UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2009, Vol. 2, No. 1

Un nuevo estimador muestral de regresión vía residuos ortogonales derivados del análisis de componentes principales

A New Sampling Estimator using Orthogonal Residuals from Principal Components Analysis

Jimmy Rico Bermúdez[a]
Jimmy.Rico@nielsen.com

Resumen

Los estimadores de regresión son herramientas que emplean técnicas estadísticas propias como el análisis de regresión para aprovechar la información auxiliar disponible. En éste documento se presentan todas las herramientas teóricas necesarias para proponer un nuevo estimador de regresión ortogonal, para el cual el ajuste realizado no sea obtenido a partir de la teoría de mínimos cuadrados y en su lugar, éste se apoye en la construcción de componentes principales que por su naturaleza minimizan las distancias ortogonales de cada uno de los puntos de la nube de observaciones a la recta que recoge la mayor inercia.

***Palabras clave*:** estimador de regresión, información auxiliar, componentes principales, linealización de Taylor.

Abstract

Regression estimators are tools that employ statistics techniques such as regression analysis in order to gain in efficiency by means of the available auxiliary information. This paper presents the theoretical approach that yields to the proposal of a new orthogonal regression estimator for which the fit is not based in the theory of classical least squares, but instead, it is based in the theory of principal components which minimizes the orthogonal distances from each point of the scatter plot to the line that incorporates most of the inertia.

***Key words*:** auxiliary information, regression estimator, principal components, Taylor linearization.

[a]Ejecutivo estadístico Senior. The Nielsen Company.

1. El estimador de regresión ortogonal

1.1. Supuestos sobre el modelo

El supuesto sobre el cual se construye el estimador de regresión ortogonal propuesto en el presente artículo consiste en que la variable de estudio Y y la información auxiliar X [1] determinan una nube de puntos que siguen un modelo que tiene las siguientes características:

1. $y_1, \ldots, y_N$ son realizaciones de las variables aleatorias independientes $Y_1, \ldots, Y_N$.
2. $E(y_k) = \beta_1 + \beta_2 x_k \ \forall \ k = 1, \ldots, N$
3. $V(y_k) = \sigma^2$

En otras palabras, el modelo bajo el cual se supone se rigen las variables X y Y resulta homocedástico con intercepto.

1.2. Construcción del estimador de regresión ortogonal

Sea S la matriz que contiene las N observaciones tomadas de una población correspondientes a la variable de interés de estudio Y y la variable auxiliar para la estimación X así:

$$S = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_N & y_N \end{bmatrix}$$

Luego la matriz $\tilde{S}$ corresponde a la matriz de las N observaciones centradas de tal forma que:

$$\tilde{S} = \begin{bmatrix} \frac{x_1 - \bar{x}_U}{\sqrt{N}} & \frac{y_1 - \bar{y}_U}{\sqrt{N}} \\ \frac{x_2 - \bar{x}_U}{\sqrt{N}} & \frac{y_2 - \bar{y}_U}{\sqrt{N}} \\ \vdots & \vdots \\ \frac{x_N - \bar{x}_U}{\sqrt{N}} & \frac{y_N - \bar{y}_U}{\sqrt{N}} \end{bmatrix}$$

de ésta manera el producto $\tilde{S}^t \tilde{S}$ corresponde a la matriz a diagonalizar Σ, definida de la siguiente forma:

[1] Las características de la información auxiliar son:

- Se encuentra disponible para todos los elementos en la población de estudio.
- Está altamente correlacionada con la variable de estudio.

