

ESTIMADORES ÓPTIMOS DE CALIBRACIÓN USANDO REGRESIÓN ORTOGONAL

ANDRÉS GUTIÉRREZ R.

RESUMEN. Los estimadores óptimos de calibración utilizan información auxiliar completa para producir estimaciones más eficientes; en estos términos y con una sola variable auxiliar, se decide utilizar, para estimar beta (el vector de parámetros en un modelo lineal), una función que minimice las distancias perpendiculares entre los valores observados y una recta de regresión.

Palabras clave: *estimadores de calibración, información auxiliar, regresión ortogonal.*

ABSTRACT. Optimal calibration estimators uses complete auxiliary information to produce more efficient estimates; in order to estimate beta, the parameters vector, we use a function what minimizes the perpendicular distances between the observed values and the regression line.

Key words: *auxiliary information, calibration estimators, orthogonal regression.*

1. INTRODUCCIÓN

En la mayoría de poblaciones donde hay una fuerte relación lineal entre la variable de estudio y y una variable auxiliar x , la línea de regresión poblacional intercepta el eje y a cierta distancia del origen. Con una sola variable x se tiene el siguiente modelo de superpoblación ξ , definido de la siguiente manera:

$$(1.1) \quad E_{\xi}(y_k) = \alpha + \beta x_k \quad V_{\xi}(y_k) = \sigma^2$$

Suponga el universo $U = (1, \dots, N)$ el conjunto de elementos en una población finita y s el conjunto de los elementos que conforman la muestra. Sean y_k , $k \in s$ y x_j , $j \in U$, los valores de las respuestas a la variable y y a la variable auxiliar x , asociados al i -ésimo elemento para y y al j -ésimo elemento de la población para x ; siendo π_k la probabilidad de inclusión de primer orden, se asume que el total $\mathbf{X} = \sum_U \mathbf{x}_k$ es conocido.

El objetivo de este artículo es construir tres formas diferentes para estimar el total de la variable de interés, apelando a la teoría de calibración. En primer lugar se desarrolla la teoría de calibración óptima de Wu, después se hace un recuento de tres métodos para la estimación de los parámetros desconocidos del modelo y por último se realiza una comparación empírica de los tres casos bajo un muestreo aleatorio simple (M.A.S.)

2. ESTIMADORES ÓPTIMOS DE CALIBRACIÓN

En la construcción de un estimador de calibración hay dos componentes básicas: una distancia Φ_s y un conjunto de restricciones. La distancia Ji-cuadrado, $\Phi_s = \sum_{k \in s} (w_k - d_k)^2 / (d_k q_k)$, es la más usada donde los factores q_k , ponderaciones en la distancia, no están correlacionados con $d_k = 1/\pi_k$ (usualmente se toma $q_k = 1$).

2.1. Construcción. La cercanía a los pesos originales d_k puede ser medida por un distancia $G(\cdot) = G(\frac{w_k}{d_k})$ que debe ser estrictamente no negativa y convexa, tal que $G(1) = G'(1) = 0$ y $G''(1) = 0$, por tanto la distancia total en toda la muestra será:

$$(2.1) \quad \sum_s d_k G\left(\frac{w_k}{d_k}\right)$$

Para hallar los nuevos pesos, el problema se reduce a minimizar (3.1) sujeto a la ecuación de calibración

$$(2.2) \quad \sum_s w_k U(x_k) = \sum_U U(x_k)$$

Utilizando un multiplicador de Lagrange, tenemos

$$(2.3) \quad \omega = \sum_s d_k G\left(\frac{w_k}{d_k}\right) - \lambda \left(\sum_s w_k U(x_k) - \sum_U U(x_k) \right)$$

Derivando (3.3) corespecto a w_k e igualando a cero

$$(2.4) \quad \frac{\partial \omega}{\partial w_k} = \sum_s d_k \frac{\partial G\left(\frac{w_k}{d_k}\right)}{\partial w_k} \left(\frac{1}{d_k}\right) - \lambda \left(\sum_s U(x_k) \right) = 0$$

