiere consultar los artículos:

Deville, Särndal y Sautory, *Generalized raking procedures in survey sampling*; Journal of the American Statistical Association, September 93, Vol. 88 nº 423.

Deville y Särndal, *Calibration estimators in survey sampling*. Journal of the American Statistical Association; June 92, Vol. 87, nº 418.