$$\Sigma = \begin{bmatrix} \frac{\sum_{i=1}^{N}(x_i-\bar{x}_U)^2}{N} & \frac{\sum_{i=1}^{N}(x_i-\bar{x}_U)(y_i-\bar{y}_U)}{N} \\ \frac{\sum_{i=1}^{N}(x_i-\bar{x}_U)(y_i-\bar{y}_U)}{N} & \frac{\sum_{i=1}^{N}(y_i-\bar{y}_U)^2}{N} \end{bmatrix}$$

en la que cada una de las entradas puede ser expresada como función de los totales t_x, t_{x^2} y N así:

$$\frac{\sum_{i=1}^{N}(x_i-\bar{x}_U)^2}{N} = \frac{\sum_{i=1}^{N}(x_i^2-2x_i\bar{x}_U+\bar{x}_U^2)}{N} = \frac{\sum_{i=1}^{N}(x_i^2)}{N} - \frac{\sum_{i=1}^{N}(2x_i\bar{x}_U)}{N} + \frac{\sum_{i=1}^{N}(\bar{x}_U^2)}{N}$$
$$= \frac{t_{x^2}}{N} - \frac{2\bar{x}_U N\bar{x}_U}{N} + \frac{N\bar{x}_U^2}{N} = \frac{t_{x^2}}{N} - \Big(\frac{1}{N}\sum_{i=1}^{N}x_i\Big)^2 = \frac{t_{x^2}}{N} - \frac{t_x^2}{N^2}$$

y de igual forma se puede verificar la siguiente equivalencia para Y:

$$\frac{\sum_{i=1}^{N}(y_i-\bar{y}_U)^2}{N} = \frac{t_{y^2}}{N} - \frac{t_y^2}{N^2}$$

finalmente la expresión faltante puede ser expresada en términos de los totales t_x, t_y, t_{xy} y N de la siguiente manera:

$$\frac{\sum_{i=1}^{N}(x_i-\bar{x}_U)(y_i-\bar{y}_U)}{N} = \frac{\sum_{i=1}^{N}x_iy_i}{N} - \frac{\sum_{i=1}^{N}x_i\bar{y}_U}{N} - \frac{\sum_{i=1}^{N}\bar{x}_Uy_i}{N} + \frac{\sum_{i=1}^{N}\bar{x}_U\bar{y}_U}{N}$$
$$= \frac{t_{xy}}{N} - 2\bar{x}_U\bar{y}_U + \bar{x}_U\bar{y}_U = \frac{t_{xy}}{N} - \frac{t_xt_y}{N^2}$$

Luego, para calcular los valores propios de la matriz Σ se deben encontrar las soluciones de $det(\Sigma - \lambda I) = 0$, es decir:

$$0 = det(\Sigma-\lambda I) = \Big[\Big(\frac{t_{x^2}}{N} - \frac{t_x^2}{N^2}\Big) - \lambda\Big]\Big[\Big(\frac{t_{y^2}}{N} - \frac{t_y^2}{N^2}\Big) - \lambda\Big] - \Big[\frac{t_{xy}}{N} - \frac{t_xt_y}{N^2}\Big]^2$$
$$= \lambda^2 - \lambda\Big[\frac{t_{x^2}}{N} - \frac{t_x^2}{N^2} + \frac{t_{y^2}}{N} - \frac{t_y^2}{N^2}\Big] + \Big(\frac{t_{x^2}}{N} - \frac{t_x^2}{N^2}\Big)\Big(\frac{t_{y^2}}{N} - \frac{t_y^2}{N^2}\Big)$$
$$- \Big(\frac{t_{xy}}{N} - \frac{t_xt_y}{N^2}\Big)^2$$

Ahora se definen:

$$S_x^2 = \frac{t_{x^2}}{N} - \frac{t_x^2}{N^2},\ S_y^2 = \frac{t_{y^2}}{N} - \frac{t_y^2}{N^2}\ y\ S_{xy} = \frac{t_{xy}}{N} - \frac{t_x t_y}{N^2}$$

de tal manera que el valor propio máximo de la matriz Σ está dado por:

$$\lambda_{MAX} = \frac{(S_x^2 + S_y^2) + \sqrt{(S_x^2 - S_y^2)^2 + 4S_{xy}^2}}{2}$$