Haciendo $g(\cdot) = \partial G(\cdot)$, (3.4) es igual a

$$(2.5) \quad \sum_s g \frac{w_k}{d_k} - \sum_s \lambda U(x_k) = 0$$

En este paso es necesario definir una función $F(\cdot)$, tal que $F(\cdot) = g^{-1}(\cdot) \ni F(g(t)) = t$, por lo tanto

$$(2.6) \quad F\left(g\left(\frac{w_k}{d_k}\right)\right) = F(\lambda U(x_k))$$

Lo que nos guía al valor de los nuevos pesos

$$(2.7) \quad w_k = d_k F(\lambda U(x_k))$$

2.1.1. Distancias $G(\cdot)$. En general hay varios tipos de distancias que pueden utilizarse en la construcción de un estimador de calibración, pero todas guían asintóticamente al mismo estimador (cabe resaltar que la minimización de las distancias no es la única forma de llegar al estimador de calibración). Las más utilizadas son las siguientes:

- *Método lineal*: $G(x) = (1/2)(x - 1)^2$; $x \in \mathcal{R}$; $F(u) = 1 + u$
- *Método multiplicativo o razón de Raking*: $G(x) = x \log(x) - x + 1$; $x >$; $F(u) = \exp(u)$
- *Método truncado lineal* Fijando dos constantes L y U y restringiendo el rango de los pesos resultantes a este intervalo con la distancia lineal

El último método se utiliza para evadir los pesos extremos o negativos, que se pueden eliminar con una buena escogencia de L y U .

En general utilizaremos la distancia *Ji cuadrado*, o *lineal*, que calcula la distancia, en toda la muestra, de los nuevos pesos a los pesos clásicos como:

$$(2.8) \quad \phi_s = \sum_s d_k G\left(\frac{w_k}{d_k}\right) = \frac{1}{2} \sum_s \frac{(w_k - d_k)^2}{d_k}$$

Ésta permite llegar a

$$(2.9) \quad w_k = d_k \lambda U(x) + d_k$$

y reemplazando en la ecuación de calibración (2.2)

$$(2.10) \quad \sum_s d_k \lambda U^2(x_k) + \sum_s d_k U(x_k) = \sum_U U(x_k)$$

y el multiplicador de Lagrange se resuelve como

$$(2.11) \quad \lambda = \frac{\sum_U U(x_k) - \sum_s d_k U(x_k)}{\sum_s d_k U^2(x_k)}$$

Así, se llega al estimador óptimo de calibración para cualquier función $U(x)$

$$(2.12) \quad \sum_s w_k y_k = \sum_s d_k \lambda U(x_k) y_k + \sum_s d_k y_k$$

Y reemplazando λ

$$(2.13) \quad \hat{Y}_{opt} = \hat{Y}_\pi + \left(\sum_U U(x_k) - \sum_U d_k U(x_k) \right) \hat{\mathbf{B}}$$

con

$$(2.14) \quad \hat{\mathbf{B}} = \left(\sum_s d_k U^2(x_k) \right)^{-1} \left(\sum_s d_k U(x_k) y_k \right)$$

Los estimadores óptimos de calibración se han estudiado y profundizado en [3] bajo un modelo asistido. Para motivar las condiciones de optimalidad se utiliza un modelo de superpoblación semiparamétrica como en (1.1). Se consideraron los estimadores de calibración para el total poblacional \mathbf{Y} usando:

1. Una distancia Ji-cuadrado con los factores de peso satisfaciendo $q_i > q$ para alguna constante $q > 0$ y $N^{-1} \sum_{i=1}^N q_i^2 = O(1)$
2. Una sola restricción, dada por una reducción de dimensión $u_i = u(\mathbf{x}_i, \theta)$, como en (2.2), donde la forma funcional $u(\cdot, \cdot)$ puede ser arbitraria.

Algunos de los resultados más importantes en [3] pueden ser resumidos como sigue:

- Sea $\hat{\theta} = (\sum_{i \in s} d_i q_i \mathbf{x}_i \mathbf{x}_i')^{-1} \sum_{i \in s} d_i q_i \mathbf{x}_i y_i$. Si se usa $u_i = \mathbf{x}_i' \theta$ como variable de calibración en (2.2) el estimador de calibración resultante es idéntico al estimador convencional de calibración que *Deville y Särndal* explican en [1]. Por tanto la clase de estimadores usados en [3] es muy general pues incluye al estimador original como un caso particular.
- Para cualquier estimador consistente de θ tal que $\hat{\theta} = \theta + o_p(1)$, si se reemplaza θ por $\hat{\theta}$, en la restricción (2.2), el estimador de calibración resultante no cambia asintóticamente.

