UNIVERSIDAD DE GRANADA

FACULTAD DE CIENCIAS

Departamento de Estadística e Investigación Operativa

CONTRIBUCIONES A LAS TECNICAS INDIRECTAS DE ESTIMACION

TESIS DOCTORAL

Antonio Arcos Cebrián

GRANADA, 1995

UNIVERSIDAD DE GRANADA
FACULTAD DE CIENCIAS

Departamento de Estadística e Investigación Operativa

CONTRIBUCIONES A LAS TECNICAS INDIRECTAS DE ESTIMACION

TESIS DOCTORAL

Antonio Arcos Cebrián

GRANADA, 1995

UNIVERSIDAD DE GRANADA
FACULTAD DE CIENCIAS

CONTRIBUCIONES A LAS TECNICAS INDIRECTAS DE ESTIMACION

ANTONIO ARCOS CEBRIAN

TESIS DOCTORAL

Sección de Matemáticas

CONTRIBUCIONES A LAS TECNICAS INDIRECTAS DE ESTIMACION

Memoria que para optar al grado de Doctor en Ciencias, sección Matemáticas, presenta Antonio Arcos Cebrián.

Vº Bº
Director de Tesis:

Prof. Dr. D. Andrés Gonzalez Carmona

Vº Bº
Director de Tesis:

Prof. Dr. Dña. María del Mar Rueda García

DEPARTAMENTO DE ESTADISTICA E INVESTIGACION OPERATIVA

FACULTAD DE CIENCIAS

UNIVERSIDAD DE GRANADA

Deseo expresar mi sincero agradecimiento a los directores de este trabajo así como al Departamento de Estadística e Investigación Operativa, por el apoyo y ayuda recibidos.

Indice

Introducción.

Las técnicas indirectas de estimación en muestreo de poblaciones finitas tratan de mejorar, en la fase de estimación, la precisión de los estimadores directos que no utilizan más que la información que proporciona la observación de la variable de interés en la muestra selecionada, mediante el conocimiento de información suplementaria aportada por una o varias variables auxiliares.

La mayor parte de los trabajos referentes a estas técnicas indirectas de estimación centran su objetivo en mejorar la precisión en la estimación de forma directa del parámetro media poblacional (o total o proporción) de una variable de interés y a través de la información de una variable auxiliar x de la que concretamente es conocida su media poblacional o a través de la información que proporcionan varias variables auxiliares $x_1, \ldots, x_k$ para las que son conocidas sus respectivas medias poblacionales.

En este sentido, la primera parte de la memoria que presentamos se dedica al estudio de estas técnicas indirectas de estimación del parámetro media poblacional cuando la información auxiliar es múltiple y que hemos llamado **Mejora de estimadores múltiples de la media** en la que se ha hecho la suposición de trabajar con un muestreo aleatorio simple, puesto que es el diseño muestral más simple y sirve de base para comparar la eficiencia de diseños muestrales más complejos. Las ideas que se proponen para la construcción de estimadores pueden utilizarse también para otros diseños muestrales, dando lugar a nuevos estimadores que bajo ciertas condiciones mejorarán los respectivos estimadores simples de cada diseño muestral.

Esta primera parte la hemos dividido a su vez en tres: estimadores múltiples de razón, estimadores múltiples de producto y estimadores múltiples de

razón–producto, en función de la técnica utilizada de acuerdo con el signo de la correlación entre las variables auxiliares disponibles: todas con correlación positiva, todas con correlación negativa o con correlación de cualquier signo, respectivamente.

Dentro de los estimadores múltiples de razón presentamos un técnica nueva que nos proporciona un estimador que hemos llamado estimador **iterado de razón con repeticiones** y que esencialmente es una técnica de estimación que, partiendo de la estimación directa de la media de la variable de interés, selecciona mediante un método forward sólo aquellas variables de entre las disponibles que producen un aumento en la precisión de la estimación tipo razón de la media, pudiendo cada variable auxiliar ser selecciona más de una vez. El estimador así construido se puede calcular siempre de forma exacta, frente a estimadores conocidos como los dados por *Olkin* (1958) o *Rueda* y otros (1992), engloba a estimadores de razón que sólo utilizan una variables auxiliar, al estimador iterado de razón dado por *Rueda* (que sólo permite seleccionar cada variable una única vez) y al estimador directo, garantizando además ser más preciso que este último, cosa que no ocurría con el estimador de *Olkin*. Posteriormente hacemos una generalización al caso en que los valores α_i (en la justificación número de veces que la variable auxiliar x_i es utilizada) no sean enteros, obteniendo un estimador que hemos llamado estimador **iterado de razón con repeticiones insesgado** que es más preciso que el directo y que cualquiera de razón univariante e insesgado, y englobando al estimador dado por *Srivastava* (1967).

Con los estimadores múltiples de producto presentamos el estimador **producto condensado** que se determina eligiendo la combinación lineal de variables auxiliares que maximice la precisión del estimador tipo producto construido con ella. El estimador obtenido, que es igual de preciso para muestras grandes que el estimador dado por *Singh* (1967), es modificado posteriormente para hacerlo insesgado, aportando además estimaciones de la combinación lineal que proporciona el estimador óptimo.

Si la correlación entre la variable de interés y las variables auxiliares es de cualquier signo se combinan las técnicas de razón y producto para conseguir una estimación más precisa de la media. Como alternativa al estimador dado por *Rao* y *Mudholkar* (1967) definimos, dentro de los estimadores múltiples de razón–producto, el estimador **iterado de razón–producto**, construido mediante un método forward que, partiendo de la estimación directa, selecciona una única vez aquella variable auxiliar que más aumente la precisión en la

estimación tipo razón o producto que se pueda realizar con ella (dependiendo del signo de la correlación con la variable principal). El estimador resultante engloba y es más preciso que el estimador directo, que cualquier estimador de razón o de producto que se pueda construir con cualquier variable auxiliar. Además conseguimos hacerlo insesgado mediante una técnica jackknife y obtenemos el estimador **iterado de razón–producto insesgado** que además engloba como casos particulares al estimador dado por *Sukhatme* y otros (1984) y al estimador dado por *Singh* y *Biradar* (1992).

En la segunda parte de la memoria se trata la estimación del parámetro varianza poblacional bajo el título **Mejora de estimadores de la varianza**. En ella se estudian las técnicas de estimación para este parámetro distinguiendo cuando el número de variables auxiliares es uno o mayor, asumiendo en cualquier caso que la varianza de la variable o variables auxiliares es conocida y que nos sirve para intentar mediante las técnicas de estimación indirectas mejorar la precisión en la estimación de la varianza de la variable de interés.

Los estimador univariantes indirectos conocidos son el estimador de razón y regresión dados por *Isaki* (1983) y las posteriores modificaciones para mejorar el estimador de razón dado por *Isaki* debidas a *Prasad* y *Singh* (1990) (para aumentar la precisión) y a los mismos autores en 1992 para disminuir el sesgo. Al margen de otras contribuciones (adaptación del estimador de razón de *Isaki* a muestreo aleatorio simple, observación que el estimador propuesto por *Prasad* y *Singh* (1992) es esencialmente el estimador de regresión), presentamos dos técnicas nuevas: la **estimación de producto** y el **método repetido de sustitución**. Comprobamos como la estimación de producto tiene sentido ser planteada, al igual que en el caso de la estimación de producto de la media, y puede mejorar en precisión a la estimación directa y de razón, presentando la ventaja sobre la estimación de regresión de que el estimador resultante teórico se puede calcular siempre. El método repetido de sustitución, análogo al dado por *Srivastava* (1967) para la estimación de la media, proporciona un estimador insesgado que para muestras grandes es igual de preciso que el de regresión y presentando la ventaja sobre éste, en caso de "normalidad", de requerir sólo de información sobre la correlación entre las variables principal y auxiliar para ser computado exactamente además de ser más preciso que el estimador directo, el de razón y el de producto en cualquier tipo de población.

Los estimadores múltiples indirectos de la varianza son tratados en el último capítulo y son debidos a *Isaki* (1983). Él desarrolla las técnica de razón y regresión múltiples bajo la suposición de que el vector $(y, x_1, \ldots, x_k)$ tenga

los mismos momentos hasta el orden cuatro que una normal multivariante y que los coeficientes de correlación entre todas las variables sean iguales. Nosotros estudiamos la técnica de estimación de razón en una población cualquiera y en muestreo aleatorio simple. Además presentamos una técnica nueva, el **método de exponenciación** como alternativa a la estimación de razón y regresión múltiple, extensión del método repetido de sustitución univariante, que proporciona un estimador, en cualquier población, insesgado, que para muestras grandes es igual de preciso que el de regresión múltiple y presentando la ventaja sobre éste, bajo el modelo de *Isaki*, de requerir sólo de información sobre la correlación común a todas las variables para ser computado exactamente además de ser más preciso que el estimador directo, el de razón múltiple y el de razón o producto construido con cualquier variable auxiliar.

Parte I

Mejora de estimadores múltiples de la media.

Capítulo 1

Variables auxiliares con correlación positiva con la variable de interés: Estimadores múltiples de razón.

§1.1 Introducción.

En las encuestas por muestreo, es frecuente disponer de una o varias variables correladas con el carácter objeto de estudio, y la información que proporcionan estas variables auxiliares puede producir estimadores más precisos. Esta información auxiliar puede utilizarse bien para modificar el diseño muestral (estratificación de la población, selección de unidades con probabilidades desiguales, etc.) o bien para modificar los estimadores usuales (que sólo utilizan la información muestral de la variable principal) mediante estimadores indirectos.

La literatura de muestreo en poblaciones finitas es abundante en ejemplos en los cuales los métodos indirectos son usados para estimar medias y totales. Ya en el año 1802 *Laplace* utilizó un estimador tipo razón para estimar la población de Francia y han sido ampliamente considerados en los últimos cincuenta años.

Entre estos métodos, el que ha centrado el interés de los investigadores ha sido el de estimación de razón puesto que su error cuadrático medio es más pequeño que la varianza del estimador de expansión simple (si el coeficiente de correlación entre las variables es positivo y alto) y es más sencillo

computacionalmente que el método de regresión.

La investigación sobre estimación de razón ha seguido fundamentalmente dos caminos:

1. Construir estimadores tipo razón que disminuyan el sesgo como los estimadores insesgados y cuasi–insesgados estudiados por *Hartley* y *Ross* (1954), *Quenouille* (1956), *Ruiz* y *Santos* (1989), y otros. (Una recopilación amplia de los estimadores insesgados indirectos puede verse en la bibliografía)

2. Construir estimadores tipo razón que disminuyan el error cuadrático medio bajo ciertas condiciones, como los estudiados por *Ray* y *Sahai* (1980), *Prasad* (1986), *Menéndez* y *Ferrales* (1989), *Srivastava* (1980), etc.

Estos desarrollos están realizados generalmente bajo un esquema de muestreo aleatorio simple. Otro camino para mejorar la precisión de los estimadores es utilizar estimadores de razón bajo otros diseños muestrales más complejos. Así hay trabajos sobre estimadores de razón en muestreo estratificado, *Hansen, Hurwitz* y *Gurney* (1946), muestreo bifásico *Rao* (1975 a, 1975 b, 1981), muestreo polietápico *Williams* (1961, 1962) y *Rao* (1964) y muestreo sistemático *Swain* (1964) y *Singh* (1966).

En este primer capítulo el ambiente en que nos movemos es la estimación de la media poblacional de una variable de interés disponiendo de la información de varias variables auxiliares correladas positivamente con la variable objeto de estudio, presentando en la segunda sección estimadores de razón múltiples conocidos, como el estimador de *Olkin* (1958), los estimadores condensados y el estimador iterado de razón y proponiendo en la tercera sección otra forma de utilizar toda la información que nos proporcionan las variables auxiliares a través del estimador iterado de razón con repeticiones.

§1.2 Estimadores múltiples de razón.

Una primera posibilidad para estimar la media poblacional de la variable de interés disponiendo de varias variables auxiliares correladas positivamente con la objeto de estudio consiste en determinar la variable con mayor correlación con la variable principal y construir a partir de ella el estimador de razón,

despreciando el resto de las variables. Sin embargo es obvio que este procedimiento implica un importante desprecio de recursos.

Olkin (1958) inició los trabajos para la utilización de varias variables auxiliares. La idea de su trabajo es construir con cada variable auxiliar el estimador de razón usual y hacer una combinación lineal de ellos con un cierto criterio de óptimo, consiguiendo así un estimador de razón múltiple.

Para intentar solventar algunos de los incovenientes que presenta el estimador de *Olkin*, *Rueda* y otros (1992) proponen un nuevo método de construir estimadores múltiples de razón mediante los estimadores **condensados** y posteriormente (*Rueda* (1993)) desarrolla otro método diferente de construcción de estimadores de razón múltiples que denomina **método iterado de razón**, que se basa en un procedimiento iterativo en el que el estimador obtenido en un paso se utiliza para construir el estimador en el paso siguiente, presentando grandes ventajas para su puesta en práctica respecto a los estimadores condensados y de *Olkin*.

1.2.1 Notación y definiciones.

Sea N el tamaño de la población, de la que se extrae una muestra de tamaño n en la que se observan la variable principal y y $x_1, x_2, \ldots, x_k$ variables auxiliares.

Notamos con:

f la fracción de muestreo.

$\overline{Y}$, $\overline{X}_i$ las medias poblacionales de las variables y y x_i, $i = 1, \ldots k$, respectivamente.

$\overline{y}$, $\overline{x}_i$ las medias muestrales de las variables y y x_i, $i = 1, \ldots k$, respectivamente.

S_y^2, $S_{x_i}^2$ las cuasivarianzas poblacionales.

s_y^2, $s_{x_i}^2$ las cuasivarianzas muestrales.

$C_{yx_i} = \frac{S_{yx_i}}{\overline{Y}\,\overline{X}_i}$; $\qquad C_{x_ix_j} = \frac{S_{x_ix_j}}{\overline{X}_i\overline{X}_j}$.

$\rho_{yx_i} = \rho_i$, $\rho_{x_ix_j} = \rho_{ij}$ los coeficientes de correlación entre las variables y, x_i y x_i, x_j, respectivamente.

$\widehat{R}_i = \dfrac{\overline{y}}{\overline{x}_i}$ el estimador del cociente $R_i = \dfrac{\overline{Y}}{\overline{X}_i}$.

$\overline{y}_{R_i} = \dfrac{\overline{y}}{\overline{x}_i}\overline{X}_i$ el estimador de razón de la media $\overline{Y}$ usando la variable auxiliar x_i.

1.2.2 El estimador múltiple de *Olkin*.

A partir de ahora consideramos un diseño de muestreo aleatorio simple.

Definición.

Se llama estimador de razón múltiple de *Olkin* de la media

$$\overline{y}_O = w_1\widehat{\overline{Y}}_{R_1} + w_2\widehat{\overline{Y}}_{R_2} + \cdots + w_k\widehat{\overline{Y}}_{R_k}$$

con $w_1, w_2, \ldots, w_k$ pesos definidos de forma que $\sum_{i=1}^{k} w_i = 1$.

Propiedades.

1. El estimador es consistente.
2. El estimador es asintóticamente normal.
3. Una cota del sesgo viene dada por la expresión:

$$|\text{ sesgo }(\overline{y}_O)\ | \leq \max_i \left(\sigma_{\widehat{R}_i} \cdot \sigma_{\overline{x}_i}\right)$$

y una aproximación del sesgo del estimador es

$$\text{sesgo }(\overline{y}_O) \simeq \frac{1-f}{n}\sum_{i=1}^{k}\left(R_i S_{x_i}^2 - \rho_i S_y S_{x_i}\right)\frac{w_i}{\overline{X}_i}$$

4. Una aproximación de orden $O(n^{-2})$ del error cuadrático medio viene dada por

$$\text{ECM }(\overline{y}_O) \simeq \frac{1-f}{n}\overline{Y}^2\sum_{i,j=1}^{k} w_i w_j a_{ij}$$

con $a_{ij} = C_y^2 - \rho_j C_y C_{x_j} - \rho_i C_y C_{x_i} + \rho_{ij} C_{x_i} C_{x_j}$ o en forma matricial

$$\mathrm{ECM}\left(\overline{y}_O\right) \simeq \frac{1-f}{n}\overline{Y}^2 W'AW$$

donde $W = (w_1, \ldots, w_k)'$ y $A = \left(a_{ij}\right)_{k\times k}$.

5. Si $A > 0$ los pesos óptimos que minimizan el error cuadrático medio del estimador vienen dados por

$$\tilde{W} = \frac{A^{-1}e}{e'A^{-1}e}$$

donde $e_{(k\times 1)} = (1, \ldots, 1)'$ y el valor mínimo del error cuadrático medio es:

$$\mathrm{ECM}\left(\overline{y}_O\right) \simeq \frac{1-f}{n}\overline{Y}^2 \frac{1}{e'A^{-1}e}$$

Ventajas.

1. El estimador de razón óptimo es al menos tan preciso como el estimador univariante.

Inconvenientes.

1. El estimador es sesgado.
2. El estimador múltiple óptimo es función de una matriz A construida a partir de parámetros poblacionales desconocidos. En la práctica se solventa este inconveniente estimando $\overline{Y}^2 A$ por la matriz $\widehat{A} = \left(\widehat{a}_{ij}\right)$ donde

$$\widehat{a}_{ij} = \frac{\sum_{t=1}^{n}\left(y_t - r_i x_{it}\right)\left(y_t - r_j x_{jt}\right)}{n-1}$$

con $r_i = \widehat{R}_i$, $i = 1, \ldots k$.

3. El cálculo del estimador de razón óptimo y de su error cuadrático medio es más complejo que el de expansión simple y sin embargo no tiene por qué ser más preciso que éste.

1.2.3 El estimador de razón condensado de mínima varianza (Estimador RCMV).

Para intentar resolver los inconvenientes del estimador de *Olkin*, *Rueda* y otros (1992) proponen un estimador cuya idea básica es la siguiente:

Si se dispone de varias variables auxiliares se pueden "condensar" en otra variable que sea una función lineal de las primeras y utilizar esta nueva variable condensada como variable auxiliar para construir un estimador tipo razón univariante.

Determinación de la variable condensada.

Considerando la función lineal de las variables auxiliares

$$z = \sum_{i=1}^{k} a_i x_i \text{ o matricialmente } z = a'x$$

siendo $a = (a_1, \ldots, a_k)'$ el vector de pesos y $x = (x_1, \ldots, x_k)$ y para obtener mínima varianza, se considera la variable condensada que haga mínimo el error cuadrático medio del estimador de razón que se puede construir a partir de ella.

Definición.

Dada la variable

$$z = a'x = \sum_{i=1}^{k} a_i x_i$$

se define a partir de ella el estimador de razón condensado de mínima varianza como

$$\overline{y}_{RCMV} = \frac{\overline{y}}{\overline{z}}\overline{Z}$$

con $\overline{z} = a'\overline{x}$ y $\overline{Z} = a'\overline{X}$.

Determinación del vector a.

$$\mathrm{ECM}\left(\overline{y}_{RCMV}\right) \simeq \overline{Y}^2 \frac{1-f}{n}\left[\frac{S_y^2}{\overline{Y}^2} + \frac{a'Sa}{\left(a'\overline{X}\right)^2} - \frac{2a'S^0}{\overline{Y}a'\overline{X}}\right]$$

o en forma alternativa

$$\mathrm{ECM}\left(\overline{y}_{RCMV}\right) \simeq \overline{Y}^2 \frac{1-f}{n} \frac{a'\left[\frac{S_y^2 \overline{X}\ \overline{X}'}{\overline{Y}^2} + S - \frac{2S^0\overline{X}'}{\overline{Y}}\right]a}{\left(a'\overline{X}\right)^2}$$

Llamando

$$B_{k\times k} = \frac{S_y^2 \overline{X}\ \overline{X}'}{\overline{Y}^2} + S - \frac{2S^0\overline{X}'}{\overline{Y}}$$

se llega a que

$$\mathrm{ECM}\left(\overline{y}_{RCMV}\right) \geq \overline{Y}^2 \frac{1-f}{n} \frac{1}{\overline{X}'B^{-1}\overline{X}}$$

y se da la igualdad para $a = CB^{-1}\overline{X}$ para alguna constante no nula C.

Así, se define:

$$z = C\overline{X}'B^{-1}x$$

$$\overline{y}_{RCMV} = \frac{\overline{y}}{\overline{X}'B^{-1}\overline{x}}\overline{X}'B^{-1}\overline{X}$$

Propiedades.

1. El estimador es consistente.
2. El estimador es asintóticamente normal.
3. El estimador es sesgado y una aproximación del sesgo viene dada por

$$\mathrm{sesgo}\left(\overline{y}_{RCMV}\right) \simeq \overline{Y}\frac{1-f}{n}\left[\frac{\overline{X}'B^{-1}SB^{-1}\overline{X}}{\left(\overline{X}'B^{-1}\overline{X}\right)^2} - \frac{\overline{X}'B^{-1}S^0}{\overline{Y}\ \overline{X}'B^{-1}\overline{X}}\right]$$

4. Es igual de preciso que el estimador de *Olkin*.

5. Si $p < q$, el error cuadrático medio del estimador de razón condensado de mínima varianza construido a partir de q variables auxiliares es menor que el del estimador construido con p variables auxiliares.