Ahora, el vector propio asociado al valor propio λ_{MAX} que resulta de hallar la solución de la ecuación $\Sigma\vec{v} = \lambda_{MAX}\vec{v}$ está dado por:

$$(S_x^2 - \lambda_{MAX})x_k^* + S_{xy}y_k^* = 0$$

con

$$x_k^* = \frac{x_k - \bar{x}_U}{\sqrt{N}}$$
$$y_k^* = \frac{y_k - \bar{y}_U}{\sqrt{N}}$$

que se puede expresar en términos de las observaciones x_k y y_k originales de la siguiente forma:

$$\begin{aligned} 0 &= (S_x^2 - \lambda_{MAX})\frac{(x_k - \bar{x}_U)}{\sqrt{N}} + S_{xy}\frac{(y_k - \bar{y}_U)}{\sqrt{N}} \\ &= (S_x^2 - \lambda_{MAX})(x_k - \bar{x}_U) + S_{xy}(y_k - \bar{y}_U) \end{aligned}$$

luego:

$$y_k = \frac{(\lambda_{MAX} - S_x^2)(x_k - \bar{x}_U)}{S_{xy}} + \bar{y}_U$$

$$y_k = \bar{y}_U - \frac{(\lambda_{MAX} - S_x^2)}{S_{xy}}\bar{x}_U + \frac{(\lambda_{MAX} - S_x^2)}{S_{xy}}x_k$$

así resultan los parámetros de la recta, definidos como:

$$B_1 = \bar{y}_U - B_1\bar{x}_U$$
$$B_2 = \frac{(\lambda_{MAX} - S_x^2)}{S_{xy}}$$

La idea principal es usar la información auxiliar para formar un conjunto de N valores cercanos a y denotados por y_k^0 con $k = 1, 2, \ldots, N$ de tal forma que éstos resultan como una combinación lineal de los valores conocidos para x_k así:

$$y_k^0 = A_1 + A_2 x_k$$

el total desconocido en la población de la variable Y puede ser escrito como:

$$t_y = \sum_U y_k = \sum_U y_k^0 + \sum_U (y_k - y_k^0) = \sum_U y_k^0 + \sum_U D_k$$

las diferencias desconocidas D_k son estimadas por $\frac{D_k}{\pi_k}$ para obtener:

$$\hat{t}_y = \sum_U y_k^0 + \sum_s \frac{D_k}{\pi_k}$$

y luego de reemplazar el valor de y_k^0 se obtiene la siguiente ecuación:

$$\begin{aligned}\hat{t}_y &= \sum_U y_k^0 + \sum_s \frac{y_k - y_k^0}{\pi_k} = \sum_U (A_1 + A_2 x_k) + \sum_s \frac{(y_k - (A_1 + A_2 x_k))}{\pi_k} \\ &= \hat{t}_{y\pi} + A_1(N - \hat{N}) + A_2(t_x - \hat{t}_{x\pi})\end{aligned}$$

en la cual los valores A_1 y A_2 propuestos en éste trabajo corresponden a los coeficientes $\hat{B}_1$ y $\hat{B}_2$ dados por:

$$\hat{B}_1 = \tilde{y}_U - \hat{B}_1 \tilde{x}_U \tag{1}$$

$$\hat{B}_2 = \frac{(\hat{\lambda}_{MAX} - \hat{S}_x^2)}{\hat{S}_{xy}} \tag{2}$$

obtenidos a partir de la técnica de componentes principales, donde:

$$\tilde{x}_U = \frac{\sum_{i=1}^{n} x_i}{\hat{N}} \quad y \quad \tilde{y}_U = \frac{\sum_{i=1}^{n} y_i}{\hat{N}}$$

y las cantidades $\hat{S}_x$, $\hat{S}_y$, $\hat{S}_{xy}$ y por consiguiente el valor propio $\hat{\lambda}_{MAX}$ se obtienen reemplazando cada uno de los totales N, t_x, t_y, t_{x^2}, t_{y^2} y t_{xy} que los componen por su respectivo π-estimador.