3. DOS FORMAS DE CALCULAR LAS DISTANCIAS OTOGONALES Y LA REGRESIÓN CLÁSICA

Se presentan dos formas para calcular el ajuste entre x y y del tipo $y = x' \beta$ bajo el modelo (1.1) y bajo una muestra aleatoria simple.

3.1. El estimador de regresión simple. Bajo un modelo llamado de regresión común simple, Sarndal(1992), el ajuste de este modelo a la población finita se da estimando α y β como:

$$(3.1) \quad a = \bar{y} - b \bar{x}$$

y

$$(3.2) \quad \beta = \frac{\sum_s (x_k - \bar{x}_s)(y_k - \bar{y}_s)}{\sum_s (x_k - \bar{x}_s)^2}$$

Por tanto el estimador de $Y = \sum_U y_k$ es

$$(3.3) \quad \hat{Y}_{rs} = \hat{Y}_\pi + \left\{ \sum_{i \in U} \mathbf{x}_i - \sum_{i \in s} \frac{\mathbf{x}_i}{\pi_i} \right\} \hat{\mathbf{B}}$$

con

$$(3.4) \quad \hat{\mathbf{B}} = \left\{ \frac{\sum_{i \in s} \frac{\mathbf{x}_i y_i}{\pi_i}}{\sum_{i \in s} \frac{\mathbf{x}_i^2}{\pi_i}} \right\}$$

Éste es un estimador de calibración clásico con una sola variable auxiliar, \mathbf{X} es una matriz particionada tal que $\mathbf{x} = (1 \mid x)$ y $U(x) = \mathbf{x}$

3.2. Ejes mayores. Usamos el principio de mínimos cuadrados y minimizamos la suma cuadrados de las distancias perpendiculares d_i^2 de los puntos (x_i, y_i) a línea de regresin. La función que deseamos minimizar es:

$$(3.5) \quad E = \sum_{i=1}^n \{(x_i - X_i)^2 (y_i - Y_i)^2\}$$

donde (X_i, Y_i) son las coordenadas ajustadas que se desean encontrar, por supuesto se trabaja bajo la siguiente función:

$$f_i = a + b x_i - y_i = 0$$

Así, no basta con encontrar simplemente una pareja de (X, Y) sino que éstas deben descansar sobre una línea recta, utilizando *un multiplicador de Lagrange*, se debe encontrar:

$$F = E + \sum_{i=1}^n \lambda_i f_i$$

Luego, igualamos el conjunto de derivadas parciales a cero y resolvemos el conjunto de ecuaciones:

$$\frac{\partial F}{\partial X_i} = \frac{\partial F}{\partial Y_i} = \frac{\partial F}{\partial a} = \frac{\partial F}{\partial b} = 0$$

Resolviendo el sistema se llega a la solución:

$$(3.6) \quad a = \bar{y} - b\bar{x}$$

y

$$(3.7) \quad b = \frac{V - U + \sqrt{(V - U)^2 + 4(C)^2}}{2C}$$

donde:

$$\begin{aligned} U &= \text{var}(X) \\ V &= \text{var}(Y) \\ C &= \text{cov}(X, Y) \end{aligned}$$

Entonces por (2.13) el estimador de $Y = \sum_U y_k$ es

$$(3.8) \quad \hat{Y}_{em} = \hat{Y}_\pi + \left\{ \sum_{i \in U} (a + b\mathbf{x}_i) - \sum_{i \in S} \frac{(a + b\mathbf{x}_i)}{\pi_i} \right\} \hat{\mathbf{B}}$$

con

$$(3.9) \quad \hat{\mathbf{B}} = \left\{ \begin{array}{l} \sum_{i \in S} \frac{(a + b\mathbf{x}_i)y_i}{\pi_i} \\ \sum_{i \in S} \frac{(a + b\mathbf{x}_i)^2}{\pi_i} \end{array} \right\}$$

Que es un estimador de calibración con $U(x) = a + bx$

3.3. Ejes mayores reducidos. Se minimiza la suma de el rea de los rectangulos definidos entre las distancias perpendiculares entre los puntos observados y la línea de regresión. Se debe minimizar la distancia definida por ΔX y ΔY , por tanto la función a minimizar es:

$$(3.10) \quad E = \sum_{i=1}^n (x_i - X_i)(y_i - Y_i)$$

Las constantes permanecen iguales ($y_i = a + bx_i$). El método de los multiplicadores de Lagrange guía al siguiente sistema $2n + 2$ de ecuaciones:

$$\frac{\partial F}{\partial X_i} = \frac{\partial F}{\partial Y_i} = \frac{\partial F}{\partial a} = \frac{\partial F}{\partial b} = 0$$