1.2.4 El estimador de razón condensado de mínima varianza e insesgado (Estimador RCMVI).

Para obtener insesgadez, *Rueda* y otros (1992) consideran la combinación lineal

$$z' = a_0 + a'x$$

Imponiendo que la recta de regresión de z' sobre y pase por el origen, se obtiene

$$a_0 = \frac{\overline{Y}a'Sa}{a'S^0} - a'\overline{X}$$

Considerando ahora el estimador

$$\overline{y}_{RCMVI} = \frac{\overline{y}}{\overline{z}'}\overline{Z}'$$

se determina

$$\mathrm{ECM}\left(\overline{y}_{RCMVI}\right) \simeq \frac{1-f}{n}\left[S_y^2 - \frac{\left(a'S^0\right)^2}{a'Sa}\right]$$

que es mínimo para $a = CS^{-1}S^0$.

Por tanto, se define el estimador de razón como:

$$\overline{y}_{RCMVI} = \frac{\overline{y}}{\left(1 + \dfrac{S^{0\prime}S^{-1}\left(\overline{x} - \overline{X}\right)}{\overline{Y}}\right)}$$

Propiedades.

1. El estimador es asintóticamente normal.
2. El estimador es insesgado.
3. El estimador es consistente.
4. Una aproximación del error cuadrático medio viene dado por la expresión

$$\mathrm{ECM}\left(\overline{y}_{RCMVI}\right) \simeq \frac{1-f}{n}\left[S_y^2 - S^{0\prime}S^{-1}S^0\right]$$

Ventajas.

1. Es el estimador más eficiente que se puede construir a partir de una función lineal de las variables auxiliares.
2. Es sencillo su cálculo y el de su precisión.
3. Es más preciso que el estimador de expansión simple.
4. El estimador es insesgado.

Inconvenientes.

1. Depende de las covarianzas poblacionales. En la práctica, éstas se sustituyen por sus valores muestrales.

1.2.5 El estimador de razón condensado (Estimador RC) y el estimador de razón condensado e insesgado (Estimador RCI).

Considerando la función lineal de las variables auxiliares

$$z' = a_0 + a'x$$

y determinando a con otro criterio distinto, el de maximizar la correlación z' con la variable principal y, *Rueda* y otros, imponiendo para la determinación de a_0 la misma condición que para el estimador RCMVI, es decir, imponiendo

$$\underset{(a_0,a)}{máx}\, \rho\,(y,z') = máx\, \frac{a'S^0}{S_y\left(a'Sa\right)^{\frac{1}{2}}}$$

$$\overline{Y} - \frac{a'S^0}{a'Sa}\overline{Z'} = 0,$$

obtienen estimadores condensados que denominan estimador de razón condensado (Estimador RC) y estimador de razón condensado e insesgado (Estimador RCI), comprobando que el estimador $\widehat{\overline{Y}}_{RCMVI}$ coincide con el estimador $\widehat{\overline{Y}}_{RCI}$, teniendo por tanto de las mismas propiedades, ventajas e inconvenientes.

1.2.6 El estimador iterado de razón (Estimador IR).

Para resolver el problema de que los coeficientes de las combinaciones lineales de los tipos de estimadores múltiples citados, para ser óptimos, deben tomar ciertos valores que dependen de parámetros desconocidos, *Rueda* (1993) introduce los estimadores iterados de razón.

Definición.

Se define el estimador de razón iterado de la media a partir de las variables $x_1, x_2, \ldots, x_k$ como

$$\overline{y}_{IR} = \overline{y}\frac{\overline{X}_1}{\overline{x}_1}\frac{\overline{X}_2}{\overline{x}_2}\cdots\frac{\overline{X}_k}{\overline{x}_k}$$

La motivación es la siguiente:

Dada la variable auxiliar x_1, el estimador de razón que se construye a partir de ella, $\overline{y}_{R_1} = \overline{y}\dfrac{\overline{X}_1}{\overline{x}_1}$ es mejor (si la correlación entre y y x_1 es alta) que el estimador de expansión simple, $\overline{y}$. Entonces para construir el estimador de razón utilizando la variable auxiliar x_2, se puede considerar $\overline{y}_{R_1}$ en vez de $\overline{y}$, y por tanto definir

$$\overline{y}_{R_{1,2}} = \overline{y}_{R_1}\frac{\overline{X}_2}{\overline{x}_2} = \overline{y}\frac{\overline{X}_1}{\overline{x}_1}\frac{\overline{X}_2}{\overline{x}_2}$$

Se puede utilizar este estimador en vez de $\overline{y}$ para construir el estimador para x_3 y obtener así:

$$\overline{y}_{R_{1,2,3}} = \overline{y}_{R_{1,2}} \frac{\overline{X}_3}{\overline{x}_3} = \overline{y} \frac{\overline{X}_1}{\overline{x}_1} \frac{\overline{X}_2}{\overline{x}_2} \frac{\overline{X}_3}{\overline{x}_3}$$

y se puede continuar el proceso hasta agotar las k variables auxiliares:

$$\overline{y}_{IR} = \overline{y}_{R_{1,2,\ldots,k}} = \overline{y} \frac{\overline{X}_1}{\overline{x}_1} \frac{\overline{X}_2}{\overline{x}_2} \cdots \frac{\overline{X}_k}{\overline{x}_k}$$

Propiedades.

1. El estimador es consistente.

2. El estimador es asintóticamente normal.

3. El estimador es sesgado y una aproximación de su sesgo viene dada por la expresión

$$\text{sesgo}\left(\overline{y}_{IR}\right) \simeq \frac{1-f}{n} \overline{Y} \left[-\sum_{i=1}^{k} C_{yx_i} + \sum_{i \neq j} C_{x_i x_j} + \sum_{i=1}^{k} C_{x_i}^2 \right]$$

4. Una expresión aproximada del error cuadrático medio del estimador de razón iterado viene dada por la expresión:

$$\text{ECM}\left(\overline{y}_{IR}\right) \simeq \overline{Y}^2 \frac{1-f}{n} \left[C_y^2 + \sum_{i=1}^{k} C_{x_i}^2 + \sum_{i \neq j} C_{x_i x_j} - 2 \sum_{i=1}^{k} C_{yx_i} \right]$$

5. Bajo ciertas condiciones, mejora al de $k-1$ variables auxiliares. En efecto,

   $\text{ECM}\left(\overline{y}_{R_{1,2,\ldots,k}}\right) \leq \text{ECM}\left(\overline{y}_{R_{1,2,\ldots,k-1}}\right)$ si se verifica:

$$2 \sum_{i=1}^{k-1} C_{x_i x_k} + C_{x_k}^2 \leq 2 C_{yx_k}$$

6. Elección de la mejor variable auxiliar para la iteración.

   Es más conveniente utilizar x_p que x_q si

$$C_{x_q}^2 + 2\sum_{j=1}^{i} C_{x_j x_q} - 2C_{y x_q} > C_{x_p}^2 + 2\sum_{j=1}^{i} C_{x_j x_p} - 2C_{y x_p}$$

Por tanto, son más deseables aquellas variables auxiliares que

(a) Tengan un coeficiente de variación más pequeño.

(b) La covarianza con el resto de las variables auxiliares sea menor.

(c) La covarianza con la variable principal sea mayor.

Para aplicar el método iterado de razón de forma óptima se plantea un método forward que permite decidir en qué orden seleccionar las variables y hasta cuando continuar el proceso.

Método forward.

1. Antes de realizar una nueva iteración, se elige la variable x_p tal que la función

$$F\left(x_p\right) = C_{x_p}^2 + 2\sum_{j=1}^{i} C_{x_j x_p} - 2C_{y x_p}$$

sea mínima (siendo x_j, $j = 1, \ldots, i$, las variables utilizadas en las iteraciones anteriores)

2. Una nueva iteración será conveniente si

$$F\left(x_p\right) < 0.$$

Ventajas.

1. El estimador iterado no depende de ningún parámetro desconocido y por tanto se puede calcular siempre.

2. El procedimiento iterativo para construir el estimador y evaluar su error cuadrático medio es computacionalmente sencillo.

3. Es siempre al menos tan preciso como el de expansión simple.

4. No utiliza todas las variables disponibles sino sólo las que producen una mejora en la precisión de la estimación.

Inconvenientes.

1. El estimador es sesgado.

2. De antemano no se sabe si es más o menos preciso que los estimadores condensado y de *Olkin*.

§1.3 Estimador iterado de razón con repeticiones.

1.3.1 Introducción.

Utilizando la información auxiliar de varias variables auxiliares correladas positivamente con la variable objeto de estudio y mediante la técnica de estimación indirecta de razón proponemos un estimador de razón múltiple que va a englobar como caso particular al estimador iterado de razón dado por *Rueda* (1993). Este estimador es una alternativa a otros estimadores de razón múltiples antes expuestos como el de *Olkin* (1958), o los estimadores condensados y los va a mejorar, pues aparte de ser insesgado aumenta la precisión hasta el punto de ser igual de preciso que el estimador de regresión múltiple.

Para el caso en que existan varias variables auxiliares correladas positivamente con la variable principal, una primera solución a los inconvenientes que plantean los estimadores de *Olkin* (1958) y *Rueda* y otros (1992) es el estimador de razón iterado anteriormente expuesto:

$$\widehat{\overline{Y}}_{IR} = \overline{y}\frac{\overline{X}_1}{\overline{x}_1}\frac{\overline{X}_2}{\overline{x}_2}\cdots\frac{\overline{X}_k}{\overline{x}_k}$$

Proponemos en esta sección un nuevo estimador múltiple de tipo razón para la media poblacional. Este estimador, que se obtiene mediante un proceso iterativo que elige en cada paso la mejor variable auxiliar para la iteración, no depende de parámetros desconocidos, se puede calcular siempre y es al menos igual de preciso que el estimador de expansión simple, englobando a éste, a los estimadores de razón univariados, al estimador iterado de razón de la sección anterior y al estimador dado por *Srivastava* (1967) con cada variable auxiliar.

1.3.2 Definición y justificación del estimador.

Definición 1.3.1 *Definimos el estimador de razón iterado con repeticiones de la media a partir de las variables auxiliares* x_1, x_2, ..., x_k *como*

$$\overline{y}_{IRR} = \overline{y}_{IRR}(\alpha_1, \alpha_2, \ldots, \alpha_k) = \overline{y}\left(\frac{\overline{X}_1}{\overline{x}_1}\right)^{\alpha_1}\left(\frac{\overline{X}_2}{\overline{x}_2}\right)^{\alpha_2}\cdots\left(\frac{\overline{X}_k}{\overline{x}_k}\right)^{\alpha_k}$$

donde α_1, α_2, ..., α_k *son enteros no negativos.*

Una posible justificación de esta definición sería la siguiente: Sea x_i la variable auxiliar que permite construir el estimador de razón $\overline{y}_{R_i}$ más preciso. Si $\overline{y}_{R_i}$ es "mejor" que $\overline{y}$ podemos utilizarlo en vez de este último para construir el estimador de razón con otra variable auxiliar x_j, obteniendo el estimador:

$$\overline{y}_{R_{ij}} = \overline{y}\frac{\overline{X}_i}{\overline{x}_i}\frac{\overline{X}_j}{\overline{x}_j}$$

Podría ocurrir que entre todas las variables auxiliares, la mejor para hacer una iteración fuera la misma variable auxiliar que ya ha sido usada, por lo que se podría obtener un estimador:

$$\overline{y}_{R_{ii}} = \overline{y}\left(\frac{\overline{X}_i}{\overline{x}_i}\right)^2$$

Del mismo modo si el estimador obtenido en el segundo paso ($\overline{y}_{R_{ij}}$ ó $\overline{y}_{R_{ii}}$) es más preciso que $\overline{y}$ podría volverse a utilizar en vez de éste para construir el estimador de razón con la variable x_k. Podría ocurrir que la mejor variable auxiliar para la iteración, x_k, fuera alguna de las variables utilizada u otra nueva por lo que al continuar el proceso se iría obteniendo un estimador de razón que puede usar diversas veces una misma variable, es decir, obtendríamos lo que hemos llamado estimador iterado de razón con repeticiones:

$$\overline{y}_{IRR} = \overline{y}\left(\frac{\overline{X}_1}{\overline{x}_1}\right)^{\alpha_1}\left(\frac{\overline{X}_2}{\overline{x}_2}\right)^{\alpha_2}\cdots\left(\frac{\overline{X}_k}{\overline{x}_k}\right)^{\alpha_k}$$

siendo α_1, α_2, ..., α_k el número de veces que se utilizan las variables x_1, x_2, ..., x_k en el procedimiento iterativo.

1.3.3 Propiedades del estimador.

Este estimador tiene las propiedades de consistencia y normalidad asintótica, como muestran las proposiciones siguientes:

Proposición 1.3.2 *El estimador $\overline{y}_{IRR}$ es consistente, en el sentido de consistencia en poblaciones finitas.*

Demostración.-

El estimador es obviamente consistente, en el sentido de consistencia en poblaciones finitas, pues es una función racional de estimadores consistentes.

Vamos a estudiar el comportamiento asintótico del estimador propuesto y para ello consideremos la población finita U como una sucesión de poblaciones $\{U_\nu\}$ donde n_ν y N_ν tienden a infinito de forma que $N_\nu - n_\nu$ también tienda a infinito cuando ν lo haga.

Teorema 1.3.1 *Si las variables $\{y_\nu\}$ y $\{x_{i\nu}\}$ $(i = 1, \ldots, k)$ satisfacen la condición de Lindeberg–Hájek (Hájek 1960), entonces para un muestreo aleatorio simple el estimador $\overline{y}_{IRR}$ es asintóticamente normal.*

Demostración.-

Según el Teorema Central del Límite para poblaciones finitas se tiene que $\overline{y}_\nu$, $\overline{x}_{1\nu}$, $\overline{x}_{2\nu}$, ..., $\overline{x}_{k\nu}$ se distribuyen nomalmente cuando $\nu \to \infty$ pues todas cumplen la condición de Lindeberg–Hájek.

Consideremos ahora la función de $\overline{y}_\nu$, $\overline{x}_{1\nu}$, $\overline{x}_{2\nu}$, ..., $\overline{x}_{k\nu}$ dada por

$$H\left(\overline{y}_\nu, \overline{x}_{1\nu}, \overline{x}_{2\nu}, \ldots, \overline{x}_{k\nu}\right) = \overline{y}_\nu \frac{\overline{B}_\nu}{\overline{b}_\nu}$$

donde

$$\overline{B}_\nu = \prod_{i=1}^{k} \overline{X}_{i\nu}^{\alpha_i} \quad \text{y} \quad \overline{b}_\nu = \prod_{i=1}^{k} \overline{x}_{i\nu}^{\alpha_i} \neq 0$$

Entonces H es una función continua con derivadas parciales de primer y segundo orden, continuas en un entorno de centro $\left(\overline{Y}_\nu, \overline{X}_{1\nu}, \overline{X}_{2\nu}, \ldots, \overline{X}_{k\nu}\right)$ y aplicando el resultado obtenido por *Cramér* (1946), $H\left(\overline{y}_\nu, \overline{x}_{1\nu}, \overline{x}_{2\nu}, \ldots, \overline{x}_{k\nu}\right)$ es asintóticamente normal.

En cuanto al sesgo tenemos la siguiente proposición:

Proposición 1.3.3 *El estimador* $\overline{y}_{IRR}$ *es sesgado y una aproximación del sesgo viene dada por la expresión*

$$\text{sesgo}\,(\overline{y}_{IRR}) \simeq \frac{1-f}{n}\left[\sum_{i=1}^{k}\alpha_i^2 C_{x_i}^2 + \sum_{i\neq j}\alpha_i\alpha_j C_{x_i}C_{x_j} - \sum_{i=1}^{k}\alpha_i C_{yx_i}\right] \qquad (1.3.1)$$

Demostración.-
Para obtener esta expresión se consideran las variables

$$a = \frac{\overline{y}-\overline{Y}}{\overline{Y}} \quad ; \quad e_i = \frac{\overline{x}_i - \overline{X}_i}{\overline{X}_i}$$

a partir de las cuales se puede expresar el estimador:

$$\overline{y}_{IRR} = \overline{Y}\,(1+a)\,(1+e_1)^{-\alpha_1}\,(1+e_2)^{-\alpha_2}\cdots(1+e_k)^{-\alpha_k}$$

Si $\mid e_i \mid < 1$ y $\mid \alpha_i e_i \mid < 1$ $i = 1,\ldots,k$, podemos desarrollar en serie cada término $(1+e_i)^{-\alpha_i}$ y obtenemos la aproximación (una justificación de este tipo de aproximaciones puede verse en *Sukhatme* y *David* 1973, 1974)

$$\overline{y}_{IRR} \simeq \overline{Y}\left(1 + a - \sum_{i=1}^{k}\alpha_i e_i - \sum_{i=1}^{k}\alpha_i a e_i + \sum_{i\neq j}^{k}\alpha_i\alpha_j e_i e_j + \sum_{i=1}^{k}\alpha_i^2 e_i^2\right) \qquad (1.3.2)$$

y entonces

$$\text{sesgo}\,(\overline{y}_{IRR}) = E\left(\overline{y}_{IRR} - \overline{Y}\right) \simeq$$

$$\simeq \overline{Y}E\left(a - \sum_{i=1}^{k}\alpha_i e_i - \sum_{i=1}^{k}\alpha_i a e_i + \sum_{i\neq j}^{k}\alpha_i\alpha_j e_i e_j + \sum_{i=1}^{k}\alpha_i^2 e_i^2\right)$$

y dado que $E(a) = 0$ y $E(e_i) = 0, i = 1,\ldots,k$, se tiene

$$\text{sesgo}\,(\overline{y}_{IRR}) \simeq \overline{Y}\left[-\sum_{i=1}^{k}\alpha_i E\,(a e_i) + \sum_{i\neq j}\alpha_i\alpha_j E\left(e_i e_j\right) + \sum_{i=1}^{k}\alpha_i^2 E\left(e_i^2\right)\right] =$$

$$= \frac{1-f}{n}\overline{Y}\left[-\sum_{i=1}^{k}\alpha_i\frac{S_{yx_i}}{\overline{Y}\,\overline{X}_i}+\sum_{i\neq j}\alpha_i\alpha_j\frac{S_{x_ix_j}}{\overline{X}_i\overline{X}_j}+\sum_{i=1}^{k}\alpha_i^2\frac{S_{x_i}^2}{\overline{X}_i^2}\right]$$

obteniéndose la expresión anterior, y puesto que en muestreo aleatorio simple

$$E\left(ae_i\right)=\frac{\mathrm{Cov}\left(\overline{y},\overline{x}_i\right)}{\overline{Y}\,\overline{X}_i};\qquad \mathrm{Cov}\left(\overline{y},\overline{x}_i\right)=\frac{1-f}{n}S_{yx_i}$$

siendo S_{yx_i} la cuasicovarianza poblacional.

Vamos a estudiar a continuación su precisión. Para ello probamos la siguiente proposición:

Proposición 1.3.4 *Una aproximación del error cuadrático medio del estimador $\overline{y}_{IRR}$ viene dada por la expresión*

$$\mathrm{ECM}\left(\overline{y}_{IRR}\right)\simeq\overline{Y}^2\frac{1-f}{n}U'AU \tag{1.3.3}$$

siendo $A=A_{(k+1)\times(k+1)}$ la matriz de coeficientes de variación y $U=U_{(k+1)}=(1,-\alpha_1,-\alpha_2,\ldots,-\alpha_k)'_{(k+1)}$.