De esta forma se introduce formalmente el estimador de regresión ortogonal denotado por $\hat{t}_{yort}$:

$$\hat{t}_{yort} = \hat{t}_{y\pi} + \hat{B}_1(N - \hat{N}) + \hat{B}_2(t_{x_1} - \hat{t}_{x_1\pi}) \tag{3}$$

1.3. Estimación de la varianza del estimador de regresión ortogonal

Para determinar la forma que toma la estimación de la varianza del estimador de regresión ortogonal se aplica la técnica de linealización de Taylor teniendo en cuenta que el estimador que resulta expresado como:

$$\hat{t}_{yort} = f(\hat{N}, \hat{t}_{x\pi}, \hat{t}_{y\pi}, \hat{t}_{x^2\pi}, \hat{t}_{y^2\pi}, \hat{t}_{xy\pi})$$

es función **no lineal** de seis totales.

La idea general de la técnica es generar un pseudo-estimador lineal $\hat{\theta}_0$ para el que la varianza es más simple de calcular y la expresión $V(\hat{\theta}_0)$ es la aproximación de la varianza del estimador $\hat{\theta}$ y en consecuencia se obtiene $\hat{V}(\hat{\theta})$.

La técnica para encontrar $\hat{\theta}_0$ consiste en la aproximación de Taylor de primer orden de la función f, expandiendo la misma alrededor del punto $(t_1, \ldots, t_q)$ de tal forma que:

$$\hat{\theta} = \hat{\theta}_0 = \theta + \sum_{j=1}^{q} a_j(\hat{t}_{j\pi} - t_j)$$

calculando los coeficientes a_j como sigue:

$$a_j = \left.\frac{\partial f}{\partial \hat{t}_{j\pi}}\right|_{(\hat{t}_{1\pi},\ldots,\hat{t}_{q\pi})=(t_1,\ldots,t_q)}$$

Así para muestras de tamaño grande los π-estimadores $(\hat{t}_{1\pi}, \ldots, \hat{t}_{q\pi})$ toman valores cercanos a los totales poblacionales $(t_1, \ldots, t_q)$ con una alta probabilidad, lo cual implica que tanto el sesgo como la varianza del parámetro pueden ser también aproximados por las cantidades obtenidas para la estadística lineal $\hat{\theta}_0$.

A continuación se presenta el cálculo de la derivada correspondiente a $\hat{N}$:

$$a_1 = \frac{\partial t_{yort}}{\partial \hat{N}} = \Big[- \frac{\hat{t}_{y\pi}}{\hat{N}^2} + \frac{1}{\left(\hat{t}_{xy\pi} - \frac{\hat{t}_{y\pi}\hat{t}_{x\pi}}{\hat{N}}\right)^2} \Big[\Big[- \frac{\hat{t}_{x\pi}\hat{t}_{x^2\pi}}{\hat{N}^2} + \frac{2\hat{t}_{x\pi}^3\hat{N}}{\hat{N}^4} + \frac{\hat{t}_{x\pi}\hat{t}_{x^2\pi}}{2f^2} - \frac{2\hat{t}_{x\pi}^3 f}{2\hat{N}^4} +$$
$$\frac{\hat{t}_{x\pi}\hat{t}_{y^2\pi}}{2\hat{N}^2} - \frac{2\hat{t}_{y\pi}^2\hat{t}_{x\pi}\hat{N}}{2\hat{N}^4} - \frac{\hat{t}_{x\pi}}{2}\Big(\Big[\frac{\hat{t}_{x^2\pi}}{\hat{N}} - \frac{\hat{t}_{x\pi}^2}{\hat{N}^2} - \frac{\hat{t}_{y^2\pi}}{\hat{N}} + \frac{\hat{t}_{y\pi}^2}{\hat{N}^2} \Big]^2 + 4\Big[\frac{\hat{t}_{xy\pi}}{\hat{N}} - \frac{\hat{t}_{y\pi}\hat{t}_{x\pi}}{\hat{N}^2} \Big]^2 \Big)^{-\frac{1}{2}} \Big[2\Big(\frac{\hat{t}_{x^2\pi}}{\hat{N}} -$$