Resolviendo el sistema se llega a la solución:

$$(3.11) \quad a = \bar{y} - b\bar{x}$$

y

$$(3.12) \quad b = \sqrt{V/U} = \frac{\sigma y}{\sigma x}$$

Entonces por (2.13) el estimador de $Y = \sum_U y_k$ es

$$(3.13) \quad \hat{Y}_{emr} = \hat{Y}_\pi + \left\{ \sum_{i \in U} (a + b\mathbf{x}_i) - \sum_{i \in s} \frac{(a + b\mathbf{x}_i)}{\pi_i} \right\} \hat{\mathbf{B}}$$

con

$$(3.14) \quad \hat{\mathbf{B}} = \left\{ \begin{array}{l} \sum_{i \in s} \frac{(a + b\mathbf{x}_i)y_i}{\pi_i} \\ \sum_{i \in s} \frac{(a + b\mathbf{x}_i)^2}{\pi_i} \end{array} \right\}$$

Que es un estimador de calibración con $U(x) = a + bx$.

Luego tenemos tres formas de trazar las líneas de regresión para la población que se pueden observar en la figura 1.

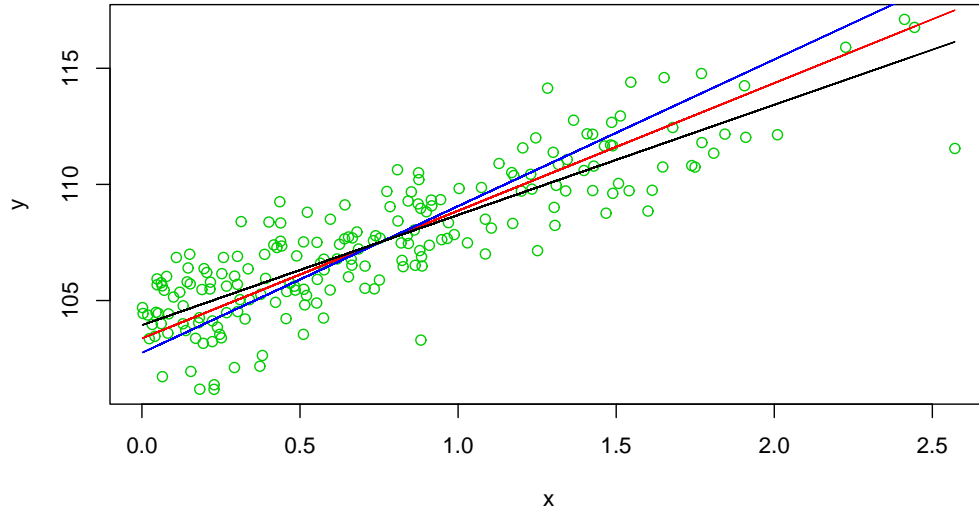


FIGURA 1. *Comportamiento de las líneas de regresión para la misma población*

4. SIMULACIÓN

En esta sección se realiza un estudio limitado por medio de una simulación, con el fin de tener un acercamiento a la optimalidad de los estimadores propuestos y desarrollados en este trabajo para un diseño (MAS).

Se simuló un total poblacional $N = 5000$ de un modelo de superpoblación, ξ . Se simuló una población con la siguiente estructura de dispersión, así:

- El modelo P , es un modelo de regresión lineal con estructura de varianza no-homogenea, así $y_i = \beta_0 + \beta_1 x_i + x_i \varepsilon_i$, para todo $i = 1, 2, \dots, N$.

En [3] se muestra que existen ciertas condiciones para que un diseño muestral sea regular¹, por tanto distribuciones de colas pesadas como la distribución log-normal y la distribución gamma con parámetros de escala muy grandes, no podrán ser usadas para generar la información auxiliar (*x-variables*). Así que se generan los valores de x de una distribución gamma con parámetro de forma 1 y parámetro de escala 2. Esta variable toma valores no negativos y es sesgada a la derecha, esto es muy común en aplicaciones reales de encuestas por muestreo [3].