Demostración.-

De la aproximación (1.3.2) obtenemos

$$\mathrm{ECM}\left(\overline{y}_{IRR}\right)=E\left(\overline{y}_{IRR}-\overline{Y}\right)^2\simeq$$

$$\simeq\overline{Y}^2E\left(a-\sum_{i=1}^{k}\alpha_ie_i-\sum_{i=1}^{k}\alpha_iae_i+\sum_{i\neq j}^{k}\alpha_i\alpha_je_ie_j+\sum_{i=1}^{k}\alpha_i^2e_i^2\right)^2$$

$$\simeq\overline{Y}^2E\left(a^2+\sum_{i=1}^{k}\alpha_i^2e_i^2-2\sum_{i=1}^{k}a\alpha_ie_i+\sum_{i\neq j}^{k}\alpha_i\alpha_je_ie_j\right)$$

reteniendo sólo los términos de orden inferior o igual a dos en a y e_i, obtenemos:

$$\mathrm{ECM}\left(\overline{y}_{IRR}\right)\simeq\overline{Y}^2\frac{1-f}{n}\left[\frac{V\left(\overline{y}\right)}{\overline{Y}^2}+\sum_{i=1}^{k}\alpha_i^2\frac{V\left(\overline{x}_i\right)}{\overline{X}_i^2}+\right.$$

$$\left.-2\sum_{i=1}^{k}\alpha_i\frac{\mathrm{Cov}\left(\overline{y},\overline{x}_i\right)}{\overline{Y}\,\overline{X}_i}+\sum_{i\neq j}\alpha_i\alpha_j\frac{\mathrm{Cov}\left(\overline{x}_i,\overline{x}_j\right)}{\overline{X}_i\overline{X}_j}\right]\simeq$$

$$\simeq \overline{Y}^2 \frac{1-f}{n} \left[C_y^2 + \sum_{i=1}^{k} \alpha_i^2 C_{x_i}^2 + \sum_{i \neq j} \alpha_i \alpha_j C_{x_i x_j} - 2 \sum_{i=1}^{k} \alpha_i C_{y x_i} \right].$$

Si llamamos $U = U_{(k+1)} = (1, -\alpha_1, -\alpha_2, \ldots, -\alpha_k)'_{(k+1)}$ podemos escribir el error cuadrático medio de la forma

$$\mathrm{ECM}\,(\overline{y}_{IRR}) \simeq \overline{Y}^2 \frac{1-f}{n} U' A U$$

siendo $A = A_{(k+1)\times(k+1)}$ la matriz de coeficientes de variación.

1.3.4 Determinación del estimador óptimo.

Según razonábamos en la justificación de este estimador las potencias α_1, α_2, ..., α_k indican el número de veces que se utiliza la variable x_i en la estimación de razón. Nos planteamos entonces el problema de ver cuales son los valores de α_1, α_2, ..., α_k que hagan óptimo el estimador iterado de razón en el sentido de mayor precisión.

El problema sería minimizar la expresión

$$F(\alpha_1, \alpha_2, \ldots, \alpha_k) = \mathrm{ECM}\,(\overline{y}_{IRR}(\alpha_1, \alpha_2, \ldots, \alpha_k)) \simeq$$

$$\simeq \overline{Y}^2 \frac{1-f}{n} \left[C_y^2 + \sum_{i=1}^{k} \alpha_i^2 C_{x_i}^2 + \sum_{i \neq j} \alpha_i \alpha_j C_{x_i x_j} - 2 \sum_{i=1}^{k} \alpha_i C_{y x_i} \right]$$

en α_1, α_2, ..., α_k con la restricción de que los α_i son enteros y no negativos. Estaríamos pues ante un problema de programación entera no lineal con restricciones, no teniendo pues garantizados ni la obtención de óptimos globales ni la convergencia en un número finito de pasos.

Para resolver el problema vamos a formular un algoritmo análogo al definido para el estimador iterado de razón de la sección anterior, que permitirá a partir del estimador de expansión simple, construir el estimador iterado de razón con repeticiones, óptimo en el sentido de mayor precisión. Para ello hemos de considerar dos cuestiones:

1. ¿Qué variable auxiliar se debe seleccionar en cada paso?

2. ¿Cuándo debe pararse el proceso?

Vamos a resolver estas cuestiones. Si denotamos por $\mathrm{ECM}(\alpha_1, \alpha_2, \ldots, \alpha_k)$ al error cuadrático medio del estimador $\overline{y}_{IRR}(\alpha_1, \alpha_2, \ldots, \alpha_k)$, tenemos que el estimador $\overline{y}_{IRR}(\alpha_1, \alpha_2, \ldots, \alpha_h + 1, \ldots, \alpha_k)$ es mejor (más preciso) que el estimador $\overline{y}_{IRR}(\alpha_1, \alpha_2, \ldots, \alpha_h, \ldots, \alpha_k)$ si

$$\mathrm{ECM}(\alpha_1, \alpha_2, \ldots, \alpha_h + 1, \ldots, \alpha_k) - \mathrm{ECM}(\alpha_1, \alpha_2, \ldots, \alpha_h, \ldots, \alpha_k) < 0$$

si y sólo si

$$\overline{Y}^2 \frac{1-f}{n} \left[\left(C_y^2 + \sum_{i=1}^{k} \alpha_i^2 C_{x_i}^2 + 2 \sum_{i<j}^{k} \alpha_i \alpha_j C_{x_i x_j} - 2 \sum_{i=1}^{k} \alpha_i C_{y x_i} \right) - \right.$$

$$- \left(C_y^2 + \sum_{i \neq h} \alpha_i^2 C_{x_i}^2 + (\alpha_h + 1)^2 C_{x_h}^2 + 2 \sum_{\substack{1<j \\ i \neq h}}^{k} \alpha_i \alpha_j C_{x_i x_j} \right.$$

$$\left. \left. + 2 \sum_{j \neq h} \alpha_j (\alpha_h + 1) C_{x_h x_j} - 2 \sum_{i \neq h} \alpha_i C_{y x_i} - 2(\alpha_h + 1) C_{y x_h} \right) \right] =$$

$$= \overline{Y}^2 \frac{1-f}{n} \left[(1 + 2\alpha_h) C_{x_h}^2 + 2 \sum_{j \neq h} \alpha_j C_{x_j x_h} - 2 C_{y x_h} \right] < 0$$

Llamando

$$F(\alpha_h) = (1 + 2\alpha_h) C_{x_h}^2 + 2 \sum_{j \neq h} \alpha_j C_{x_j x_h} - 2 C_{y x_h}$$

en cada paso se elegirá la variable auxiliar x_h que haga $F(\alpha_h)$ lo menor posible. Además será conveniente realizar la iteración sólo si $F(\alpha_h) < 0$.

1.3.5 Método Forward.

Para conseguir el estimador iterado de razón con repeticiones de forma óptima procederemos de la siguiente forma:

1. **Paso 1.-**

   Partimos del vector $\alpha^0 = (\alpha_1, \alpha_2, \ldots, \alpha_k) = (0, 0, \ldots, 0)$, es decir, del estimador $\overline{y}$.

2. **Paso 2.-**

   Se elige la variable x_h tal que $F(\alpha_h)$ es mínimo. Además:

   (a) Si $F(\alpha_h) < 0$ se realiza la iteración.

   (b) Si $F(\alpha_h) \geq 0$ se detiene el proceso.

3. **Paso 3.-**

   El vector α^0 se actualiza sumando una unidad a la componente correspondiente a la variable encontrada y se vuelve al paso 2.

El procedimiento para, ya que $\forall x_h$ existe n_h entero tal que $n_h > \frac{C_{yx_h}}{C^2_{x_h}}$ y para ese n_h se tiene que $F(n_h) > 0\ \forall \alpha_i\ i \neq h$, lo que implica que el número de iteraciones con la variable x_h (α_h) no puede ser mayor de n_h. Por tanto todos los α_h son finitos.

Esta función $F(\alpha_1, \alpha_2, \ldots, \alpha_k)$ depende de los coeficientes de variación que en general son desconocidos. No obstante es fácil de estimar sustituyendo los coeficientes de variación poblacionales por sus respectivos valores muestrales.

1.3.6 Mejoras.

Este método tiene una serie de ventajas como son:

1. El estimador iterado con repeticiones no depende de ningún parámetro desconocido y por tanto se puede calcular siempre.

2. El procedimiento iterativo para construir el estimador y evaluar su error cuadrático medio es computacionalmente sencillo.

3. Está garantizado que es siempre al menos tan preciso como el estimador directo.

4. No utiliza forzosamente todas las variables auxiliares disponibles sino sólo las que producen una mejora en la precisión de la estimación.

5. Engloba como casos particulares el estimador directo, los estimadores de razón univariados y al estimador de razón iterado:

(a) Si $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ el estimador $\overline{y}_{IRR}$ resulta ser el estimador de expansión simple

(b) Si $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 1$ el estimador $\overline{y}_{IRR}$ resulta ser el estimador de razón iterado.

(c) Si $\alpha_i = 0 \ \forall i \neq j$ y $\alpha_j = 1$ el estimador $\overline{y}_{IRR}$ resulta ser el estimador de razón univariado $\widehat{\overline{Y}}_{R_j}$.

1.3.7 El estimador iterado de razón con repeticiones insesgado.

De una forma más general podemos definir un estimador iterado de razón con repeticiones en el que los exponentes no sean necesariamente enteros.

Definición 1.3.5 *Definimos el estimador de razón iterado con repeticiones de la media a partir de las variables auxiliares x_1, x_2, ..., x_k como*

$$\overline{y}_{IRR} = \overline{y}_{IRR}(\alpha_1, \alpha_2, \ldots, \alpha_k) = \overline{y} \left(\frac{\overline{X}_1}{\overline{x}_1}\right)^{\alpha_1} \left(\frac{\overline{X}_2}{\overline{x}_2}\right)^{\alpha_2} \cdots \left(\frac{\overline{X}_k}{\overline{x}_k}\right)^{\alpha_k}$$

donde α_1, α_2, ..., α_k son números reales.

Las propiedades de consistencia y normalidad asintótica son análogas, siendo también válidas las aproximaciones del sesgo y del error cuadrático medio obtenidas. Sin embargo, en la determinación del estimador óptimo se obtienen unos valores de $\alpha_1, \ldots, \alpha_k$ reales que hacen al estimador insesgado, como demostramos en la proposición siguiente:

Proposición 1.3.6 *Para $\alpha_0 = C^{-1}C_0$ el estimador $\overline{y}_{IRR}$ es óptimo, en el sentido de mayor precisión, insesgado y su varianza aproximada viene dada por la expresión:*

$$V\left(\overline{y}_{IRR}\right) \simeq \overline{Y}^2 \frac{1-f}{n} \left[C_y^2 - C_0' C^{-1} C_0\right]$$

Demostración.-
Reescribiendo

$$\text{ECM}\,(\overline{y}_{IRR}) \simeq \overline{Y}^2 \frac{1-f}{n}\left[C_y^2 + \sum_{i=1}^{k} \alpha_i^2 C_{x_i}^2 + \sum_{i\neq j} \alpha_i \alpha_j C_{x_i x_j} - 2\sum_{i=1}^{k} \alpha_i C_{y x_i}\right]$$

en la forma

$$\text{ECM}\,(\overline{y}_{IRR}) \simeq \overline{Y}^2 \frac{1-f}{n}\left[C_y^2 + \alpha' C \alpha - 2\alpha' C_0\right] \tag{1.3.4}$$

siendo $C = C_{(k\times k)} = (c_{ij})$ con $c_{ij} = C_{x_i x_j}$, $C_0 = C_{0(k\times 1)} = (c_{0i})$ con $c_{0i} = C_{y x_i}$ y $\alpha = \alpha_{(k\times 1)} = (\alpha_1, \alpha_2, \ldots, \alpha_k)'$, minimizar en α la expresión (1.3.4) equivale a minimizar en α la expresión

$$\alpha' C \alpha - 2\alpha' C_0$$

Asumiendo que C tiene inversa, esta expresión tiene su mínimo en

$$\alpha_0 = C^{-1} C_0$$

y para dicho valor de α el estimador iterado es **insesgado**, puesto que

$$\text{sesgo}\,(\overline{y}_{IRR}) \simeq \frac{1-f}{n}\left[\sum_{i=1}^{k} \alpha_i^2 C_{x_i}^2 + \sum_{i\neq j} \alpha_i \alpha_j C_{x_i} C_{x_j} - \sum_{i=1}^{k} \alpha_i C_{y x_i}\right] =$$

$$= \frac{1-f}{n}\left[\alpha' C \alpha - \alpha' C_0\right],$$

y para $\alpha = \alpha_0 = C^{-1} C_0$ se tiene

$$\text{sesgo}\,(\overline{y}_{IRR}) \simeq \frac{1-f}{n}\left[C_0' C^{-1} C C^{-1} C_0 - C_0' C_{-1} C_0\right] = 0.$$

y su varianza aproximada viene dada por

$$V\,(\overline{y}_{IRR}) \simeq \overline{Y}^2 \frac{1-f}{n}\left[C_y^2 - C_0' C^{-1} C_0\right]$$

Es evidente pues, que el estimador obtenido es siempre más preciso que el estimador de expansión simple, el estimador de razón univariante para cada variable y el estimador de razón iterado, puesto que estos estimadores se obtiene como casos particulares para determinados valores de $\alpha_1, \ldots, \alpha_k$. Pero además, como demostramos en el apartado siguiente, la precisión del estimador propuesto es tan buena como la del estimador de regresión.

1.3.8 Comparación con el estimador de regresión.

Cuando existen varias variables auxiliares $x_1, \ldots, x_k$ se define (véase *Krishnaiah* y *Rao* (1988)) el estimador de regresión mediante la fórmula:

$$\overline{y}_{reg} = \overline{y} + \sum_{i=1}^{k} b_i \left(\overline{X}_i - \overline{x}_i\right)$$

siendo b_i constantes fijadas.

Si el vector $b = b_{(k\times 1)} = (b_1, \ldots, b_k)'$ es el vector de coeficientes de regresión de y sobre x_i $(i = 1, \ldots, k)$ el estimador obtenido es óptimo, en el sentido de mayor precisión, insesgado y su varianza aproximada viene dada por la expresión:

$$V\left(\overline{y}_{reg}\right) = \frac{1-f}{n} \frac{y'_N \left[I_N - x_N \left(x'_N x_N\right)^{-1}\right)^{-1} x'_N\right] y_N}{N-1} \tag{1.3.5}$$

siendo $x_N = x_{N(N\times k)} = \left(\tilde{x}_{ji}\right)$ con $\tilde{x}_{ji} = \left(x_{ji} - \overline{X}_i\right)$, $y_N = y_{N(N\times 1)} = \left(\tilde{y}_j\right)$ con $\tilde{y}_j = \left(y_j - \overline{Y}\right)$ para $i = 1, \ldots, k$ y $j = 1, \ldots, N$.

Proposición 1.3.7 *El estimador iterado de razón con repeticiones óptimo,* $\overline{y}_{IRR}$ *con* $\alpha_0 = C^{-1}C_0$, *es igual de preciso que el estimador de regresión múltiple.*

Demostración.-
Reescribiendo la expresión (1.3.5) en la forma:

$$V\left(\overline{y}_{reg}\right) = \frac{1-f}{n}\left[S_y^2 - S'_0 S^{-1} S_0\right]$$

siendo $S_0 = S_{0(k\times 1)} = (S_i)$ con $S_i = S_{yx_i}$ y $S = S_{(k\times k)} = \left(S_{ij}\right)$ con $S_{ij} = S_{x_i x_j}$, considerando ahora la expresión de la varianza aproximada del estimador iterado de razón con repeticiones óptimo dada en la proposición 1.3.6

$$V\left(\overline{y}_{IRR}\right) \simeq \overline{Y}^2 \frac{1-f}{n}\left[C_y^2 - C'_0 C^{-1} C_0\right]$$

y llamando $T = T_{(k\times k)} = \text{diag}\left(\overline{X}_1^{-1}, \ldots, \overline{X}_k^{-1}\right)$, obtenemos $C = T'ST$ y $C_0 = \dfrac{T'S_0}{\overline{Y}}$.

Entonces

$$C_0'C^{-1}C_0 = \frac{1}{\overline{Y}^2}S_0'TT^{-1}S^{-1}T^{-1}TS_0 = \frac{S_0'S^{-1}S_0}{\overline{Y}^2}$$

de donde se deduce que

$$V(\overline{y}_{IRR}) = \frac{1-f}{n}\left[S_y^2 - S_0'S^{-1}S_0\right]$$

que coincide con la varianza del estimador de regresión.

1.3.9 Mejoras.

Este estimador tiene una serie de ventajas como son:

1. Es insesgado.

2. Es igual de preciso que el estimador de regresión.

3. Engloba como casos particulares el estimador directo, los estimadores de razón univariados, el estimador de razón iterado y el estimador dado por *Srivastava* (1967) con cada variable auxiliar:

   (a) Si $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ el estimador $\overline{y}_{IRR}$ resulta ser el estimador de expansión simple

   (b) Si $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 1$ el estimador $\overline{y}_{IRR}$ resulta ser el estimador de razón iterado.

   (c) Si $\alpha_i = 0 \ \forall i \neq j$ y $\alpha_j = 1$ el estimador $\overline{y}_{IRR}$ resulta ser el estimador de razón univariado $\widehat{\overline{Y}}_{R_j}$.

   (d) Si $\alpha_i = 0 \ \forall i \neq j$ y $\alpha_j \neq 0$ el estimador $\overline{y}_{IRR}$ resulta ser el estimador de *Srivastava* (1967) con la variable auxiliar x_j.

   siendo por tanto más preciso que todos ellos.

1.3.10 Ejemplos numéricos.

Incluimos aquí los resultados de la aplicación de los métodos de razón anteriormente expuestos para tres ejemplos clásicos de la literatura de muestreo en poblaciones finitas.

En cada ejemplo se especifican las variables principal y auxiliares consideradas y se da una tabla con el error cuadrático medio de cada estimador (expansión simple, Olkin, iterado de razón con repeticiones e iterado), su eficiencia respecto al de expansión simple y si tiene o no sesgo.

1. *Olkin* (1958)

   Se quiere estimar el número de habitantes de las 200 ciudades más grandes de Estados Unidos a partir de una muestra de tamaño 50.

   y = Número de habitantes en 1950
   x_1 = Número de habitantes en 1940
   x_2 = Número de habitantes en 1930

Estimador	ECM	Eficiencia	Sesgo
Expansión simple	$N^2\frac{1-f}{n}\cdot 40260.4$	1	NO
Olkin	$N^2\frac{1-f}{n}\cdot 918.8$	43.8	SI
$\overline{y}_{IRR}$	$N^2\frac{1-f}{n}\cdot 842.9$	47.7	NO
Iterado	$N^2\frac{1-f}{n}\cdot 2303$	17.4	SI

2. *Sukhatme* y otros (1984)

   Se quiere estimar la cosecha de guayaba en el distrito de Allahabad (India) a partir de una muestra de 27 pueblos elegida mediante muestreo aleatorio simple de entre 153 pueblos.

   y = producción total de guayaba
   x_1 = número de árboles de guayaba plantados
   x_2 = superficie cultivada de guayaba

Estimador	ECM	Eficiencia	Sesgo
Expansión simple	$N^2\frac{1-f}{n}\cdot 1268$	1	NO
Olkin	$N^2\frac{1-f}{n}\cdot 540$	2.35	SI
$\overline{y}_{IRR}$	$N^2\frac{1-f}{n}\cdot 302.4$	4.19	NO
Iterado	$N^2\frac{1-f}{n}\cdot 403.8$	3.1	SI

3. *Singh* y otros (1983)

   Los datos provienen del Instituto Indio de Investigación Estadística para la Agricultura y corresponden a

   y = peso por parcela de azúcar de caña
   x_1 = altura de las cañas
   x_2 = número de cañas por parcela
   x_3 = diámetro de las cañas

Estimador	ECM	Eficiencia	Sesgo
Expansión simple	$N^2\frac{1-f}{n}\cdot 51.744$	1	NO
Olkin	$N^2\frac{1-f}{n}\cdot 25.278$	2.046	SI
$\overline{y}_{IRR}$	$N^2\frac{1-f}{n}\cdot 24.924$	2.076	NO
Iterado	$N^2\frac{1-f}{n}\cdot 34.413$	1.503	SI

En los ejemplos se observa como aumenta la eficiencia con respecto al estimador de expansión simple, siendo el estimador iterado de razón con repeticiones el más preciso en los tres ejemplos. La eficiencia con respecto a los otros dos estimadores, de Olkin e iterado, es variable, siendo especialmente significativa con respecto al estimador iterado en el primer ejemplo y con respecto al estimador de Olkin en el segundo.

Capítulo 2

Variables auxiliares con correlación negativa con la variable de interés: Estimadores múltiples de producto.

§2.1 Introducción.

En este segundo capítulo el ambiente en que nos movemos es la estimación de la media poblacional de una variable de interés disponiendo de la información de varias variables auxiliares correladas negativamente con la variable objeto de estudio, presentando en la segunda sección el estimador de producto múltiple más conocido, el estimador de *Singh* (1967), estudiando sus propiedades y proponiendo en la tercera sección un estimador alternativo para la media poblacional siguiendo la idea de *Rueda* y otros (1992), condensando la información en una única variable para construir con ella un estimador múltiple tipo producto.

Es conocido (véase *Murthy* (1964)) que el uso del estimador producto construido con la variable x_i:

$$\overline{y}_{P_i} = \overline{y}\frac{\overline{x}_i}{\overline{X}_i},$$

cuando existe una relación del tipo $y_i x_i = \text{cte.}$, proporciona una mejora en la precisión de la estimación respecto al estimador directo.