$$\frac{\hat{t}_{x\pi}^2}{\hat{N}^2}-\frac{\hat{t}_{y^2\pi}}{\hat{N}}+\frac{\hat{t}_{y\pi}^2}{\hat{N}^2}\Big)\Big(-\frac{\hat{t}_{x^2\pi}}{\hat{N}^2}+\frac{2\hat{t}_{x\pi}^2\hat{N}}{\hat{N}^4}+\frac{\hat{t}_{y^2\pi}}{\hat{N}^2}-\frac{2\hat{N}\hat{t}_{y\pi}^2}{\hat{N}^4}\Big)+8\Big(\frac{\hat{t}_{xy\pi}}{\hat{N}}-\frac{\hat{t}_{y\pi}\hat{t}_{x\pi}}{\hat{N}^2}\Big)\Big(-\frac{\hat{t}_{xy\pi}}{\hat{N}^2}+\frac{2\hat{t}_{y\pi}\hat{t}_{x\pi}\hat{N}}{\hat{N}^4}\Big)\Big]\Big]\Big(\hat{t}_{xy\pi}-\frac{\hat{t}_{y\pi}\hat{t}_{x\pi}}{\hat{N}}\Big)-\Big[\hat{t}_{x\pi}\Big(\frac{\hat{t}_{x^2\pi}}{\hat{N}}-\frac{\hat{t}_{x\pi}^2}{\hat{N}^2}\Big)-\frac{\hat{t}_{x\pi}}{2}\Big(\frac{\hat{t}_{x^2\pi}}{\hat{N}}-\frac{\hat{t}_{x\pi}^2}{\hat{N}^2}+\frac{\hat{t}_{y^2\pi}}{\hat{N}}-\frac{\hat{t}_{y\pi}^2}{\hat{N}^2}\Big)-\hat{t}_{x\pi}\sqrt{\Big(\frac{\hat{t}_{x^2\pi}}{\hat{N}}-\frac{\hat{t}_{x\pi}^2}{\hat{N}^2}-\frac{\hat{t}_{y^2\pi}}{\hat{N}}+\frac{\hat{t}_{y\pi}^2}{\hat{N}^2}\Big)^2+4\Big(\frac{\hat{t}_{xy\pi}}{\hat{N}}-\frac{\hat{t}_{y\pi}\hat{t}_{x\pi}}{\hat{N}^2}\Big)^2}\Big]\Big(\frac{\hat{t}_{y\pi}\hat{t}_{x\pi}}{\hat{N}^2}\Big)\Big]\Big]\Big(N-\hat{N}\Big)-B_1$$

que luego de ser evaluada en: $(\hat{N},\hat{t}_{x\pi},\hat{t}_{y\pi},\hat{t}_{x^2\pi},\hat{t}_{y^2\pi},\hat{t}_{xy\pi})=(N,t_x,t_y,t_{x^2},t_{y^2},t_{xy})$ toma un valor de $=-B_1$

Los resultados para las derivadas restantes se presentan a continuación:

Para $t_{x\pi}$:

$$a_2=\frac{\partial t_{yort}}{\partial t_{x\pi}}\bigg|_{(\hat{N},\hat{t}_{x\pi},\hat{t}_{y\pi},\hat{t}_{x^2\pi},\hat{t}_{y^2\pi},\hat{t}_{xy\pi})=(N,t_x,t_y,t_{x^2},t_{y^2},t_{xy})}=-B_2$$