El valor del parámetro β_1 se fijó en uno y el valor de β_0 se fijó convenientemente tal que $y_i > 0$ para ambos modelos. Los ε_i son independientes e idénticamente distribuidos como $N(0, \sigma_0^2)$. Se dispuso de cuatro valores para σ_0^2 tal que el coeficiente de correlación, ρ , en la población finita, entre y y x para el modelo P , fueran 0.9, 0.8, 0.7, y 0.6, respectivamente.

En cada corrida de la simulación se tomó una muestra aleatoria simple de $n = 500$. Los parámetros (β_0, β_1) se estimaron usando 1) mínimos cuadrados ponderados 2) ejes mayores y 3) ejes mayores reducidos para el modelo P . Así, se calcularon los siguientes estimadores para el total poblacional \mathbf{Y}

- $\hat{\mathbf{Y}}_\pi$ El estimador de Horvitz-Thompson para MAS.
- $\hat{\mathbf{Y}}_{Cal}$ El estimador de calibración convencional propuesto por [2] (3.3).
- $\hat{\mathbf{Y}}_{Opt1}$ El estimador de óptimo de calibración propuesto en (3.7).
- $\hat{\mathbf{Y}}_{Opt2}$ El estimador de óptimo de calibración propuesto en (3.11).

El proceso se repitió $B = 1000$ veces. La simulación fue programada en el software libre R v 2.1.1.². En la simulación, el desempeño de un estimador $\hat{\mathbf{Y}}$ fue evaluado usando su sesgo relativo, SR y su eficiencia relativa, ER , definidas por [3], como:

$$(4.1) \quad \mathbf{SR} = B^{-1} \sum_{b=1}^B \frac{\hat{\mathbf{Y}}_b - \mathbf{Y}}{\mathbf{Y}}$$

$$(4.2) \quad \mathbf{ER} = \frac{ECM(\hat{\mathbf{Y}}_{\pi^*})}{ECM(\hat{\mathbf{Y}})},$$

donde

$$(4.3) \quad ECM(\hat{\mathbf{Y}}) = B^{-1} \sum_{b=1}^B (\hat{\mathbf{Y}}_b - \mathbf{Y})^2$$

y $\hat{\mathbf{Y}}_b$ se calculó en la b -ésima muestra simulada. Como se puede notar el estimador de Horvitz-Thompson, $\hat{\mathbf{Y}}_\pi$, fue utilizado como línea base de comparación. Grandes valores para $ER (> 1)$ representan alta eficiencia del estimador $\hat{\mathbf{Y}}$ en comparación al estimador $\hat{\mathbf{Y}}_{\pi^*}$.

¹Un diseño muestral es regular si satisface: (i) $\max_{i \in s} n d_i = O(1)$, (ii) La distribución asintótica del estimador HT es normal

ρ	Modelo P			
	\hat{Y}_π	\hat{Y}_{Cal}	\hat{Y}_{Opt1}	\hat{Y}_{Opt2}
0.6	1.0	1.49	1.17	0.56
0.7	1.0	1.88	1.60	0.98
0.8	1.0	2.71	2.46	1.88
0.9	1.0	5.24	5.07	4.62

TABLA 1. *Eficiencia relativa de los estimadores para el modelo P*

En términos de sesgo relativo SR , todos los valores correspondientes son menores del 1.5 por ciento. Aunque el estimador \hat{Y}_{Opt2} tiene un sesgo aún mucho menor. En general cuando la correlación aumenta, el sesgo decrece junto con el MSE y la eficiencia relativa crece en ambos modelos. Se nota que los estimadores óptimos creados funcionan bien cuando la correlación es mayor a 0,8, siendo el estimador de los ejes mayores reducidos el más deficiente cuando no se alcanza dicha medida.

REFERENCIAS

- [1] Särndal, C.E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling* Springer, New York.
- [2] Deville, J.C. and Särndal, C.E. (1992) *Calibration Estimators in Survey Sampling*. Journal of the American Statistical Association, 87, 376-382.
- [3] Wu, C. (2002) *Optimal Calibration Estimators in Survey Sampling*. Working Paper 2002-01. Department of Statistics and Actuarial Science, University of Waterloo, Canada.
- [4] Wu, C., and Sitter, R.R. (2001) *A Model Calibration Approach to Using Complete Auxiliary Information From Survey Data*. Journal of the American Statistical Association, 96, pp.185-193

CALLE 5 NO. 12-25 BL 5 APTO 603, BOGOTÁ, COLOMBIA
 E-mail address: hagutierrezro@unal.edu.co
 E-mail address: psirusteam@yahoo.com