El sesgo exacto de este estimador producto viene dado por la expresión

$$\text{sesgo}\left(\overline{y}_{P_i}\right) = \frac{1-f}{n}\frac{S_{yx_i}}{\overline{X}_i}$$

y su tiene por error cuadrático medio aproximado es

$$\text{ECM}\left(\overline{y}_{P_i}\right) \simeq \frac{1-f}{n}\left(S_y^2 + R_i^2 S_{x_i}^2 + 2R_i\rho_{yx_i}S_{x_i}S_y\right).$$

Este estimador proporciona un aumento en precisión con respecto al estimador de expansión simple (si $\rho_{yx_i} < -\frac{C_{x_i}}{2C_y}$) e incluso respecto al estimador de razón $\overline{y}_{R_i} = \overline{y}\frac{\overline{X}_i}{\overline{x}_i}$ (véase *Chaubey, Dwivedi* y *Singh* (1984)).

§2.2 Estimadores múltiples de producto.

De acuerdo con lo anterior y siguiendo la idea de *Olkin* (1958) se pueden construir estimadores tipo producto cuando se disponga de varias variables auxiliares correladas negativamente con la variable objeto de estudio, como el estimador dado por *Singh* (1967), que a continuación exponemos, señalando sus propiedades más importantes.

2.2.1 El estimador múltiple de *Singh*.

Definición.

Se llama estimador de producto múltiple de *Singh* de la media a

$$\overline{y}_S = w_1\overline{y}_{P_1} + w_2\overline{y}_{P_2} + \cdots + w_k\overline{y}_{P_k} \tag{2.2.1}$$

con w_1, w_2, ..., w_k pesos definidos de forma que $\sum_{i=1}^{k} w_i = 1$.

Propiedades.

1. El estimador es consistente.

2. El estimador es asintóticamente normal.

3. El estimador es sesgado y su sesgo exacto viene dado por la expresión

$$\text{sesgo}\,(\overline{y}_S) = \frac{1-f}{n}\sum_{i=1}^{k} w_i \frac{S_{yx_i}}{\overline{X}_i}$$

4. Llamando $D = D_{(k\times k)} = (d_{ij})$ con $d_{ij} = S_y^2 + R_j S_{yx_j} + R_i S_{yx_i} + R_i R_j S_{x_i x_j}$, $i, j = 1, \ldots, k$ una aproximación del error cuadrático medio del estimador es

$$\text{ECM}\,(\overline{y}_S) \simeq \frac{1-f}{n} W'DW$$

donde $W = (w_1, \ldots, w_k)'$.

5. Si $D > 0$ los pesos óptimos que minimizan el error cuadrático medio del estimador vienen dados por

$$\tilde{W} = \frac{D^{-1}e}{e'D^{-1}e}$$

donde $e_{(k\times 1)} = (1, \ldots, 1)'$ y el valor mínimo del error cuadrático medio es:

$$\text{ECM}\,(\overline{y}_S) \simeq \frac{1-f}{n}\frac{1}{e'D^{-1}e} \tag{2.2.2}$$

6. El estimador múltiple óptimo es función de una matriz D construida a partir de parámetros poblacionales desconocidos. En la práctica se solventa este inconveniente estimando D por la matriz $\widehat{D} = \left(\widehat{d}_{ij}\right)$ donde

$$\widehat{d}_{ij} = s_y^2 + r_j s_{yx_j} + r_i s_{yx_i} + r_i r_j s_{x_i x_j}; \qquad i, j = 1, \ldots, k$$

donde

$$r_i = \frac{\overline{y}}{\overline{x}_i} \quad ; \quad s_y^2 = \frac{1}{n-1}\sum_{t=1}^{n}(y_t - \overline{y})^2$$

$$s_{yx_i} = \frac{1}{n-1}\sum_{t=1}^{n}(y_t - \overline{y})(x_{i_t} - \overline{x}_i) \; ; \; s_{x_i x_j} = \frac{1}{n-1}\sum_{t=1}^{n}(x_{i_t} - \overline{x}_i)\left(x_{j_t} - \overline{x}_j\right).$$

§2.3 Estimador producto condensado.

Como el estimador producto propuesto por *Singh* (1967) es una combinación lineal de estimadores producto univariantes, es lógico que para que tenga buenas propiedades debería de existir una relación inversa entre la variable principal y las variables auxiliares. Parece díficil que en la práctica se de esta situación, pero es más creíble que pueda existir una relación inversa entre la variable principal y una combinación lineal de variables auxiliares. Esta nueva variable puede ser adecuada para construir el estimador producto condensado, que presentamos a continuación.

2.3.1 Definición del estimador.

Sean $x_1, \ldots, x_k$ variables auxiliares correladas negativamente con la variable objeto de estudio y.

Consideremos la variable condensada z definida de la forma

$$z = \sum_{i=1}^{k} a_i x_i = a'x \text{ con } a = (a_1, \ldots, a_k)' \ , \ x = (x_1, \ldots, x_k)'$$

y por tanto

$$\overline{z} = a'\overline{x} \ , \ \overline{x} = (\overline{x}_1, \ldots, \overline{x}_k)'$$

$$\overline{Z} = a'\overline{X} \ , \ \overline{X} = \left(\overline{X}_1, \ldots, \overline{X}_k\right)'$$

Sean las matrices

$$S = S_{(k\times k)} = \left(S_{ij}\right) \ , \ S_{ij} = \text{Cov}\left(x_i, x_j\right)$$

$$S^0 = S^0_{(k\times 1)} = \left(S_{yx_1}, \ldots, S_{yx_k}\right) \ , \ S_{yx_j} = \text{Cov}\left(y, x_j\right)$$

Construimos el estimador producto condensado de la forma

$$\overline{y}_{PC} = \overline{y}\frac{\overline{z}}{\overline{Z}} \tag{2.3.1}$$

2.3.2 Propiedades.

Ahora estudiamos dos propiedades de este estimador.

Estudio del error cuadrático medio.

Proposición 2.3.1 *Una aproximación de orden* $O\left(n^{-2}\right)$ *del error cuadrático medio del estimador producto condensado viene dada por expresión:*

$$\mathrm{ECM}\left(\overline{y}_{PC}\right) \simeq \overline{Y}^2 \frac{(1-f)}{n}\left[\frac{S_y^2}{\overline{Y}^2} + \frac{a'Sa}{\left(a'\overline{X}\right)^2} + \frac{2a'S^0}{\overline{Y}a'\overline{X}}\right] \tag{2.3.2}$$

Demostración.-
Consideramos las variables

$$o_1 = \frac{\overline{y}-\overline{Y}}{\overline{Y}} \; ; \quad o_2 = \frac{\overline{z}-\overline{Z}}{\overline{Z}}.$$

Dependiendo de estas variables podemos expresar el estimador $\overline{y}_{PC}$ de la siguiente forma:

$$\overline{y}_{PC} = \overline{Y}\left(1+o_1\right)\left(1+o_2\right)$$

Entonces

$$\overline{y}_{PC} - \overline{Y} = \left(o_1 + o_2 + o_1 o_2\right)\overline{Y}.$$

Elevando al cuadrado y quedándonos sólo con los términos de orden igual o inferior a dos en o_1 y o_2 obtenemos la aproximación:

$$\left(\overline{y}_{PC} - \overline{Y}\right)^2 \simeq \overline{Y}^2\left(o_1^2 + o_2^2 + 2o_1 o_2\right)$$

y así

$$\mathrm{ECM}\left(\overline{y}_{PC}\right) \simeq \overline{Y}^2\left[E\left(o_1^2\right) + E\left(o_2^2\right) + 2E\left(o_1 o_2\right)\right].$$

Sustituyendo los valores de $E\left(o_1^2\right)$, $E\left(o_2^2\right)$ y $E\left(o_1 o_2\right)$ en muestreo aleatorio simple:

$$E\left(o_1^2\right) = \frac{S_y^2}{\overline{Y}^2}; \quad E\left(o_2^2\right) = \frac{a'Sa}{\left(a'\overline{X}\right)^2}; \quad E\left(o_1 o_2\right) = \frac{a'S^0}{\overline{Y}a'\overline{X}},$$

obtenemos una aproximación de orden $O(n^{-2})$ del error cuadrático medio:

$$\mathrm{ECM}\left(\overline{y}_{PC}\right) \simeq \overline{Y}^2 \frac{(1-f)}{n}\left[\frac{S_y^2}{\overline{Y}^2} + \frac{a'Sa}{\left(a'\overline{X}\right)^2} + \frac{2a'S^0}{\overline{Y}a'\overline{X}}\right]$$

El orden de la aproximación nos lo proporciona el siguiente lema:

Lema 2.3.1 *Sea una sucesión de poblaciones* $\{S_n\}$ *de tamaños* $\{N_n\}$ *extraídas mediante muestreo aleatorio simple de una superpoblación infinita* $S = \{U_i,\ V_i\}$. *Sea* $\{N_n\}$ *estrictamente creciente y tal que*

$$\lim_{n\to\infty} t_n = \lim_{n\to\infty} \frac{n}{N_n} = t \quad 0 < t < 1 \ y\ 0 < t_n < 1\ \forall n$$

Suponiendo que para algunos enteros positivos fijados r *y* s, *los momentos de orden* $r+s$ *e inferiores de* S_n *permanecen acotados. Sean* $\overline{u}$ *y* $\overline{v}$ *las medias muestrales basadas en una muestra de tamaño* n *extraída de* S_n *mediante muestreo aleatorio simple. Entonces si* $\overline{U}$ *y* $\overline{V}$ *son las medias poblacionales para* S_n, *se tiene*

$$E\left[\left(\overline{u}-\overline{U}\right)^r\left(\overline{v}-\overline{V}\right)^s\right] = \begin{cases} O\left(n^{-\frac{r+s}{2}}\right) & si\ r+s\ es\ par \\ O\left(n^{-\frac{r+s+1}{2}}\right) & si\ r+s\ es\ impar \end{cases}$$

En nuestro caso, llamando $\mathrm{ECM}_2\left(\overline{y}_{PC}\right)$ a la aproximación tomada (que según el lema contiene los términos de o_1 y o_2 hasta los de orden $O\left(n^{-2}\right)$), obtenemos

$$\mathrm{ECM}\left(\overline{y}_{PC}\right) - \mathrm{ECM}_2\left(\overline{y}_P\right) = \overline{Y}^2\left(E\left(o_1^2o_2^2\right) + 2E\left(o_1^2o_2\right) + 2E\left(o_1o_2^2\right)\right)$$

cuyos sumandos son de orden $O\left(n^{-2}\right)$.

Estudio del sesgo.

Proposición 2.3.2 *El sesgo exacto del estimador producto condensado viene dado por la expresión*

$$\mathrm{sesgo}\left(\overline{y}_{PC}\right) = \frac{(1-f)}{n}\frac{a'S^0}{a'\overline{X}}$$

Demostración.-
Calculamos el sesgo del estimador dado en (2.3.1):

$$\text{sesgo}\,(\overline{y}_{PC}) = E\,(\overline{y}_{PC}) - \overline{Y} = \overline{Y}E\,(o_1 + o_2 + o_1 o_2) =$$

$$= \overline{Y}E\,(o_1 o_2) = \overline{Y}\frac{(1-f)}{n}\frac{\text{Cov}(z,y)}{\overline{Y}\,\overline{Z}} = \frac{(1-f)}{n}\frac{a'S^0}{a'\overline{X}}$$

Podemos observar que el sesgo tiende a cero cuando n aumenta.

2.3.3 El estimador óptimo.

Hasta ahora hemos supuesto que la variable condensada es conocida. Pero si esto no es cierto, podemos seleccionar la mejor combinación lineal. Dos vías son posibles:

1. Considerar la variable condensada z que maximice la correlación (en valor absoluto) con la variable principal.
2. Considerar la variable condensada que proporcione el estimador más preciso.

Obviamente seleccionamos el segundo camino.

Determinación del vector a.

Vamos a determinar el vector a de forma que haga mínimo el error cuadrático medio del estimador $\overline{y}_{PC}$ construido a partir de la variable condensada.

De (2.3.2), minimizar en a el valor $\text{ECM}\,(\overline{y}_{PC})$ equivale a minimizar en a la expresión

$$\frac{S_y^2}{\overline{Y}^2} + \frac{a'Sa}{\left(a'\overline{X}\right)^2} + \frac{2a'S^0}{\overline{Y}a'\overline{X}}$$

o equivalentemente

$$\frac{a'\left[\dfrac{S_y^2\overline{X}\;\overline{X}'}{\overline{Y}^2} + S + \dfrac{2S^0\overline{X}'}{\overline{Y}}\right]a}{\left(a'\overline{X}\right)^2} = \frac{a'Ba}{\left(a'\overline{X}\right)^2}$$

donde

$$B = B_{(k\times k)} = \frac{S_y^2 \overline{X}\ \overline{X}'}{\overline{Y}^2} + S + \frac{2S^0\overline{X}'}{\overline{Y}}$$

Si la matriz B es definida positiva, aplicando la desigualdad extendida de Cauchy–Schwartz (véase al final de este apartado), obtenemos

$$\frac{\left(a'\overline{X}\right)^2}{a'Ba} \leq \overline{X}'B^{-1}\overline{X}$$

Por tanto el error cuadrático medio alcanza su valor mínimo en $a = CB^{-1}\overline{X}$ para alguna constante no nula C y ese mínimo vale

$$\mathrm{ECM}\,(\overline{y}_{PC}) = \overline{Y}^2 \frac{1-f}{n} \frac{1}{\overline{X}'B^{-1}\overline{X}}$$

Entonces definimos

$$z = C\overline{X}'B^{-1}x$$

con lo que obtenemos

$$\overline{z} = C\overline{X}'B^{-1}\overline{x}\ ,\quad \overline{Z} = C\overline{X}'B^{-1}\overline{X}$$

y así elegimos el estimador de producto

$$\overline{y}_{PC} = \frac{\overline{y}}{\overline{X}'B^{-1}\overline{X}}\overline{X}'B^{-1}\overline{x} \tag{2.3.3}$$

cuyo error cuadrático medio viene dado por

$$\mathrm{ECM}\,(\overline{y}_{PC}) = \overline{Y}^2 \frac{1-f}{n} \frac{1}{\overline{X}'B^{-1}\overline{X}}$$

Desigualdad extendida de Cauchy–Schwartz.

Sean $b_{(k\times 1)}$ y $d_{(k\times 1)}$ vectores y $B_{(k\times k)}$ definida positiva. Entonces

$$(b'd)^2 \leq (b'Bb)\left(d'B^{-1}d\right) \tag{2.3.4}$$

y la igualdad es cierta si y sólo si $b = cB^{-1}d$ para alguna constante c

Comparación con el estimador producto múltiple.

En este apartado comparamos la precisión del estimador de producto múltiple dado por *Singh* con la del estimador alternativo que proponemos, concluyendo que ambos son igual de precisos, para muestras grandes. Para muestras de tamaño moderado presentamos los resultados de un estudio de simulación en el que se comprueba que el estimador de producto condensado presenta una precisión mayor que el estimador de *Singh*.

Proposición 2.3.3 *Si n es suficientemente grande, el estimador producto condensado es igual de eficiente que el estimador múltiple de Singh.*

Demostración.-

Una aproximación del error cuadrático medio del estimador producto múltiple

$$\overline{y}_S = \sum_{i=1}^{k} w_i \overline{y}_{P_i} \ , \ \sum_{i=1}^{k} w_i = 1 \ , \ \overline{y}_{P_i} = \overline{y}\frac{\overline{x}_i}{\overline{X}_i} (i = 1, \ldots, k)$$

la podemos obtener reescribiendo la expresión (2.2.2) en la forma:

$$\mathrm{ECM}\left(\overline{y}_S\right) \simeq \frac{1-f}{n}\overline{Y}^2 \frac{1}{e' A^{-1} e}$$

siendo $e' = (1, \ldots, 1)$, $A = A_{(k\times k)} = \left(a_{ij}\right)$ con

$$a_{ij} = C_y^2 + \rho_j C_y C_{x_j} + \rho_i C_y C_{x_i} + \rho_{ij} C_{x_i} C_{x_j}$$

La aproximación del error cuadrático medio del estimador propuesto, $\overline{y}_{PC}$, puede tranformarse de la forma

$$\mathrm{ECM}\left(\overline{y}_{PC}\right) \simeq \overline{Y}^2 \frac{1-f}{n}\left[\frac{S_y^2}{\overline{Y}^2} + \frac{a'Sa}{\left(a'\overline{X}\right)^2} + \frac{2a'S^0}{\overline{Y}a'\overline{X}}\right] =$$

$$= \overline{Y}^2 \frac{1-f}{n} \frac{a'\overline{X}\frac{S_y^2}{\overline{Y}^2}\overline{X}'a + a'Sa + \frac{a'S^0\overline{X}'a}{\overline{Y}} + \frac{a'\overline{X}S^{0\prime}a}{\overline{Y}}}{\left(a'\overline{X}\right)^2} = \overline{Y}^2 \frac{1-f}{n} \frac{a'Ca}{\left(a'\overline{X}\right)^2},$$

donde

$$C = \frac{S_y^2}{\overline{Y}^2}\overline{X}\,\overline{X}' + S + \frac{S^0\overline{X}'}{\overline{Y}} + \frac{\overline{X}S^{0\prime}}{\overline{Y}}.$$

Si asumimos que C es definida positiva (es como mínimo semidefinida) podemos aplicar la desigualdad extendida de Cauchy–Schwartz y obtenemos que el valor mínimo de ECM $(\overline{y}_{PC})$ adopta la forma

$$\text{ECM}\,(\overline{y}_{PC}) \simeq \overline{Y}^2 \frac{1-f}{n} \frac{1}{\overline{X}'C^{-1}\overline{X}}$$

Ahora bien, llamando $M = M_{(k\times k)} = \text{diag}\left(\overline{X}_1^{-1}, \overline{X}_2^{-1}, \ldots, \overline{X}_k^{-1}\right)$, obtenemos $C = MAM'$ y por tanto

$$\overline{X}'C^{-1}\overline{X} = \overline{X}'M'^{-1}A^{-1}M^{-1}\overline{X} = e'A^{-1}e$$

Entonces concluimos que ECM $(\overline{y}_{PC}) \simeq$ ECM $(\overline{y}_S)$ y ambos estimadores son igual de eficientes.

Estimación del error cuadrático medio.

De acuerdo con la proposición anterior, directamente de (2.2.2) y siguiendo *Olkin*, un estimador del error cuadrático medio viene dado por

$$\widehat{\text{ECM}}\,(\overline{y}_{PC}) \simeq \widehat{\text{ECM}}\,(\overline{y}_S) \simeq \frac{(1-f)}{n}\frac{1}{e'\widehat{D}e}$$

con $\widehat{D} = \left(\hat{d}_{ij}\right)$ un estimator de la matriz D es como sigue:

$$\hat{d}_{ij} = s_y^2 + \hat{R}_j S_{yx_j} + \hat{R}_i S_{yx_i} + \hat{R}_i\hat{R}_j S_{x_ix_j}.$$

Generalmente, este estimador es sesgado y su sesgo tiene orden $O(n^{-1})$ (véase *Cochran* (1971)).

Estimación de los pesos.

El estimador $\overline{y}_{PC}$ de (2.3.3) depende de B, que es función de momentos poblacionales desconocidos, pero en la práctica pueden ser estimados a partir de los datos muestrales de la forma:

$$\widehat{B} = \widehat{B}_{(k\times k)} = c_y^2 \overline{X}\ \overline{X}' + \widehat{S} + 2\widehat{C}^0 \overline{X}'$$

con

$$\widehat{S} = \left(\widehat{S}_{ij}\right); \quad \widehat{S}_{ij} = s_{x_i x_j}; \quad c_y^2 = \frac{s_y^2}{\overline{y}^2}$$

$$\widehat{C}^0 = \left(c_{yx_i}\right) \quad ; \quad c_{yx_i} = \frac{s_{x_i}}{\overline{y}\ \overline{x}_i}$$

Sustituyendo la matriz $\widehat{B}$ en (2.3.3) obtenemos una aproximación del estimador óptimo teórico.

Para estudiar el efecto de la estimación de los pesos consideramos el estimador producto condensado

$$\overline{y}_{PC} = \frac{\overline{y}}{\overline{X}' B^{-1} \overline{X}} \overline{X}' B^{-1} \overline{x} = \overline{y} f\left(B, \overline{x}\right)$$

donde B es un parámetro desconocido. Siguiendo *Fuller* (1976; Ch 5) tenemos que

$$f\left(\widehat{B}, \overline{x}\right) - f\left(B, \overline{x}\right) = O_p\left(n^{-\frac{1}{2}}\right)$$

y por tanto el estimador obtenido de

$$\overline{y}'_{PC} = \overline{y} f\left(\widehat{B}, \overline{x}\right)$$

despreciando los términos de orden $O_p\left(n^{-\frac{3}{2}}\right)$ tiene la misma varianza que $\overline{y}_{PC}$, costruido con la variable condensada z.

De esta forma, el efecto producido al sustituir $\widehat{B}$ en lugar de B es asintóticamente despreciable.