Para $t_{y\pi}$:

$$a_3=\frac{\partial t_{yort}}{\partial t_{y\pi}}\bigg|_{(\hat{N},\hat{t}_{x\pi},\hat{t}_{y\pi},\hat{t}_{x^2\pi},\hat{t}_{y^2\pi},\hat{t}_{xy\pi})=(N,t_x,t_y,t_{x^2},t_{y^2},t_{xy})}=1$$

para $t_{x^2\pi}$:

$$a_4=\frac{\partial t_{yort}}{\partial t_{x^2\pi}}\bigg|_{(\hat{N},\hat{t}_{x\pi},\hat{t}_{y\pi},\hat{t}_{x^2\pi},\hat{t}_{y^2\pi},\hat{t}_{xy\pi})=(N,t_x,t_y,t_{x^2},t_{y^2},t_{xy})}=0$$

para $t_{y^2\pi}$:

$$a_5=\frac{\partial t_{yort}}{\partial t_{y^2\pi}}\bigg|_{(\hat{N},\hat{t}_{x\pi},\hat{t}_{y\pi},\hat{t}_{x^2\pi},\hat{t}_{y^2\pi},\hat{t}_{xy\pi})=(N,t_x,t_y,t_{x^2},t_{y^2},t_{xy})}=0$$

y finalmente para $t_{xy\pi}$:

$$a_6=\frac{\partial t_{yort}}{\partial t_{xy\pi}}\bigg|_{(\hat{N},\hat{t}_{x\pi},\hat{t}_{y\pi},\hat{t}_{x^2\pi},\hat{t}_{y^2\pi},\hat{t}_{xy\pi})=(N,t_x,t_y,t_{x^2},t_{y^2},t_{xy})}=0$$

Con cada una de las derivadas se construye la transformada u_k, de la siguiente manera (Cassel et al. 1976):

$$u_k=\sum_{j=1}^{q}a_jy_{jk}$$

que para éste caso toma la forma:

$$u_k = y_k - B_1 - B_2 x_k$$

Que como se observa, no se distancia de la transformada u_k del modelo homocedástico con intercepto cuando la recta se construye vía mínimos cuadrados, donde finalmente la aproximación de la varianza de $\hat{t}_{yort}$ será definida por:

$$AV(\hat{t}_{yort}) = V(\hat{t}_{pseudoort}) = V\left(\sum\nolimits_s \frac{u_k}{\pi_k}\right) = \sum\sum\nolimits_U \Delta_{kl}\frac{u_k}{\pi_k}\frac{u_l}{\pi_l}$$

dado que u_k depende de los a_j y éstos últimos de los totales poblacionales desconocidos, es necesario trabajar con $\hat{u}_k$ determinado por $\hat{a}_j$, quién a su vez resulta de reemplazar dichos totales desconocidos por sus correspondientes $\pi - estimadores$.

Así $\hat{u}_k$ para el presente caso resulta como:

$$\hat{u}_k = y_k - \hat{B}_1 - \hat{B}_2 x_k$$

Obteniendo así la estimación de la varianza del estimador de regresión ortogonal propuesto, expresada por:

$$\hat{V}(\hat{t}_{yort}) = \hat{V}\left(\sum\nolimits_s \frac{\hat{u}_k}{\pi_k}\right) = \sum\sum\nolimits_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{\hat{u}_k}{\pi_k}\frac{\hat{u}_l}{\pi_l}$$

Si se observa nuevamente la forma que tienen cada uno de los u_k para t_{yort} resulta:

$$u_k = y_k - B_1 - B_2 x_k$$

esto es:

$$\begin{aligned} u_k &= y_k - B_1 - B_2 x_k \\ &= y_k - (B_1 + B_2 x_k) \\ &= y_k - y_k^0 \\ &= E_k \end{aligned}$$

en otras palabras, cada u_k necesario para calcular la aproximación de la varianza del estimador t_{yort} resultan ser las diferencias E_k obtenidas al trazar la recta que teniendo toda la población minimiza las distancias ortogonales de cada una de las observaciones de la nube de puntos a la misma.