El estimador producto condensado e insesgado.

Construimos a continuación un estimador producto condensado que no tenga sesgo. El procedimiento que utilizamos para ello es muy simple.

Según la proposición 2.3.2, el estimador condensado es en general sesgado y su sesgo exacto viene dado por la expresión

$$\text{sesgo}\,(\overline{y}_{PC}) = \frac{1-f}{n}\frac{\overline{X}'B^{-1}S^0}{\overline{X}'B^{-1}\overline{X}}$$

Entonces podemos modificar el estimador $\overline{y}_{PC}$ de la forma

$$\overline{y}_{PCI} = \overline{y}_{PC} - \frac{1-f}{n}\frac{\overline{X}'B^{-1}S^0}{\overline{X}'B^{-1}\overline{X}} = \frac{\overline{X}'B^{-1}}{\overline{X}'B^{-1}\overline{X}}\left[\overline{x}\;\overline{y} - \frac{1-f}{n}S^0\right]$$

que evidentemente es un estimador sin sesgo de la media poblacional, cuya precisión es la misma del estimador producto condensado dado en 2.3.3.

Distribución asintótica.

Vamos a estudiar el comportamiento asintótico del estimador propuesto y para ello consideremos la población finita U como una sucesión de poblaciones $\{U_\nu\}$ donde n_ν y N_ν tienden a infinito de forma que $N_\nu - n_\nu$ también tienda a infinito cuando ν lo haga.

Teorema 2.3.1 *Si las variables $\{y_\nu\}$ y $\{x_{i\nu}\}$ $(i = 1, \ldots, k)$ satisfacen la condición de Lindeberg–Hájek (Hájek 1960), entonces para un muestreo aleatorio simple el estimador $\overline{y}_{PC}$ es asintóticamente normal.*

Demostración.-

Análogamente a la demostración dada en el teorema 1.3.1 y considerando ahora la función de $\overline{y}_\nu$, $\overline{x}_{1\nu}$, $\overline{x}_{2\nu}$, ..., $\overline{x}_{k\nu}$ dada por

$$H\,(\overline{y}_\nu, \overline{x}_{1\nu}, \overline{x}_{2\nu}, \ldots, \overline{x}_{k\nu}) = \overline{y}_\nu \frac{\overline{b}_\nu}{\overline{B}_\nu}$$

donde

$$\overline{b}_\nu = \sum_{i=1}^{k} a_{i\nu}\overline{x}_{i\nu} \quad \text{y} \quad \overline{B}_\nu = \sum_{i=1}^{k} a_{i\nu}\overline{X}_{i\nu} \neq 0$$

se obtiene el resultado enunciado. Evidentemente, el resultado también es cierto para el estimador $\overline{y}_{PCI}$.

Capítulo 3

Variables auxiliares con correlación positiva y negativa con la variable de interés: Estimadores múltiples de razón–producto.

§3.1 Introducción.

En este tercer capítulo estudiaremos cómo utilizar la información auxiliar que proporcionan varias variables auxiliares con correlación de cualquier signo con la variable objeto de estudio, para la estimación de la media poblacional de ésta, presentando en la segunda sección el estimador múltiple más usual dado por *Rao* y *Mudholkar* (1967) con sus propiedades más relevantes y aportando en la tercera sección un estimador alternativo llamado estimador iterado de razón–producto.

Cuando sólo son dos las variables auxiliares, x con correlación positiva con la variable principal y y la variable auxiliar z con correlación negativa, *Singh* (1967a) dió una primera forma de combinar la información auxiliar proporcionada por las dos variables, mediante un estimador de razón–producto. Notar que estos dos estimadores indirectos (de razón y de producto) son, por separado y bajo ciertas condiciones, (si $\rho_{xy} > \frac{C_x}{2C_y}$ y si $\rho_{zy} < \frac{-C_z}{2C_y}$, respectivamente, donde C_t es el coeficiente de variación de la variable t) más precisos que el estimador de expansión simple.

Singh (1967a) define el estimador de razón–producto de la media poblacional como

$$\overline{y}_{RPS} = \overline{y}\left(\frac{\overline{X}}{\overline{x}}\right)\left(\frac{\overline{z}}{\overline{Z}}\right)$$

Este estimador es sesgado y aproximaciones del sesgo y del error cuadrático medio vienen dadas por las expresiones:

$$\text{sesgo}\left(\overline{y}_{RPS}\right) \simeq \frac{N-n}{nN}\overline{Y}\left\{C_x^2 - \rho_{yx}C_yC_x + \rho_{yz}C_yC_z - \rho_{xz}C_xC_z\right\}$$

$$\text{ECM}\left(\overline{y}_{RPS}\right) \simeq \frac{N-n}{nN}\overline{Y}^2\left\{C_y^2 + C_x^2 + C_z^2 - 2\rho_{yx}C_yC_x + 2\rho_{yz}C_yC_z - 2\rho_{xz}C_xC_z\right\}$$

Singh y *Biradar* (1992), usando la técnica jackknife, formulan una clase general de estimadores de razón–producto cuasi–insesgados

$$\overline{y}_{SB} = \left[\left\{1 - \lambda\left(1-\theta\right)\right\}\overline{y} + \lambda\overline{y}_{RPS} - \lambda\theta\overline{y}_{RPSJ}\right]$$

donde $\theta = \frac{(N-n)(n-m)}{n(N-n+m)}$ e $\overline{y}_{RPSJ}$ es el estimador jackknife obtenido apartir del estimador de razón–producto al dividir la muestra de tamaño n en g muestras de tamaño m, es decir,

$$\overline{y}_{RPSJ} = \frac{1}{g}\sum_{j=1}^{g}\overline{y}_j'\left(\frac{\overline{X}}{\overline{x}_j'}\right)\left(\frac{\overline{z}_j'}{\overline{Z}}\right),$$

siendo $\overline{t}_j'$ la media muestral de la variable t obtenida omitiendo el j–ésimo grupo, obteniendo después el óptimo en esta clase.

Otra forma de combinar la información de las variables x y z es sumando los respectivos estimadores de razón y de producto obtenidos con cada una de las variables y ponderándolos de forma que la precisión sea máxima. Es decir:

$$\overline{y}_A = w_1\overline{y}\frac{\overline{X}}{\overline{x}} + w_2\overline{y}\frac{\overline{z}}{\overline{Z}}$$

con $w_1 + w_2 = 1$ y eligiéndolos de forma que el error cuadrático medio del estimador $\overline{y}_A$ sea mínimo.

Esta idea de combinar estimadores de razón y de producto adecuadamente se puede utilizar para el caso en que exista más de una variable de cada tipo, como veremos a continuación.

§3.2 Estimadores múltiples de razón–producto.

Sean x_1, x_2, ..., x_k las variables auxiliares disponibles y supongamos que

$$\rho_{yx_i} > 0 \quad i = 1, \ldots, p \quad \text{y}$$
$$\rho_{yx_i} < 0 \quad i = p+1, \ldots, k$$

3.2.1 El estimador de *Rao* y *Muldholkar*.

Siguiendo la última idea del apartado anterior, *Rao* y *Muldholkar* (1967) consideran el estimador

$$\overline{y}_{RM} = \sum_{i=1}^{p} w_i \overline{y} \frac{\overline{X}_i}{\overline{x}_i} + \sum_{i=p+1}^{k} w_i \overline{y} \frac{\overline{x}_i}{\overline{X}_i}$$

con $\sum_{i=1}^{k} w_i = 1$.

El sesgo de este estimador es

$$\text{sesgo}\,(\overline{y}_{RM}) = -\sum_{i=1}^{p} w_i \,\text{Cov}\left(\frac{\overline{y}}{\overline{x}_i}, \overline{x}_i\right) + \sum_{i=p+1}^{p+q} \frac{w_i}{\overline{X}_i} \,\text{Cov}\,(\overline{y}, \overline{x}_i)$$

y una aproximación de orden $O(n^{-2})$ viene dada por la expresión

$$\text{sesgo}\,(\overline{y}_{RM}) \simeq \frac{1-f}{n}\overline{Y}\left[\sum_{i=1}^{p} w_i \left(C_y^2 - \rho_i C_y C_{x_i}\right) + \sum_{i=p+1}^{k} w_i \rho_i C_y C_{x_i}\right]$$

En cuanto al error cuadrático medio del estimador, una aproximación es

$$\text{ECM}\,(\overline{y}_{RM}) \simeq \frac{1-f}{n}\overline{Y}^2 \sum_{i,j=1}^{k} w_i w_j d_{ij} = \frac{1-f}{n}\overline{Y}^2 W' D W$$

donde $W = (w_1, \ldots, w_k)'$ y $D = \left(d_{ij}\right)_{(k\times k)}$ con

$$d_{ij} = C_y - \rho_i C_y C_{x_i} - \rho_j C_y C_{x_j} + \rho_{ij} C_{x_i} C_{x_j} \text{ si } i, j = 1, \ldots, p$$

$$d_{ij} = C_y + \rho_i C_y C_{x_i} + \rho_j C_y C_{x_j} + \rho_{ij} C_{x_i} C_{x_j} \text{ si } i,j = p+1,\ldots,k$$

$$d_{ij} = C_y - \rho_i C_y C_{x_i} + \rho_j C_y C_{x_j} - \rho_{ij} C_{x_i} C_{x_j} \text{ si } i = 1,\ldots,p \; ; \; j = p+1,\ldots,k$$

Asumiendo que la matriz de coeficientes de variación es definida positiva, el vector de pesos óptimos viene dado por

$$W = \frac{D^{-1}e}{e'D^{-1}e}$$

donde $e_{(k\times 1)} = (1,\ldots,1)'$ y el valor mínimo del error cuadrático medio es:

$$\text{ECM}\,(\overline{y}_{RM}) \simeq \overline{Y}^2 \frac{1-f}{n} \frac{1}{e'D^{-1}e}$$

Evidentemente este estimador va a tener los mismos inconvenientes que el de *Olkin* especialmente en cuanto a su sesgo y a que su óptimo no se puede calcular al depender de parámetros poblacionales en general desconocidos. Para resolver algunos de estos problemas proponemos un estimador alternativo en la siguiente sección.

§3.3 Estimador iterado de razón–producto.

El método empleado por *Rao* y *Muldholkar* para construir estimadores múltiples combinando variables con correlación positiva y con correlación negativa requiere el conocimiento de ciertos parámetros, en general desconocidos, como los coeficientes de variación entre todas las variables. Además no garantiza que la precisión sea mayor que la de los estimadores de expansión simple.

Para resolver estos problemas vamos a definir un estimador de razón–producto que se va a poder construir siempre, puesto que no va a depender de ningún parámetro desconocido y es siempre como mínimo igual de preciso que el de expansión simple.

3.3.1 Definición del estimador

Sea una población de tamaño N de la que extraemos una muestra aleatoria simple de tamaño n.

Sean $x_1, x_2, \ldots, x_k$ las variables auxiliares disponibles e y la variable principal.

Dada la variable auxiliar x_i, el estimador de razón o el estimador producto que se construye a partir de ella, $\overline{y}_{R_i} = \overline{y}\frac{\overline{X}_i}{\overline{x}_i}$ o $\overline{y}_{P_i} = \overline{y}\frac{\overline{x}_i}{\overline{X}_i}$, es mejor (bajo ciertas condiciones) que el estimador de expansión simple, $\overline{y}$. Además, la correlación entre la variable x_i y la variable principal y nos indica si es mejor construir el estimador de razón o el producto. Suponiendo x_i correlada positivamente con y, construimos $\overline{y}_{R_i}$.

Suponiendo ahora la variable auxiliar x_j correlada negativamente con la variable principal, podemos considerar $\overline{y}_{R_i}$ para construir el estimador producto en vez de $\overline{y}$, y por tanto definiríamos

$$\overline{y}_{RP_{i,j}} = \overline{y}_{R_i}\frac{\overline{x}_j}{\overline{X}_j} = \overline{y}\frac{\overline{X}_i}{\overline{x}_i}\frac{\overline{x}_j}{\overline{X}_j}$$

Ahora bien, si este estimador es mejor que $\overline{y}$, podemos repetir el procedimiento sustituyéndo en la expresión del estimador de razón o producto (según sea positiva o negativa la correlación) que se construye con la variable auxiliar x_l y continuar el proceso hasta agotar las k variables auxiliares, obteniendo, reordenando las variables si fuese necesario, el estimador iterado de razón–producto de la media a partir de las variables $x_1, x_2, \ldots, x_k$

$$\overline{y}_{IRP} = \overline{y}_{RP_{1,2,\ldots,k}} = \overline{y}\frac{\overline{X}_1}{\overline{x}_1}\frac{\overline{X}_2}{\overline{x}_2}\cdots\frac{\overline{X}_p}{\overline{x}_p}\frac{\overline{x}_{p+1}}{\overline{X}_{p+1}}\frac{\overline{x}_{p+2}}{\overline{X}_{p+2}}\cdots\frac{\overline{x}_k}{\overline{X}_k} \tag{3.3.1}$$

donde $x_1, x_2, \ldots, x_p$ están correladas positivamente con la variable principal y $x_{p+1}, x_{p+2}, \ldots, x_k$ están correladas negativamente con la variable principal, y.

3.3.2 Propiedades.

Consistencia.

Proposición 3.3.1 *El estimador $\overline{y}_{IRP}$ es consistente, en el sentido de consistencia en poblaciones finitas.*

Demostración.-
Es inmediato que el estimador es consistente puesto que para $n = N$ se tiene

$$\overline{y}_{IRP} = \overline{Y}$$

Distribución asintótica del estimador iterado.

Teorema 3.3.1 *Si las variables* $\{y_\nu\}$ *y* $\{x_{i\nu}\}$ $(i = 1, \ldots, k)$ *satisfacen la condición de Lindeberg–Hájek, entonces para un muestreo aleatorio simple el estimador iterado de razón–producto es asintóticamente normal.*

Demostración.-
Es análoga a la del teorema 1.3.1 considerando la función

$$H\left(\overline{y}_\nu, \overline{x}_{1\nu}, \overline{x}_{2\nu}, \ldots, \overline{x}_{k\nu}\right) = \overline{y}_\nu \frac{\overline{X}_{1\nu}}{\overline{x}_{1\nu}} \frac{\overline{X}_{2\nu}}{\overline{x}_{2\nu}} \cdots \frac{\overline{X}_{p\nu}}{\overline{x}_{p\nu}} \frac{\overline{x}_{p+1\nu}}{\overline{X}_{p+1\nu}} \frac{\overline{x}_{p+2\nu}}{\overline{X}_{p+2\nu}} \cdots \frac{\overline{x}_{k\nu}}{\overline{X}_{k\nu}}$$

Sesgo.

El estimador iterado de razón–producto es sesgado, como lo prueba la siguiente proposición:

Proposición 3.3.2 *El estimador* $\overline{y}_{IRP}$ *es sesgado y una aproximación del sesgo viene dada por la expresión*

$$\text{sesgo}\left(\overline{y}_{IRP}\right) \simeq \frac{1-f}{n}\overline{Y}\left[-\sum_{i=1}^{k}\left|C_{yx_i}\right| + \sum_{i\neq j=1}^{k} C_{x_ix_j} + \sum_{i=1}^{p} C_{x_i}^2 - 3 \sum_{\substack{1\leq i\leq p \\ p+1\leq j\leq k}} C_{x_ix_j}\right]$$

Demostración.-
Consideramos las variables:

$$a = \frac{\overline{y} - \overline{Y}}{\overline{Y}} \; ; \quad e_i = \frac{\overline{x}_i - \overline{X}_i}{\overline{X}_i} \; i = 1, \ldots, k$$

con las que el estimador iterado de razón–producto adopta la forma:

$$\overline{y}_{IRP} = \overline{Y}(1+a)\prod_{i=1}^{p}(1+e_i)^{-1}\prod_{i=p+1}^{k}(1+e_i)$$

Si $|e_i| < 1$, $i = 1, \ldots p$, podemos desarrollar en serie de Taylor cada término $(1+e_i)^{-1}$

$$\overline{y}_{IRP} = \overline{Y}(1+a)\prod_{i=1}^{p}\left(1 - e_i + e_i^2 + \cdots\right)\prod_{i=p+1}^{k}(1+e_i)$$

y conservando sólo los términos de grado inferior o igual a dos en a y e_i $(i = 1, \ldots, k)$, obtenemos la aproximación:

$$\overline{y}_{IRP} \simeq \overline{Y}\left(1 + a - \sum_{i=1}^{p} e_i + \sum_{i=p+1}^{k} e_i - \sum_{i=1}^{p} ae_i + \right.$$

$$\left. + \sum_{i=p+1}^{k} ae_i + \sum_{i\neq j=1}^{k} e_i e_j - 3\sum_{\substack{1\leq i\leq p \\ p+1\leq j\leq k}} e_i e_j + \sum_{i=1}^{p} e_i^2\right) \qquad (3.3.2)$$

Entonces

$$\text{sesgo}\,(\overline{y}_{IRP}) = E\left[\overline{y}_{IRP} - \overline{Y}\right] \simeq \overline{Y}\left[-\sum_{i=1}^{k}|E(ae_i)| + \right.$$

$$\left. + \sum_{i\neq j=1}^{k} E\left(e_i e_j\right) + \sum_{i=1}^{p} E\left(e_i^2\right) - 3\sum_{\substack{1\leq i\leq p \\ p+1\leq j\leq k}} E\left(e_i e_j\right)\right]$$

sustituyendo los valores de

$$E\left(e_i e_j\right) = \frac{\text{Cov}\left(\overline{x}_i, \overline{x}_j\right)}{\overline{X}_i \overline{X}_j} \;;\quad E(ae_i) = \frac{\text{Cov}\,(\overline{y}, \overline{x}_i)}{\overline{Y}\,\overline{X}_i} \;;\quad E\left(e_i^2\right) = \frac{V(\overline{x}_i)}{\overline{X}_i^2} \qquad (3.3.3)$$

en un muestreo aleatorio simple se tiene la siguiente expresión aproximada del sesgo

$$\text{sesgo}\,(\overline{y}_{IRP}) \simeq \frac{1-f}{n}\overline{Y}\left[-\sum_{i=1}^{k}\left|C_{yx_i}\right| + \sum_{i\neq j=1}^{k} C_{x_i x_j} + \sum_{i=1}^{p} C_{x_i}^2 - 3\sum_{\substack{1\leq i\leq p \\ p+1\leq j\leq k}} C_{x_i x_j}\right]$$

Error cuadrático medio.

Estudiemos el error cuadrático medio del estimador.

Proposición 3.3.3 *Una aproximación del error cuadrático medio del estimador* $\overline{y}_{IRP}$ *viene dada por la expresión*

$$\mathrm{ECM}\left(\overline{y}_{IRP}\right) \simeq \overline{Y}^2 \frac{1-f}{n}\left[C_y^2 + \sum_{i=1}^{k} C_{x_i}^2 + \sum_{i\neq j=1}^{k} C_{x_i x_j} - 2\sum_{i=1}^{k}\left|C_{yx_i}\right| - 4\sum_{\substack{1\leq i\leq p \\ p+1\leq j\leq k}} C_{x_i x_j}\right]$$

Demostración.-
Según la aproximación (3.3.2) podemos escribir:

$$\mathrm{ECM}\left(\overline{y}_{IRP}\right) = E\left[\overline{y}_{IRP} - \overline{Y}\right]^2 \simeq$$

$$\simeq \overline{Y}^2 E\left[a^2 + \sum_{i=1}^{k} e_i^2 - 2\sum_{i=1}^{p} a e_i + 2\sum_{i=p+1}^{k} a e_i + \sum_{i\neq j=1}^{k} e_i e_j - 4\sum_{\substack{1\leq i\leq p \\ p+1\leq j\leq k}} e_i e_j\right]$$

reteniendo sólo los términos de orden inferior o igual a dos en a y e_i, y sustituyendo los valores deducidos de (3.3.3) para muestreo aleatorio simple obtenemos la expresión:

$$\mathrm{ECM}\left(\overline{y}_{IRP}\right) \simeq \overline{Y}^2 \frac{1-f}{n}\left[C_y^2 + \sum_{i=1}^{k} C_{x_i}^2 + \sum_{i\neq j=1}^{k} C_{x_i x_j} - 2\sum_{i=1}^{k}\left|C_{yx_i}\right| - 4\sum_{\substack{1\leq i\leq p \\ p+1\leq j\leq k}} C_{x_i x_j}\right]$$

3.3.3 Determinación del estimador óptimo.

Estudiamos el problema de determinar si es mejor utilizar todas las variables auxiliares o habrá algún paso en el que no convenga hacer una nueva iteración. El problema es importante pues es claro que el estimador de razón o el estimador producto no siempre mejoran el estimador de expansión simple, y por tanto el pasar de $\overline{y}_{RP_{1,2,\ldots,k-1}}$ a $\overline{y}_{RP_{1,2,\ldots,k}}$ no tiene por qué producir estimadores más precisos.

Sean $x_1, x_2, \ldots x_k$ variables auxiliares y sean $\overline{y}_{RP_{1,2,\ldots,k-1}}$ e $\overline{y}_{RP_{1,2,\ldots,k}}$ los estimadores iterados de razón-producto construidos a partir de las variables $x_1, x_2, \ldots x_{k-1}$ y $x_1, x_2, \ldots x_k$, respectivamente.

De la aproximación del error cuadrático medio se deduce la expresión:

$$\mathrm{ECM}\left(\overline{y}_{RP_{1,2,\ldots,k}}\right) \simeq \mathrm{ECM}\left(\overline{y}_{RP_{1,2,\ldots,k-1}}\right) +$$

$$+\overline{Y}^2 \frac{1-f}{n}\left[-2\left|C_{yx_k}\right| + C_{x_k}^2 + 2\sum_{i=1}^{k-1} C_{x_i x_k} - 4G(x_k)\right]$$

con

$$G(z) = \begin{cases} \sum_{i=1}^{p} C_{x_i z} & \text{si } \rho_{yz} < 0 \\ \sum_{i=p+1}^{k-1} C_{x_i z} & \text{si } \rho_{yz} > 0 \end{cases}$$

donde se supone que se han realizado $k-1$ iteraciones: $i = 1, \ldots, p$ de correlación positiva con la variable principal e $i = p+1, \ldots, k-1$ de correlación negativa.

Esta expresión da las condiciones para las cuales el estimador iterado con k variables es mejor que el que utiliza sólo $k-1$ variables. Así el estimador $\overline{y}_{RP_{1,2,\ldots,k}}$ es más preciso que el estimador $\overline{y}_{RP_{1,2,\ldots,k-1}}$ si se verifica:

$$2\sum_{i=1}^{k-1} C_{x_i x_k} + C_{x_k}^2 < 2\left|C_{yx_k}\right| + 4\sum_{i=1}^{t} C_{x_i x_j} \qquad \text{si } \rho_{yx_k} < 0,$$

o bien

$$2\sum_{i=1}^{k-1} C_{x_i x_k} + C_{x_k}^2 < 2\left|C_{yx_k}\right| + 4\sum_{i=t+1}^{k} C_{x_i x_j} \qquad \text{si } \rho_{yx_k} > 0.$$

Consideramos ahora el problema de qué variable auxiliar elegir entre las disponibles para realizar una nueva iteración. Es decir, si

$$\overline{y}_{RP_{1,2,\ldots,u}} = \overline{y}\prod_{j=1}^{p}\frac{\overline{X}_j}{\overline{x}_j}\prod_{j=p+1}^{u}\frac{\overline{x}_j}{\overline{X}_j}$$

y notando las u iteraciones anteriores de la misma forma, $i = 1, \ldots, p$ de correlación positiva con la variable principal e $i = p+1, \ldots, k-1$ de correlación

negativa, aunque podrían no haberse utilizado todas las variables de cada tipo de correlación, ¿cuál de las $k-u$ variables que no han sido utilizadas será la mejor para construir el nuevo estimador? La contestación a esta pregunta se deduce también de la expresión aproximada del error cuadrático medio:

Será más conveniente el uso de la variable x_p que de la variable x_q si

$$F\left(x_p\right) < F\left(x_q\right)$$

donde

$$F(z) = -2\left|C_{yz}\right| + C_z^2 + 2\sum_{i=1}^{u} C_{x_i z} - 4G(z)$$

con

$$G(z) = \begin{cases} \sum_{i=1}^{p} C_{x_i z} & \text{si} \quad \rho_{yz} < 0 \\ \sum_{i=p+1}^{u} C_{x_i z} & \text{si} \quad \rho_{yz} > 0 \end{cases}$$

Por tanto serían tanto más deseables las variables x_p en cuanto que

1. Tengan un menor coeficiente de variación.
2. La covarianza con el resto de las variables auxiliares sea más pequeña, en valor absoluto.
3. La covarianza con la variable principal sea más grande, en valor absoluto.

En vista de los resultados anteriores parece lógico que para plantear el método iterado de razón–producto de forma óptima en el sentido de menor error cuadrático medio, hemos de considerar dos cuestiones:

1. En qué orden seleccionar las variables
2. Hasta cuándo continuar el proceso

Esto nos ha llevado al siguiente método forward:

3.3.4 Método forward.

Para conseguir el estimador iterado de forma óptima en el sentido de menor error cuadrático medio, debemos proceder de la forma siguiente:

1. Antes de realizar una nueva iteración, elegir la variable x_p tal que la función $F(z)$ sea menor (siendo x_j, $j = 1, \ldots, u$, las variables utilizadas en las iteraciones anteriores)

2. Una nueva iteración será conveniente si

$$F\left(x_p\right) < 0.$$

Esta función $F(z)$ depende de los coeficientes de variación que en general son desconocidos. No obstante se pueden estimar por sus valores muestrales.

3.3.5 Mejoras.

El método forward para la determinación del estimador óptimo tiene una serie de ventajas, como son:

1. El estimador iterado no depende de ningún parámetro desconocido y por tanto se puede calcular siempre.

2. El procedimiento iterativo para construir el estimador y evaluar su error cuadrático medio es computacionalmente sencillo.

3. Es siempre al menos tan preciso como el de expansión simple.

4. No utiliza todas las variables disponibles sino sólo las que producen una mejora en la precisión de la estimación. Además decide en cada paso la mejor variable para realizar la iteración, qué estimador, razón o producto, construir e itera si este estimador mejora en precisión al anterior.

5. En el caso particular en que todas las variables estén correladas positivamente con la variable principal se obtiene el estimador iterado de razón.

6. En el caso particular en que todas las variables estén correladas negativamente con la variable principal se obtiene un estimador iterado de producto.

3.3.6 Estimador iterado de razón–producto insesgado.

A continuación modificamos el estimador iterado de razón–producto mediante la técnica jackknife para hacerlo insesgado.

El estimador iterado de razón-producto se va construyendo mediante un procedimiento iterativo que parte del estimador de expansión simple y elige en cada paso la variable que al iterar produzca un estimador más preciso. El estimador así obtenido

- no depende de ningún parámetro desconocido (y por tanto se puede calcular siempre),
- es siempre al menos tan preciso como el estimador de expansión simple y
- no utiliza todas las variables disponibles sino sólo las que producen una mejora en la precisión del estimador.

Así pues, este estimador resuelve los dos inconvenientes considerados del estimador múltiple dado por *Rao* y *Mudholkar*. Sin embargo, ambos estimadores son sesgados.

A continuación vamos a utilizar la técnica jackknife introducida por *Quenouille* (1956) para modificar el estimador iterado de razón–producto haciendo que sea insesgado.

Formación del estimador iterado de razón–producto insesgado.

Consideremos una muestra de tamaño n extraida mediante muestreo aleatorio simple de una población finita de tamaño N. Sea el estimador de razón–producto iterado dado en 3.3.1 (haciendo la observación que para simplificar la notación en la definición de este estimador intervienen sólo las variables seleccionadas por el anterior método forward):

$$\overline{y}_{IRP} = \overline{y} \prod_{i=1}^{p} \frac{\overline{X}_i}{\overline{x}_i} \prod_{i=p+1}^{k} \frac{\overline{x}_i}{\overline{X}_i} \quad \text{con} \quad \begin{array}{lll} \rho_{yx_i} & > & 0 \quad i = 1, \ldots, p \\ \rho_{yx_i} & < & 0 \quad i = p+1, \ldots, k \end{array}$$

Según la proposición 3.3.2, este estimador es sesgado y una expresión aproximada de su sesgo es

$$\text{sesgo}\,(\overline{y}_{IRP}) \simeq \frac{1-f}{n}\overline{Y}A$$

con

$$A = \left[-\sum_{i=1}^{k} \left| C_{yx_i} \right| + \sum_{i \neq j=1}^{k} C_{x_i x_j} + \sum_{i=1}^{p} C_{x_i}^2 - 3 \sum_{\substack{1 \leq i \leq p \\ p+1 \leq j \leq k}} C_{x_i x_j} \right] \tag{3.3.4}$$

Consideremos ahora la muestra dividida aleatoriamente en g grupos de tamaño m ($n = mg$). Definimos el estimador iterado de razón–producto jackknife de la forma:

$$\overline{y}_{IRPJ} = \frac{1}{g}\sum_{j=1}^{g} \overline{y}_{IRP(j)} \quad \text{con} \quad \overline{y}_{IRP(j)} = \overline{y}_{(j)} \prod_{i=1}^{p} \frac{\overline{X}_i}{\overline{x}_{i(j)}} \prod_{i=p+1}^{k} \frac{\overline{x}_{i(j)}}{\overline{X}_i}$$

siendo $\overline{y}_{(j)}$ y $\overline{x}_{i(j)}$ las medias muestrales basadas en las $n-m$ unidades obtenidas omitiendo el j–ésimo grupo. Por tanto:

$$\overline{y}_{(j)} = \frac{n\overline{y} - m\overline{y}_j}{n-m}$$

donde $\overline{y}_j$ es la media muestral del subgrupo j.

Calculamos seguidamente el sesgo de este nuevo estimador de forma aproximada:

$$\text{sesgo}\,(\overline{y}_{IRPJ}) = \frac{1}{g}\sum_{i=1}^{g} \text{sesgo}\left(\overline{y}_{IRP(j)}\right) \simeq \frac{N-(n-m)}{N(n-m)}\overline{Y}A$$

Entonces

$$\text{sesgo}\,(\overline{y}_{IRP}) = \frac{(N-n)(n-m)}{n(N-n+m)}\,\text{sesgo}\,(\overline{y}_{IRPJ})$$

Definimos el estimador iterado de razón–producto insesgado

$$\overline{y}_{IRPI} = \frac{N-n+m}{N} g\overline{y}_{IRP} - \frac{N-n}{N}(g-1)\overline{y}_{IRPJ}$$

que es insesgado como se puede comprobar fácilmente.

Precisión.

Estudiemos su precisión.

Proposición 3.3.4 *En un primer grado de aproximación, el estimador iterado de razón-producto, $\overline{y}_{IRP}$, el estimador iterado de razón-producto jackknife, $\overline{y}_{IRPJ}$ y el estimador iterado de razón-producto insesgado, $\overline{y}_{IRPI}$ tienen la misma precisión.*

Demostración.-

$$V\left(\overline{y}_{IRPI}\right) = \left(\frac{N-n+m}{N}\right)^2 g^2 \,\mathrm{ECM}\left(\overline{y}_{IRP}\right) +$$

$$+ \left(\frac{N-n}{N}\right)^2 (g-1)^2 \,\mathrm{ECM}\left(\overline{y}_{IRPJ}\right) -$$

$$- 2\frac{N-n+m}{N}\,\frac{N-n}{N} g(g-1) E\left(\overline{y}_{IRPJ} - \overline{Y}\right)\left(\overline{y}_{IRP} - \overline{Y}\right) \qquad (3.3.5)$$

En la proposición 3.3.3 comprobamos que una aproximación del error cuadrático medio del estimador $\overline{y}_{IRP}$ es

$$\mathrm{ECM}\left(\overline{y}_{IRP}\right) \simeq \overline{Y}^2 \frac{1-f}{n}\left[C_y^2 + \sum_{i=1}^{k} C_{x_i}^2 + \right.$$

$$\left. + \sum_{i\neq j=1}^{k} C_{x_i x_j} - 2\sum_{i=1}^{k}\left|C_{y x_i}\right| - 4 \sum_{\substack{1\leq i\leq p \\ p+1\leq j\leq k}} C_{x_i x_j}\right] \qquad (3.3.6)$$

Se puede comprobar también, siguiendo un procedimiento análogo al dado por *Sukhatme* y otros (1984) para el estimador de razón, que en un primer orden de aproximación

$$\mathrm{ECM}\left(\overline{y}_{IRPJ}\right) = E\left(\overline{y}_{IRPJ} - \overline{Y}\right)\left(\overline{y}_{IRP} - \overline{Y}\right) = \mathrm{ECM}\left(\overline{y}_{IRP}\right) \qquad (3.3.7)$$

Por tanto, sustituyendo 3.3.6 y 3.3.7 en 3.3.5 y simplificando se llega a que

$$V(\overline{y}_{IRPI}) = \frac{N-n}{Nn}\overline{Y}^2 \left[C_y^2 + \sum_{i=1}^{k} C_{x_i}^2 + \right.$$

$$\left. + \sum_{i \neq j=1}^{k} C_{x_i x_j} - 2\sum_{i=1}^{k} \left| C_{yx_i} \right| - 4 \sum_{\substack{1 \leq i \leq p \\ p+1 \leq j \leq k}} C_{x_i x_j} \right]$$

y por tanto los estimadores $\overline{y}_{IRP}$, $\overline{y}_{IRPJ}$ y $\overline{y}_{IRPI}$ tienen igual precisión.

3.3.7 Mejoras.

Hemos obtenido un estimador indirecto que utiliza información auxiliar de variables con correlación de cualquier signo con la variable principal y con las siguientes características:

1. Es insesgado.
2. Es siempre más preciso que el estimador de expansión simple.
3. Es siempre más preciso que el estimador de razón, el estimador de producto y el estimador de razón–producto que se pueda construir utilizando cualquiera de las variables auxiliares.
4. No depende de ningún parámetro desconocido, por lo cual se puede calcular siempre.
5. Para el caso $p = 1$ y $k = 1$, el estimador coincide con el estimador de razón jackknife $\overline{y}_Q$ considerado por *Sukhatme* et al. (1984).
6. Para el caso $p = 1$ y $k = 2$, el estimador coincide con un caso particular de los estimadores propuestos por *Singh* y *Biradar* (1992).

Por último hacer notar que todo este estudio de extensión múltiple se ha hecho bajo la suposición de trabajar con un muestreo aleatorio simple, puesto que es el diseño muestral más simple y sirve de base para comparar la eficiencia de diseños muestrales más complejos. Las ideas que se proponen para la construcción de estimadores pueden utilizarse también para otros diseños muestrales, dando lugar a nuevos estimadores que bajo ciertas condiciones mejorarán los respectivos estimadores simples de cada diseño muestral.

Parte II

Mejora de estimadores de la varianza.

Capítulo 4

Estimadores univariantes.

§4.1 Introducción.

El uso de información auxiliar proporcionada por una variable x altamente correlada con la variable principal y es común en la estimación de medias o totales poblacionales. Los estimadores de razón y regresión utilizan la información que proporciona la variable x para modificar los estimadores directos, consiguiendo estimadores más precisos del parámetro en cuestión.

De una forma similar, es razonable suponer que estas técnicas de estimación se puedan utilizar, bajo las condiciones adecuadas, para proporcionar estimadores eficientes de la varianza.

Fuller (1970) propone un estimador de regresión de la varianza del estimador de Horvitz–Thompson del total poblacional usando como variable auxiliar x, las cantidades $\pi_i\pi_j - \pi_{ij}$ y $(\pi_i\pi_j - \pi_{ij})(i-j)^2$, siendo π_i y π_{ij} las probabilidades de inclusión individual y conjunta de cada unidad, respectivamente.

Ogus y *Clark* (1971) proponen el uso de estimadores de razón y diferencia de la varianza bajo un diseño de muestreo de Poisson, con el propósito de reducir el efecto del tamaño muestral aleatorio, en la estimación de la varianza.

Bajo un diseño muestral en el que se selecciona una unidad en cada estrato con probabilidad proporcional al tamaño (PPS), *Hansen*, *Hurwitz* y *Madow* (1953) proponen el uso de una variable correlada junto con la técnica de estratos colapsados, para estimar la varianza, y prueban que el estimador está positivamente sesgado.

Bajo el mismo diseño muestral, *Hartley*, *Rao* y *Kiefer* (1969) proponen un estimador de la varianza basado en suponer una buena regresión entre las

verdaderas medias de los estratos y algunas variables auxiliares. Sus ejemplos, que utilizan una sóla variable, indican una mejora considerable en términos del sesgo absoluto respecto al estimador propuesto por Hansen, Hurwitz y Madow. No obstante, el primer método es más sencillo de aplicar. Además el estimador de Hartley no se ha comprobado que sea no negativo bajo todas las condiciones.

Posteriormente, *Shapire* y *Bateman* (1978) consideran la reducción del sesgo del estimador de la varianza en un diseño con una unidad por estrato, utilizando como estimador de la varianza el estimador de Yates–Grundy, para un diseño con dos unidades por estrato con probabilidades π_{ij} calculadas en base a un esquema de muestreo de *Durbin* (1967).

En este capítulo vamos a estudiar el problema de definir estimadores indirectos de la varianza poblacional, utilizando la información que proporciona alguna variable auxiliar x para la cual es conocida su varianza poblacional.

En la primera sección se estudia el método de estimación de regresión introducido por *Isaki* (1983) en la estimación de la varianza poblacional.

En la segunda sección se hace lo mismo con distintos estimadores de razón de la varianza: el estimador de *Isaki* (1983) para muestreo aleatorio simple y para muestreo con probabilidades iguales con reemplazamiento y las posteriores mejoras de éste debidas a *Prasad* y *Singh* (1990, 1992) para aumentar la precisión y disminuir el sesgo.

En la tercera sección presentamos el método de estimación de producto en la estimación de la varianza poblacional.

En la cuarta sección presentamos un estimador de razón alternativo al dado por *Isaki* para estimar la varianza poblacional, bajo muestreo con reemplazo, con las propiedades de ser siempre más preciso que el estimador de expansión simple (la cuasivarianza muestral), igual de preciso que el estimador de regresión y englobando como caso particular al estimador de *Isaki* y al estimador producto.

Finalmente, en la última sección comparamos algunos de los métodos indirectos de estimación de la varianza en general y particularizamos a poblaciones "normales".

§4.2 Estimación de regresión.

Consideremos una población finita $U = (U_1, \ldots, U_N)$ de unidades de las cuales nos interesa el estudio de una variable principal y. Supongamos que se extrae una muestra aleatoria simple (suponiendo grande el tamaño de la población para poder prescindir del factor de corrección por finitud) o una muestra aleatoria con reemplazo, de tamaño n, en cuyas unidades se observa la variable y. El problema es estimar la varianza poblacional

$$S_y^2 = \frac{N}{N-1}\sigma_y^2 \quad \text{con} \quad \sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2 .$$

El método de estimación directa proporciona un estimador, la varianza muestral, dado por

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 .$$

4.2.1 Definición del estimador de regresión.

Sea x una variable auxiliar correlada con y, de la que se conoce la varianza poblacional

$$S_x^2 = \frac{N}{N-1}\sigma_x^2 \quad \text{con} \quad \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2 .$$

Si escribimos

$$\beta_2(y) = \frac{\mu_{40}}{\mu_{20}^2} \quad ; \quad \beta_2(x) = \frac{\mu_{04}}{\mu_{02}^2} \quad ; \quad \theta = \frac{\mu_{22}}{\mu_{20}\mu_{02}}$$

donde

$$\mu_{rs} = \frac{1}{N}\sum\left(y_j - \overline{Y}\right)^r\left(x_j - \overline{X}\right)^s$$

en un primer grado de aproximación (véase *Kendall* y *Stuart* (1977)):

$$V\left(s_y^2\right) \simeq \frac{1}{n}S_y^4\left(\beta_2(y) - 1\right) \quad ; \quad V\left(s_x^2\right) \simeq \frac{1}{n}S_x^4\left(\beta_2(x) - 1\right)$$

y

$$\text{Cov}\left(s_y^2, s_x^2\right) \simeq \frac{S_y^2 S_x^2}{n}(\theta - 1).$$

Definición 4.2.1 *Se define el estimador de regresión de la varianza poblacional de la forma*

$$\widehat{S}_{reg}^2 = s_y^2 + b\left(S_x^2 - s_x^2\right)$$

donde s_y^2 y s_x^2 son las varianzas muestrales de las variables x e y, respectivamente.

4.2.2 Propiedades.

Consistencia.

Obvia, pues para $n = N$, $\widehat{S}_{reg}^2 = S_y^2$.

Sesgo.

El estimador $\widehat{S}_{reg}^2$ es insesgado puesto que $E\left(\widehat{S}_{reg}^2\right) = S_y^2$.

Precisión.