de ésta forma la aproximación de la varianza de t_{yort} toma la forma:

$$AV(\hat{t}_{yort}) = \sum\sum\nolimits_{U} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \tag{4}$$

nótese que siguiendo el mismo razonamiento para $\hat{u}_k$ debe incluirse un subíndice m a la variable e_k puesto que:

$$\begin{aligned} \hat{u}_k &= y_k - \hat{B}_1 - \hat{B}_2 x_k \\ &= y_k - (\hat{B}_1 + \hat{B}_2 x_k) \\ &= y_k - \hat{y}_k \end{aligned}$$

en ésta instancia cada $\hat{y}_k$ que compone a $\hat{u}_k$ depende de la muestra m tomada de la población de estudio pues para cada una de ellas resultan $\hat{B}_j$ diferentes, luego cada $\hat{u}_k$ queda expresado como:

$$\hat{u}_k = e_{km} \tag{5}$$

y por ende, la expresión correspondiente a la estimación de la varianza puede ser fácilmente expresada en términos de los errores e_{km} así:

$$\hat{V}(\hat{t}_{yort}) = \sum\sum\nolimits_{s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_{km}}{\pi_k} \frac{e_{lm}}{\pi_l}$$

por último, y como resultado adicional, una vez calculada la estimación de la Varianza se obtiene el correspondiente intervalo de confianza para el total de la variable Y de la siguiente manera:

$$\hat{t}_{yort} \pm z_{1-\frac{\alpha}{2}} [\hat{V}(\hat{t}_{yort})]^{\frac{1}{2}}$$

donde $z_{1-\frac{\alpha}{2}}$ es el $1 - \frac{\alpha}{2}$ cuantil de la distribución normal estándar.

2. Lógica de programación

Algoritmo (Comparación de $\hat{t}_{yort}$ frente a algunos estimadores clásicos)

Entrada:

- K = Número de muestras a seleccionar.

- n = Tamaño de muestra.
- N = Tamaño de la Población de estudio.

Salida:

- $\bar{\hat{t}}$ = Promedio de las estimaciones del total para la variable y.
- $S^2_{\hat{t}}$ = Varianza de las estimaciones del total para y.
- $\bar{\hat{V}}_g$ = Promedio de las varianzas ponderadas del estimador del total para y.
- ECR_g = Tasa de cobertura empírica basada en la varianza ponderada.
- $\bar{\hat{V}}_{sim}$ = Promedio de las varianzas simples del estimador del total para y.
- ECR_{sim} = Tasa de cobertura empírica basada en la varianza simple.
- AV = Aproximación de la varianza del estimador para el total de y.

Iteración:

- Se selecciona una muestra de tamaño n empleando un diseño M.A.S.
- Para la muestra seleccionada se calculan los coeficientes de la componente principal $\hat{B}_1$ y $\hat{B}_2$ como se indica en (1) y (2).
- Se obtiene la estimación $\hat{t}_{yort}$ para el total de la variable y como se indica en (3).
- Se obtienen los valores de la transformada $\hat{u}_k$ como se indica en (5).
- Se calcula la varianza estimada del estimador ortogonal.
- Se establece el intervalo de confianza del 95 % con las estimaciones obtenidas.
- Se crea una variable z_k definida como sigue:

$$z_k = \begin{cases} 1, & \text{si el intervalo de confianza contiene al total poblacional de } y. \\ 0, & \text{en caso contrario.} \end{cases}$$

- El anterior procedimiento se repite K veces.

Para la implementación práctica del anterior algoritmo se hizo uso de la población MU281 que puede ser consultada en el apendice B del libro (Särndal et al. 1992) conformada por 281 municipios de Suecia. Con éste conjunto de datos se calculó el estimador ortogonal para $y = RMT85 \times 10^{-4}$: Dinero recibido por los municipios en impuestos, empleando como información auxiliar tanto la variable $x_1 = CS82$:

Número de curules del partido conservador en el consejo municipal como $x_2 =$ $SS82$: Número de curules del partido social demócrata en el consejo municipal.