Para medir su precisión comprobamos que su varianza aproximada es:

$$V\left(\widehat{S}_{reg}^2\right) \simeq \frac{1}{n}\left(S_y^4(\beta_2(y) - 1) + b^2 S_x^4(\beta_2(x) - 1) - 2bS_y^2 S_x^2(\theta - 1)\right).$$

Escribimos

$$V\left(\widehat{S}_{reg}^2\right) = V\left(s_y^2\right) + b^2 V\left(S_x^2 - s_x^2\right) + 2b\,\text{Cov}\left(s_y^2, S_x^2 - s_x^2\right) =$$

$$= V\left(s_y^2\right) + b^2 V\left(s_x^2\right) - 2b\,\text{Cov}\left(s_y^2, s_x^2\right) \simeq$$

$$= \frac{1}{n}S_y^4(\beta_2(y) - 1) + b^2\frac{1}{n}S_x^4(\beta_2(x) - 1) - 2b\frac{1}{n}S_y^2 S_x^2(\theta - 1) =$$

$$= \frac{1}{n}\left(S_y^4(\beta_2(y) - 1) + b^2 S_x^4(\beta_2(x) - 1) - 2bS_y^2 S_x^2(\theta - 1)\right)$$

Además, en el caso general, el valor de b que minimiza la varianza viene dado por la expresión:

$$b = \frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)}{V\left(s_x^2\right)} = \frac{S_y^2(\theta - 1)}{S_x^2(\beta_2(x) - 1)} \tag{4.2.1}$$

valor que se obtiene de

$$\frac{\partial V\left(\widehat{S}_{reg}^2\right)}{\partial b} = 2bV\left(s_x^2\right) - 2\,\mathrm{Cov}\left(s_y^2, s_x^2\right) = 0$$

y de

$$\frac{\partial^2 V\left(\widehat{S}_{reg}^2\right)}{\partial b^2} = 2V\left(s_x^2\right) \geq 0.$$

Así el estimador óptimo viene dado por la expresión

$$\widehat{S}_{reg_{b_0}}^2 = s_y^2 + \frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)}{V\left(s_x^2\right)}\left(S_x^2 - s_x^2\right) \tag{4.2.2}$$

Para este valor la varianza tiene por mínimo

$$V_{mín}\left(\widehat{S}_{reg}^2\right) \simeq \frac{1}{n} S_y^4 \left((\beta_2(y) - 1) - \frac{(\theta - 1)^2}{(\beta_2(x) - 1)}\right)$$

pues sustituyendo se obtiene

$$V_{mín}\left(\widehat{S}_{reg}^2\right) \simeq \frac{1}{n}\left(S_y^4\left(\beta_2(y) - 1\right) +\right.$$

$$\left. + \frac{S_y^4(\theta - 1)^2}{S_x^4(\beta_2(x) - 1)^2} S_x^4\left(\beta_2(x) - 1\right) - 2\frac{S_y^2(\theta - 1)}{S_x^2(\beta_2(x) - 1)} S_y^2 S_x^2(\theta - 1)\right) =$$

$$= \frac{1}{n} S_y^4 \left((\beta_2(y) - 1) + \frac{(\theta - 1)^2}{(\beta_2(x) - 1)} - 2\frac{(\theta - 1)^2}{(\beta_2(x) - 1)}\right) =$$

$$= \frac{1}{n} S_y^4 \left((\beta_2(y) - 1) - \frac{(\theta - 1)^2}{(\beta_2(x) - 1)}\right).$$

Precisión en caso de normalidad.

Comprobamos que si los momentos de la distribución (y, x) son los mismos de una normal bivariante hasta el orden cuatro el valor mínimo de la varianza del estimador de regresión es

$$Vmín\left(\widehat{S}^2_{reg}\right) \simeq \frac{2}{n} S_y^4 \left(1-\rho^4\right),$$

donde $\rho = \rho(x, y)$.

En efecto, si los momentos de la distribución (y, x) son los mismos de una normal bivariante hasta el orden cuatro se tiene

$$\theta = \frac{\mu_{22}}{\mu_{20}\mu_{02}} = 1 + 2\rho^2 \;\; ; \;\; \beta_2(x) = \frac{\mu_{04}}{\mu_{02}^2} = 3$$

y entonces

$$b = \frac{S_y^2(\theta - 1)}{S_x^2(\beta_2(x) - 1)} = \frac{S_y^2\left(1 + 2\rho^2 - 1\right)}{S_x^2\left(3 - 1\right)} = \frac{S_y^2}{S_x^2}\rho^2 = \beta^2$$

siendo β el coeficiente de regresión de y sobre x.

En este caso

$$Vmín\left(\widehat{S}^2_{reg}\right) \simeq \frac{1}{n} S_y^4 \left(2 - \frac{4\rho^4}{2}\right) = \frac{2}{n} S_y^4 \left(1-\rho^4\right)$$

Comparación con el estimador de expansión simple.

Comprobamos que el estimador de regresión óptimo es siempre más preciso que el estimador de expansión simple, comparando sus respectivas varianzas.

En el caso más general

$$Vmín\left(\widehat{S}^2_{reg}\right) - V\left(s_y^2\right) \simeq \frac{1}{n} S_y^4 \left(-\frac{(\theta-1)^2}{(\beta_2(x)-1)}\right) \leq 0,$$

puesto que $\beta_2(x) \geq 1$.

Notar que si hay "normalidad", ambos estimadores son igual de precisos en caso de independencia de las variables puesto que

$$Vmín\left(\widehat{S}^2_{reg}\right) - V\left(s_y^2\right) = \frac{2}{n} S_y^4(-\rho^4) = 0.$$

§4.3 Estimación de razón.

Dedicamos esta sección al estudio de los estimadores indirectos de razón de la varianza poblacional mas usuales. Entre ellos, el estimador de *Isaki* (1983) en diseños de probabilidades iguales sin y con reemplazamiento y los estimadores dados por *Prasad* y *Singh* (1990, 1992).

4.3.1 El estimador de *Isaki* bajo diseño SRSWOR.

Supongamos que una muestra de tamaño n se selecciona mediante un diseño SRSWOR (muestreo aleatorio simple sin reemplazamiento), en la cual se observa la variable principal y, y una variable auxiliar x. Supongamos además que la varianza poblacional de x, S_x^2, es conocida.

El problema es estimar la varianza poblacional

$$S_y^2 = \frac{N}{N-1}\sigma_y^2 \quad \text{con} \quad \sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2.$$

Puesto que podemos expresar en forma alternativa

$$\sigma_y^2 = \frac{1}{N(N-1)}\sum_{i\neq j}^{N}\left(y_i - y_j\right)^2,$$

parece razonable pensar que si la relación entre los pares $(y_i - y_j)^2$ y $(x_i - x_j)^2$ es grande, entonces la utilización del cociente $\frac{S_x^2}{s_x^2}$ en la modificación del estimador directo de S_y^2 (s_y^2) puede producir mejoras en la estimación de la varianza de y, al igual que ocurre en el caso de la estimación de razón para la media o el total de y.

De esta manera se obtiene el estimador de razón de la varianza poblacional propuesto por *Isaki* (1983):

$$\widehat{S}_I^2 = s_y^2 \frac{S_x^2}{s_x^2}. \tag{4.3.1}$$

A continuación procedemos al estudio de sus propiedades bajo el diseño muestral dado (SRSWOR), que nos servirán para proponer modificaciones del método que supongan, bajo ciertas condiciones, mejoras bien en su sesgo o en su precisión.

Propiedades.

Consistencia.

Es clara la consistencia del estimador, en el sentido de consistencia en poblaciones finitas.

Sesgo.

El estimador de Isaki es sesgado y su sesgo exacto es

$$\text{sesgo}\left(\widehat{S}_I^2\right) = -\operatorname{Cov}\left(\frac{s_y^2}{s_x^2}, s_x^2\right). \tag{4.3.2}$$

En efecto,

$$\text{sesgo}\left(\widehat{S}_I^2\right) = E\left(s_y^2\frac{S_x^2}{s_x^2}\right) - S_y^2 = E\left(\frac{s_y^2}{s_x^2}\right)E\left(s_x^2\right) - E\left(s_y^2\right) = -\operatorname{Cov}\left(\frac{s_y^2}{s_x^2}, s_x^2\right).$$

Sin embargo, esta expresión exacta del sesgo es poco práctica puesto que la covarianza que aparece en la expresión es desconocida. A continuación damos una expresión aproximada del sesgo, más util desde el punto de vista práctico.

Una expresión aproximada del sesgo del estimador $\widehat{S}_I^2$ es, continuando con la notación de la sección anterior,

$$\text{sesgo}\left(\widehat{S}_I^2\right) \simeq \frac{(N-n)}{(N-2)n} K S_y^2 \left(\beta_2(x) - \theta\right),$$

donde

$$K = \frac{(N-1)(Nn-N-n-1)}{(n-1)N(N-3)}. \tag{4.3.3}$$

Para comprobarlo, consideramos las variables

$$e_0 = \frac{s_y^2 - S_y^2}{S_y^2}; \qquad e_1 = \frac{s_x^2 - S_x^2}{S_x^2},$$

mediante las cuales podemos expresar el estimador de razón en la forma

$$\widehat{S}_I^2 = S_y^2 \frac{(1+e_0)}{(1+e_1)},$$

que podemos desarrollar en serie de Taylor $(1+e_1)$ si $\mid e_1 \mid < 1$, (en este caso la condición requerida es equivalente a que se verifique la condición $s_x^2 < 2S_x^2$ para todas las posibles muestras de tamaño n) obteniendo

$$\widehat{S}_I^2 = S_y^2(1+e_0)(1-e_1+e_1^2-\cdots).$$

Reteniendo ahora los términos de orden inferior o igual a dos en e_0 y e_1,

$$\widehat{S}_I^2 \simeq S_y^2(1-e_1+e_1^2+e_0-e_1e_0),$$

y tomando esperanzas queda:

$$E\left(\widehat{S}_I^2\right) \simeq S_y^2\left(E\left(e_1^2\right)-E\left(e_1e_0\right)\right).$$

Ahora bien, siguiendo el método dado por *Kendall* y *Stuart* (1977), se tiene que en muestreo aleatorio simple

$$E\left(e_1^2\right) = \frac{V(s_x^2)}{S_x^4} = \frac{N-n}{(N-2)n}\left(K\beta_2(x)-M\right),$$

$$E\left(e_0e_1\right) = \frac{\mathrm{Cov}(s_y^2,s_x^2)}{S_y^2S_x^2} = \frac{N-n}{(N-2)n}\left(K\theta+K_1\theta_1-K_2\right),$$

donde K está dada en (4.3.3), $\theta_1 = \dfrac{\mu_{11}^2}{\mu_{20}\mu_{02}}$,

$$K_1 = \frac{2(N-1)(N-n-1)}{(n-1)N(N-3)},$$

$$K_2 = \frac{(N^2n-2Nn-N^2+2N-n-1)}{(n-1)N(N-3)},$$

y

$$M = \frac{(N^2n-3N^2+6N-3n-3)}{(n-1)N(N-3)}. \tag{4.3.4}$$

Sustituyendo se llega a

$$\text{sesgo}\left(\widehat{S}_I^2\right) \simeq S_y^2\frac{(N-n)}{(N-2)n)}\left(K\left(\beta_2(x)-\theta\right)-M-K_1\theta_1+K_2\right),$$

expresión que se aproxima a la que queríamos comprobar cuando la correlación entre las variables es alta, pues en tal caso

$$M + K_1\theta_1 - K_2 \simeq 0.$$

Precisión.

Para estudiar la precisión de este estimador, damos una expresión aproximada de su error cuadrático medio:

$$\mathrm{ECM}\left(\widehat{S}_I^2\right) \simeq \frac{(N-n)}{(N-2)n} K S_y^4 \left(\beta_2(y) + \beta_2(x) - 2\theta\right).$$

Para comprobarlo se procede como en el caso de la obtención de la aproximación del sesgo, obteniendo

$$\widehat{S}_I^2 - S_y^2 \simeq S_y^2\left(e_1^2 + e_0 - e_1 - e_1 e_0\right),$$

que elevando al cuadrado y reteniendo los términos de orden dos, como máximo, en e_0 y e_1, queda

$$\left(\widehat{S}_I^2 - S_y^2\right)^2 \simeq S_y^4\left(e_0^2 + e_1^2 - 2e_0e_1\right)$$

y tomando esperanzas y operando se llega a

$$\mathrm{ECM}\left(\widehat{S}_I^2\right) \simeq S_y^4\left(\frac{V\left(s_y^2\right)}{S_y^4} + \frac{V\left(s_x^2\right)}{S_x^4} - 2\frac{\mathrm{Cov}\left(s_x^2, s_y^2\right)}{S_x^2 S_y^2}\right) =$$

$$= \frac{N-n}{(N-2)n} S_y^4 \left(K\left(\beta_2(y) + \beta_2(x) - 2\theta\right) - 2\left(M + K_1\theta_1 - K_2\right)\right) \simeq$$

$$\simeq \frac{N-n}{(N-2)n} S_y^4 K \left(\beta_2(y) + \beta_2(x) - 2\theta\right),$$

si ρ_{yx} es alta.

Comparación con el estimador de expansión simple.

Veamos bajo que condiciones el estimador de razón de *Isaki* mejora en precisión al estimador de expansión simple, s_y^2. Se verifica que cuando

$$\rho(s_x^2, s_y^2) > \frac{1}{2}\frac{CV(s_x^2)}{CV(s_y^2)}$$

el estimador de Isaki es más preciso que el estimador directo, s_y^2.

En efecto, la eficiencia relativa entre los dos estimadores viene dada por la expresión

$$\frac{V\left(s_y^2\right)}{\mathrm{ECM}\left(\widehat{S}_I^2\right)} \simeq \frac{V\left(s_y^2\right)}{V\left(s_y^2\right) + S_y^4 \dfrac{V\left(s_x^2\right)}{S_x^4} - 2\,\mathrm{Cov}\left(s_x^2, s_y^2\right)\dfrac{S_y^2}{S_x^2}} =$$

$$= \frac{1}{1 + R_S^2 \frac{V(s_x^2)}{V(s_y^2)} - 2R_S \frac{\mathrm{Cov}(s_x^2, s_y^2)}{V(s_y^2)}},$$

donde $R_S = \frac{S_y^2}{S_x^2}$.

Por tanto, el estimador de razón será más eficiente si la eficiencia relativa es mayor que 1, es decir, si

$$\rho(s_x^2, s_y^2) > \frac{1}{2}\frac{CV(s_x^2)}{CV(s_y^2)}.$$

Caso particular de normalidad.

Si los momentos de la distribución (y, x) son los mismos de una normal bivariante hasta el orden cuatro, el sesgo y el error cuadrático medio del estimador de Isaki tienen por expresiones aproximadas

$$\mathrm{sesgo}\left(\widehat{S}_I^2\right) \simeq \frac{(N-n)}{(N-2)n} 2K S_y^2 \left(1 - \rho^2\right)$$

$$\mathrm{ECM}\left(\widehat{S}_I^2\right) \simeq \frac{(N-n)}{(N-2)n} 4K S_y^4 \left(1 - \rho^2\right).$$

Comparando con la precisión del estimador de expansión simple que en este caso tiene la expresión

$$V\left(s_y^2\right) = \frac{(N-n)}{(N-2)n} S_y^4 (3K - M),$$

se obtiene que el estimador de razón $\widehat{S}_I^2$ es más eficiente que s_y^2 si

$$K\left(1 - 4\rho^2\right) < -M,$$

si y sólo si

$$\rho^2 > \frac{M+K}{4K} = \frac{1}{4}\left(1 + \frac{N^2 n - 3N^2 + 6N - 3n - 3}{(N-1)(Nn - N - n - 1)}\right),$$

que conduce a la condición

$$\rho^2 > \frac{1}{2}$$

cuando N es grande.

4.3.2 El estimador de *Isaki* bajo diseño SRSWR.

Si la muestra de tamaño n de la población de tamaño N se extrae mediante un diseño de probabilidades iguales con reemplazamiento, es fácil deducir, siguiendo análogos procedimientos a los dados anteriormente, las propiedades del estimador de Isaki

$$\widehat{S}_I^2 = s_y^2 \frac{S_x^2}{s_x^2},$$

teniendo en cuenta que en este caso

$$E\left(e_1^2\right) = \frac{1}{n}(\beta_2(x) - 1); \quad E\left(e_0^2\right) = \frac{1}{n}(\beta_2(y) - 1); \quad E(e_0 e_1) = \frac{1}{n}(\theta - 1)$$

y por tanto, bajo este esquema de muestreo, podemos resumir a continuación:

Propiedades.

Consistencia.

Es clara la consistencia del estimador, en el sentido de consistencia en poblaciones finitas.

Sesgo.

Una expresión aproximada del sesgo del estimador $\widehat{S}_I^2$ es

$$\text{sesgo}\left(\widehat{S}_I^2\right) \simeq \frac{1}{n} S_y^2 \left(\beta_2(x) - \theta\right). \tag{4.3.5}$$

Precisión.

Una expresión aproximada del error cuadrático medio es:

$$\text{ECM}\left(\widehat{S}_I^2\right) \simeq \frac{1}{n} S_y^4 \left(\beta_2(y) + \beta_2(x) - 2\theta\right). \tag{4.3.6}$$

Comparación con el estimador de expansión simple.

El estimador de razón de *Isaki* mejora en precisión al estimador de expansión simple, s_y^2 cuando

$$\rho(s_x^2, s_y^2) > \frac{1}{2} \frac{CV(s_x^2)}{CV(s_y^2)}.$$

Caso particular de normalidad.

Si los momentos de la distribución (y, x) son los mismos de una normal bivariante hasta el orden cuatro, el sesgo y el error cuadrático medio del estimador de Isaki tienen por expresiones aproximadas

$$\text{sesgo}\left(\widehat{S}_I^2\right) \simeq \frac{1}{n} 2 S_y^2 \left(1 - \rho^2\right)$$

$$\text{ECM}\left(\widehat{S}_I^2\right) \simeq \frac{1}{n} 4 S_y^4 \left(1 - \rho^2\right).$$

Comparando con la precisión del estimador de expansión simple que en este caso tiene la expresión

$$V\left(s_y^2\right) = \frac{2}{n} S_y^4,$$

se obtiene que el estimador de razón $\widehat{S}_I^2$ es más eficiente que s_y^2 si y sólo si

$$\rho^2 > \frac{1}{2}.$$

4.3.3 Los estimadores de *Prasad* y *Singh*.

Al observar que el estimador de Isaki, $\widehat{S}_I^2 = s_y^2 \frac{S_x^2}{s_x^2}$ depende de s_x^2 y s_y^2, estimadores insesgados de S_x^2 y S_y^2, *Singh*, *Pandey* e *Hirano* (1973) sugieren modificar el estimador de la varianza poblacional S_y^2 de la forma $s_y^{2\,*} = As_y^2$, donde A se determina con el criterio de minimizar el error cuadrático medio del estimador $s_y^{2\,*}$. De esta forma obtienen el estimador óptimo

$$s_y^{2\,*} = \frac{1}{1 + \frac{1}{n}(\beta_2(y) - 1)} s_y^2,$$

para el que se obtiene un error cuadrático medio mínimo de

$$\text{ECM}\left(s_y^{2\,*}\right) = \frac{\frac{1}{n}(\beta_2(y) - 1)}{1 + \frac{1}{n}(\beta_2(y) - 1)} S_y^4,$$

que mejora al estimador directo s_y^2 puesto que $V\left(s_y^2\right) = \frac{1}{n} S_y^4 (\beta_2(y) - 1)$.

Por tanto, el disponer de información a priori sobre $\beta_2(y)$, el coeficiente de curtosis de y, permite construir estimadores de la varianza poblacional S_y^2 más precisos que el estimador directo s_y^2.

Disponiendo de esta información a priori, *Prasad* y *Singh* construyen un estimador tipo razón

$$\widehat{V}_R = \frac{s_y^{2\,*}}{s_x^2}$$

para el cociente de varianzas $V_R = \frac{S_y^2}{S_x^2}$ y un estimador tipo razón

$$t_2 = s_y^{2\,*} \frac{S_x^2}{s_x^2}$$

para la varianza poblacional S_y^2.