El anterior procedimiento se ejecutó mediante un programa SAS versión 9.0 para el que se tomaron como datos iniciales:

- K = 5000.
- n = 100.
- N = 281.

Los resultados obtenidos se resumen en la siguiente tabla:

Comparación de $\hat{t}_{yort}$ frente a algunos estimadores clásicos

Estimador	$\hat{\bar{t}}$	$S^2_{\hat{t}}$	$\hat{V}_g$	ECR_g	$\hat{V}_{sim}$	ECR_{sim}	AV
$\hat{t}_{y\pi}$	5.31	0.204	-	-	0.203	93.6	0.204
$\hat{t}_{yra}(X_1)$	5.31	0.121	0.120	93.1	0.121	93.2	0.121
$\hat{t}_{yra}(X_2)$	5.31	0.141	0.141	93.9	0.141	93.8	0.142
$\hat{t}_{yreg}(X_1)$	5.30	0.119	0.115	93.1	0.114	92.5	0.116
$\hat{t}_{yreg}(X_2)$	5.30	0.119	0.118	93.9	0.116	93.4	0.117
$\hat{t}_{yreg}(X_1, X_2)$	5.31	0.054	0.052	93.2	0.050	92.5	0.052
$\hat{t}_{yort}(X_1)$	5.31	0.122	-	-	0.117	93.8	0.119
$\hat{t}_{yort}(X_2)$	5.29	0.120	-	-	0.114	93.4	0.115

Conclusiones

- El estimador de regresión ortogonal toma valores cercanos al parámetro poblacional (5.315), aunque con X_2 presenta un poco de sesgo en la estimación obtenida, de otra parte el estimador de regresión clásico tanto con X_1 como con X_2 se encuentra por debajo del parámetro poblacional y sólo empleando las dos variables auxiliares se acerca a su objetivo, lo que muestra un buen desempeño por parte del estimador propuesto en éste trabajo.
- Los valores para $S^2_{\hat{t}}$ son cercanos a los que presenta el estimador de regresión clásico, aunque éste último presenta una menor varianza en las estimaciones que se obtienen a partir de la técnica de mínimos cuadrados que sugiere una leve ventaja en éste aspecto del mismo frente al estimador de regresión ortogonal propuesto en éste documento.
- La varianza que presenta el estimador ortogonal frente a la que muestra el estimador de regresión clásico no refleja mayores diferencias para ninguno de los dos estimadores, esto sólo evidencia que las dos técnicas son eficaces en la búsqueda de precisión al estimar el total de una variable de interés.

- La confiabilidad estimada que presenta el estimador de regresión ortogonal frente a la que arroja la simulación para el estimador de regresión clásico sugiere una mejora en la cobertura del parámetro poblacional, aspecto que teniendo en cuenta las anteriores observaciones presenta al estimador de regresión ortogonal como una buena alternativa en la estimación del total de una variable de estudio.

Agradecimientos

Agradezco en primer instancia al gestor de este aporte, a la persona que con su incansable esfuerzo llenaba el muestreo de nuevos planteamientos, profundas críticas y valorables puntos de vista; paz en la tumba del profesor Leonardo Bautista. A mis padres, a Tatiana Escobar, amigos y compañeros otro gran agradecimiento pues su compañía, sus consejos y su ánimo le dan sentido y forma a estos pequeños logros que brinda una linda profesión y llenan mi vida de mucha satisfacción.

Recibido: 10 de febrero de 2008
Aceptado: 12 de mayo de 2009

Referencias

Bautista, L. (1998), *Diseños de muestreo estadístico*, Universidad Nacional de Colombia.

Cassel, C., Särndal, C. & Wretman, J. (1976), 'Some results on generalized difference estimation and generalized regression estimation for finite populations.', *Biometrika* **63**, 615–620.

Särndal, C., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer, New York.