Para comparar la eficiencia de este estimador respecto a los estimadores directo y de Isaki, se estudia su error cuadrático medio. En este sentido, una aproximación de orden $O\left(n^{-1}\right)$ del error cuadrático medio del estimador t_2 como estimador de la varianza poblacional viene dada por la expresión

$$\mathrm{ECM}\left(t_2\right) \simeq \frac{1}{n} S_y^4 \left[\frac{\beta_2(y) - 1 - 2(\theta - 1)}{1 + \frac{1}{n}\left(\beta_2(y) - 1\right)} + \beta_2(x) - 1 \right].$$

Para su cálculo, escribimos el error cuadrático medio del estimador del cociente de varianzas V_R:

$$\mathrm{ECM}\left(\widehat{V}_R\right) = E\left(\widehat{V}_R - V_R\right)^2 = E\left(\frac{s_y^{2\,*}}{s_x^2} - \frac{S_y^2}{S_x^2}\right)^2 =$$

$$= E\left(\frac{s_y^{2\,*} - V_R s_x^2}{s_x^2}\right)^2 = E\left(\frac{s_y^{2\,*} - V_R s_x^2}{S_x^2\left(1 + \frac{s_x^2 - S_x^2}{S_x^2}\right)}\right)^2 =$$

$$= \frac{1}{S_x^4} E\left[\left(s_y^{2\,*} - V_R s_x^2\right)^2 \left(1 + \frac{s_x^2 - S_x^2}{S_x^2}\right)^{-2}\right] =$$

(siempre que $\left| \frac{s_x^2 - S_x^2}{S_x^2} \right| < 1$)

$$= \frac{1}{S_x^4} E\left[\left(s_y^{2\,*} - V_R s_x^2\right)^2 \left(1 - 2\left(\frac{s_x^2 - S_x^2}{S_x^2}\right) + 3\left(\frac{s_x^2 - S_x^2}{S_x^2}\right)^{-2} - \cdots\right)\right] \simeq$$

(reteniendo sólo el primer término que es de orden $O\left(n^{-1}\right)$)

$$\simeq \frac{1}{S_x^4} E\left[\left(s_y^{2\,*} - V_R s_x^2\right)^2\right] = \frac{1}{S_x^4} E\left[\left(s_y^{2\,*} - S_y^2\right) - V_R\left(s_x^2 - S_x^2\right)\right]^2 =$$

$$= \frac{1}{S_x^4}\left[E\left(s_y^{2\,*} - S_y^2\right)^2 + V_R^2 E\left(s_x^2 - S_x^2\right)^2 - 2V_R \,\mathrm{Cov}\left(s_y^{2\,*}, s_x^2\right)\right] =$$

$$= \frac{1}{S_x^4}\left[E\left(s_y^{2\,*} - S_y^2\right)^2 + V_R^2 E\left(s_x^2 - S_x^2\right)^2 - \right.$$

$$\left. -2\frac{1}{1+\frac{1}{n}\left(\beta_2(y)-1\right)} V_R \operatorname{Cov}\left(s_y^2, s_x^2\right)\right].$$

Así, sustituyendo los valores

$$\operatorname{ECM}\left(s_y^{2\,*}\right) = E\left(s_y^{2\,*} - S_y^2\right)^2 = \frac{\frac{1}{n}\left(\beta_2(y)-1\right)}{1+\frac{1}{n}\left(\beta_2(y)-1\right)} S_y^4,$$

$$E\left(s_x^2 - S_x^2\right)^2 = V\left(s_x^2\right) = \frac{1}{n} S_x^4\left(\beta_2(x)-1\right) \quad \text{y} \quad \operatorname{Cov}\left(s_y^2, s_x^2\right) = \frac{1}{n} S_x^2 S_y^2\left(\theta - 1\right),$$

operando y teniendo en cuenta que $\operatorname{ECM}(t_2) = S_x^4 \operatorname{ECM}\left(\widehat{V}_R\right)$, se llega a la expresión antes indicada.

De esta forma, comparando el estimador t_2 con el estimador de Isaki, $\widehat{S}_I^2$, se obtiene que si

$$\theta < \frac{\beta_2(y)+1}{2},$$

entonces $\operatorname{ECM}(t_2) < \operatorname{ECM}\left(\widehat{S}_I^2\right)$.

En efecto, la diferencia $\operatorname{ECM}\left(\widehat{S}_I^2\right) - \operatorname{ECM}(t_2)$ se expresa como

$$\operatorname{ECM}\left(\widehat{S}_I^2\right) - \operatorname{ECM}(t_2) = \frac{1}{n} S_y^4 \left[\beta_2(y) - 2\theta + 1 - \frac{\beta_2(y) - 1 - 2(\theta - 1)}{1+\frac{1}{n}\left(\beta_2(y)-1\right)}\right] =$$

$$= \left(\frac{1}{n}\right)^2 S_y^4 \frac{\beta_2(y)^2 - 2\theta\beta_2(y) + 2\theta - 1}{1+\frac{1}{n}\left(\beta_2(y)-1\right)}$$

y es positiva si se verifica la condición anterior.

Además, como el estimador de Isaki es más preciso que el estimador de expansión simple s_y^2 siempre que se verique

$$\theta > \frac{\beta_2(x)+1}{2},$$

se obtiene que si

$$\frac{\beta_2(x)+1}{2} < \theta < \frac{\beta_2(y)+1}{2},$$

entonces

$$\mathrm{ECM}\,(t_2) < \mathrm{ECM}\left(\widehat{S}_I^2\right) < V\left(s_y^2\right).$$

En el caso particular de normalidad antes expuesto, las condiciones anteriores por separado indican:

$$\text{Si} \quad \rho^2 < \frac{1}{2} \qquad \text{entonces} \qquad \mathrm{ECM}\,(t_2) < \mathrm{ECM}\left(\widehat{S}_I^2\right), \quad \text{y}$$

$$\text{si} \quad \rho^2 > \frac{1}{2} \qquad \text{entonces} \qquad \mathrm{ECM}\left(\widehat{S}_I^2\right) < V\left(s_y^2\right).$$

Posteriormente, en vez de condiderar un estimador específico ($s_y^{2\,*}$) para la varianza poblacional, Prasad y Singh consideran una clase de estimadores tipo razón

$$t_A = A s_y^2 \frac{S_x^2}{s_x^2}$$

para S_y^2.

Procediendo como en el caso de t_2, tomando $\widehat{V}_R^* = \dfrac{A s_y^2}{s_x^2}$ como un estimador del cociente de varianzas $\dfrac{S_y^2}{S_x^2}$, obtienen el error cuadrático medio aproximado del estimador

$$\mathrm{ECM}\,(t_A) \simeq S_y^4 \left[A^2\left(1+\frac{1}{n}\left(\beta_2(y)-1\right)\right) -\right.$$

$$\left. -2A\left(1+\frac{1}{n}\left(\theta-1\right)\right)+\frac{1}{n}\left(\beta_2(x)-1\right)+1\right],$$

y las condiciones bajo las cuales este estimador mejora al estimador de Isaki, que son

$$1 < A < \frac{1 + \frac{1}{n}(2\theta - \beta_2(y) - 1)}{1 + \frac{1}{n}(\beta_2(y) - 1)},$$

o bien,

$$\frac{1 + \frac{1}{n}(2\theta - \beta_2(y) - 1)}{1 + \frac{1}{n}(\beta_2(y) - 1)} < A < 1.$$

Para obtener el estimador óptimo de la clase, derivando respecto de A en la expresión del error cuadrático medio $\mathrm{ECM}(t_A)$, se obtiene el mínimo para

$$A = \frac{1 + \frac{1}{n}(\theta - 1)}{1 + \frac{1}{n}(\beta_2(y) - 1)},$$

para el cual el estimador óptimo

$$t_3 = \frac{1 + \frac{1}{n}(\theta - 1)}{1 + \frac{1}{n}(\beta_2(y) - 1)} s_y^2 \frac{S_x^2}{s_x^2}$$

tiene por error cuadrático medio aproximado

$$\mathrm{ECM}(t_3) \simeq \frac{1}{n} S_y^4 \left[\beta_2(x) - 1 + \frac{\beta_2(y) - (\theta - 1)\left(\frac{1}{n}(\theta - 1) + 2\right) - 1}{1 + \frac{1}{n}(\beta_2(y) - 1)} \right].$$

La relación

$$\mathrm{ECM}\left(\widehat{S}_I^2\right) = \mathrm{ECM}(t_3) + \frac{(\beta_2(y) - \theta)^2}{1 + \frac{1}{n}(\beta_2(y) - 1)} \left(\frac{1}{n}\right)^2 S_y^4,$$

que se verifica siempre, indica que $\mathrm{ECM}(t_3) < \mathrm{ECM}\left(\widehat{S}_I^2\right)$ para cualesquiera valores de $\beta_2(y)$, $\beta_2(x)$ y θ, e indica además la ganancia en precisión con el estimador t_3 frente a $\widehat{S}_I^2$.

Con respecto al estimador t_2, la relación

$$\mathrm{ECM}(t_2) = \mathrm{ECM}(t_3) + \frac{(\theta-1)^2}{1+\frac{1}{n}(\beta_2(y)-1)}\left(\frac{1}{n}\right)^2 S_y^4,$$

que también se verifica siempre, indica $\mathrm{ECM}(t_3) < \mathrm{ECM}(t_2)$ y la ganancia en precisión conseguida.

Así, t_3 es siempre más preciso que t_2, mejor que $\widehat{S}_I^2$ si

$$0 < \theta < \frac{\beta_2(y)+1}{2}$$

y $\widehat{S}_I^2$ es mejor que s_y^2 si

$$\theta > \frac{\beta_2(x)+1}{2}.$$

Por tanto, si

$$\frac{\beta_2(x)+1}{2} < \theta < \frac{\beta_2(y)+1}{2}$$

t_3 es un estimador de S_y^2 más eficiente que s_y^2, $\widehat{S}_I^2$ y t_2, pues en este caso

$$\mathrm{ECM}(t_3) < \mathrm{ECM}(t_2) < \mathrm{ECM}\left(\widehat{S}_I^2\right) < V\left(s_y^2\right).$$

Finalmente, Prasad y Singh, en el mismo trabajo, definen otra clase de estimadores para la varianza poblacional de la forma

$$t_\delta = \delta \frac{s_y^2}{s_x^2} S_x^2$$

y determinan δ minimizando $\mathrm{ECM}(t_\delta)$. Indicar que la diferencia entre t_3, el estimador óptimo para t_A y el estimador óptimo para t_δ, $\widehat{S}_{PS}^2$ radica en que antes de obtener la aproximación de los respectivos errores cuadráticos medios para despues minimizar se han hecho dos aproximaciones en el caso de t_A y una en el caso de t_δ. Sus propiedades más importantes las estudiamos a continuación.

Propiedades.

Consistencia.

Es clara la consistencia del estimador, en el sentido de consistencia en poblaciones finitas.

Sesgo.

El sesgo exacto del estimador t_δ tiene por expresión

$$\text{sesgo}\,(t_\delta) = -\delta\,\text{Cov}\left(\frac{s_y^2}{s_x^2}, s_x^2\right) + (\delta - 1)S_y^2.$$

En efecto, operando y teniendo en cuenta (4.3.2), se tiene

$$\text{sesgo}\left(\widehat{S}_{PS}^2\right) = E\left(\widehat{S}_{PS}^2 - S_y^2\right) = E\left(\delta s_y^2 \frac{S_x^2}{s_x^2} - S_y^2\right) =$$

$$= E\left(\delta\widehat{S}_I^2 - \delta S_y^2 + \delta S_y^2 - S_y^2\right) = \delta\,\text{sesgo}\left(\widehat{S}_I^2\right) + (\delta - 1)S_y^2 =$$

$$= -\delta\,\text{Cov}\left(\frac{s_y^2}{s_x^2}, s_x^2\right) + (\delta - 1)S_y^2.$$

Precisión.

Al ser el estimador sesgado, su precisión se medirá pues por su error cuadrático medio:

$$\text{ECM}\,(t_\delta) = E\left(t_\delta - S_y^2\right)^2 = E\left(\delta\left(\widehat{S}_I^2 - S_y^2\right) + (\delta - 1)S_y^2\right) =$$

$$= E\left(\delta^2\left(\widehat{S}_I^2 - S_y^2\right)^2 + (\delta - 1)^2 S_y^4 + 2\delta(\delta - 1)\left(\widehat{S}_I^2 - S_y^2\right)S_y^2\right) =$$

$$= \delta^2\,\text{ECM}\left(\widehat{S}_I^2\right) + (\delta - 1)^2 S_y^4 + 2\delta(\delta - 1)S_y^2\,\text{sesgo}\left(\widehat{S}_I^2\right).$$

Sustituyendo los valores obtenidos en (4.3.6) y (4.3.5) como aproximaciones del sesgo y del error cuadrático medio del estimador de Isaki (y por tanto sólo se realiza una aproximación antes de minimizar el error cuadrático medio, y

no dos como en el caso de t_A) obtenemos la aproximación del error cuadrático medio del estimador t_δ bajo diseño de muestreo aleatorio con reemplazamiento:

$$\mathrm{ECM}\,(t_\delta) \simeq$$

$$\simeq S_y^4\left[\delta^2\left(1+\frac{1}{n}\left(\beta_2(y)+3\beta_2(x)-4\theta\right)\right)-2\delta\left(1+\frac{1}{n}\left(\beta_2(x)-\theta\right)\right)+1\right].$$

Minimizando ahora en δ este error cuadrático medio, tenemos

$$\frac{\partial\,\mathrm{ECM}\,(t_\delta)}{\partial\delta}=S_y^4 2\delta\left(1+\frac{1}{n}\left(\beta_2(y)+3\beta_2(x)-4\theta\right)-2\left(1+\frac{1}{n}\left(\beta_2(x)-\theta\right)\right)\right],$$

igualando a cero se obtiene un mínimo para

$$\delta_1=\frac{1+\frac{1}{n}\left(\beta_2(x)-\theta\right)}{1+\frac{1}{n}\left(\beta_2(y)+3\beta_2(x)-4\theta\right)}.$$

Considerando ahora el estimador óptimo, con este valor de δ,

$$\widehat{S}_{PS}^2=\frac{1+\frac{1}{n}\left(\beta_2(x)-\theta\right)}{1+\frac{1}{n}\left(\beta_2(y)+3\beta_2(x)-4\theta\right)}s_y^2\frac{S_x^2}{s_x^2}$$

una aproximación de su error cuadrático medio es

$$\mathrm{ECM}\left(\widehat{S}_{PS}^2\right)\simeq\frac{1}{n}S_y^4\frac{\beta_2(y)+\beta_2(x)-2\theta\frac{1}{n}\left(\beta_2(x)-\theta\right)^2}{1+\frac{1}{n}\left(\beta_2(y)+3\beta_2(x)-4\theta\right)}.$$

Comparación con el estimador de Isaki.

Si

$$\beta_2(x)-1>\frac{4(\theta-1)}{3}$$

entonces $\mathrm{ECM}\left(\widehat{S}_{PS}^2\right)<\mathrm{ECM}\,(t_3)$.

De esta forma, si

$$\theta < \frac{\beta_2(y)+1}{2} \quad \text{y} \quad \theta < \frac{3\beta_2(x)+1}{4}$$

se tiene la relación

$$\text{ECM}\left(\widehat{S}^2_{PS}\right) < \text{ECM}\left(t_3\right) < \text{ECM}\left(t_2\right) < \text{ECM}\left(\widehat{S}^2_{I}\right).$$

Si además

$$\theta > \frac{\beta_2(x)+1}{2}$$

entonces

$$\text{ECM}\left(\widehat{S}^2_{PS}\right) < \text{ECM}\left(t_3\right) < \text{ECM}\left(t_2\right) < \text{ECM}\left(\widehat{S}^2_{I}\right) < \text{ECM}\left(s^2_y\right).$$

Caso particular de normalidad.

Si los momentos de la distribución (y, x) son los mismos de una normal bivariante hasta el orden cuatro, los errores cuadráticos medios de los estimadores t_3 y $\widehat{S}^2_{PS}$ tienen por expresiones aproximadas

$$\text{ECM}\left(t_3\right) \simeq \frac{4}{n} S^4_y \left(1-\rho^2\right)\left(1-\frac{(1-\rho^2)}{n+2}\right)$$

$$\text{ECM}\left(\widehat{S}^2_{PS}\right) \simeq \frac{4}{n} S^4_y \left(1-\rho^2\right).$$

Comparándolas se puede deducir que el estimador $\widehat{S}^2_{PS}$ es más preciso que t_3 siempre que

$$\rho^2 < \frac{3}{4}.$$

Por tanto, para $\rho^2 < \frac{1}{2}$, el estimador t_4 mejora al estimador de Isaki.

En otro trabajo, *Prasad* y *Singh* (1992) proponen un estimador insesgado alternativo al de *Isaki*. El razonamiento que les lleva a construir este estimador es que en determinadas situaciones prácticas el sesgo del estimador de Isaki puede ser importante, además de que en general este sesgo exacto no puede ser computado en la práctica, teniendo que recurrir a aproximaciones. Para

solventar estos problemas, estos autores sugieren un estimador insesgado de S_y^2, calculan su varianza exacta y estudian sus propiedades de precisión comparándolas con el estimador de Isaki y con el estimador insesgado usual, s_y^2. El estimador es el siguiente:

$$d_a = s_y^2 - a\frac{s_x^2}{S_x^2} + a. \tag{4.3.7}$$

Obviamente $E(d_a) = S_y^2$. Su varianza exacta viene dada por la expresión:

$$V(d_a) = V\left(s_y^2\right) + a^2\frac{V(s_x^2)}{S_x^4} - 2a\frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)}{S_x^2},$$

a la que se llega sin más que escribir

$$V(d_a) = E\left(d_a - S_y^2\right)^2 = E\left(s_y^2 - S_y^2 - a\frac{(s_x^2 - S_x^2)}{S_x^2}\right)^2 =$$

$$= E\left(s_y^2 - S_y^2\right)^2 + a^2\frac{E(s_x^2 - S_x^2)^2}{S_x^4} - 2a\frac{E\left(s_y^2 - S_y^2\right)(s_x^2 - S_x^2)}{S_x^2}.$$

De la expresión de la varianza se sigue que el estimador d_a es más preciso que el estimador de expansión simple s_y^2 si

$$0 < a < \frac{2S_x^2\,\mathrm{Cov}\left(s_y^2, s_x^2\right)}{V(s_x^2)}$$

o bien

$$\frac{2S_x^2\,\mathrm{Cov}\left(s_y^2, s_x^2\right)}{V(s_x^2)} < a < 0,$$

puesto que la diferencia entre varianzas se expresa

$$V(d_a) - V\left(s_y^2\right) = a^2\frac{V(s_x^2)}{S_x^4} - 2a\frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)}{S_x^2}.$$

Así, para poblaciones en las que se cumplan la igualdad entre los momentos antes citada, las condiciones anteriores se traducen, en un primer grado de aproximación, en que el estimador d_a mejora en precisión al estimador simple s_y^2 si

$$0 < a < 2\rho^2 S_x^2$$

o bien

$$2\rho^2 S_x^2 < a < 0,$$

pues en tales poblaciones

$$V\left(s_x^2\right) \simeq \frac{1}{n} S_x^4 \left(\beta_2(x) - 1\right) \simeq \frac{2}{n} S_x^4$$

y

$$\mathrm{Cov}\left(s_y^2, s_x^2\right) \simeq \frac{1}{n} S_x^2 S_y^2 (\theta - 1) \simeq \frac{2}{n} S_x^2 S_y^2 \rho^2.$$

Con respecto al estimador de Isaki, comparando sus precisiones se obtiene que el estimador d_a mejora al estimador de Isaki siempre que

$$S_y^2 < a < 2\frac{S_x^2 \,\mathrm{Cov}\left(s_y^2, s_x^2\right)}{V\left(s_x^2\right)} - S_y^2$$

o bien

$$2\frac{S_x^2 \,\mathrm{Cov}\left(s_y^2, s_x^2\right)}{V\left(s_x^2\right)} - S_y^2 < a < S_y^2.$$

Notar que en este caso, llamando

$$\alpha_1 = S_y^2; \quad \alpha_2 = 2\frac{S_x^2 \,\mathrm{Cov}\left(s_y^2, s_x^2\right)}{V\left(s_x^2\right)} - S_y^2,$$

se tiene la igualdad

$$V\left(d_a\right) - V\left(s_y^2\right) = \frac{V\left(s_x^2\right)}{S_x^4}\left[a^2 - a\left(\alpha_1 + \alpha_2\right) + \alpha_1\alpha_2\right].$$

Ahora bien, sustituyendo los valores que en muestreo aleatorio simple tienen las expresiones

$$V(s_x^2) = S_x^4 \frac{N-n}{(N-2)n}\left(K\beta_2(x) - M\right),$$

$$\mathrm{Cov}(s_y^2, s_x^2) = S_y^2 S_x^2 \frac{N-n}{(N-2)n} \left(K\theta + K_1\theta_1 - K_2\right),$$

donde K está dada en (4.3.3), y θ_1 K_1, K_2 y M están dadas en (4.3.4), de las condiciones anteriores resulta que el estimador t_a mejora al estimador de Isaki siempre que

$$S_y^2 < a < (2b-1)S_y^2$$

o bien

$$(2b-1)S_y^2 < a < S_y^2,$$

donde

$$b = \frac{K\theta + K_1\theta_1 - K_2}{K\beta_2(x) - M}.$$

Así, para poblaciones normales (en el el contexto anterior), las condiciones anteriores se traducen, en un primer grado de aproximación, en que el estimador d_a mejora en precisión al estimador simple s_y^2 si

$$S_y^2 < a < \left(2\rho^2 - 1\right) S_y^2$$

o bien

$$\left(2\rho^2 - 1\right) S_y^2 < a < S_y^2.$$

En su trabajo *Prasad* y *Singh* no obtienen el óptimo de la clase de estimadores que presentan, d_a. El estudio de este óptimo lo tratamos en el siguiente apartado, dando una respuesta a porqué no determinaron el estimador óptimo de la clase.

Comparación con el estimador de regresión.

Hay que hacer notar que si se determina el estimador óptimo de la clase de estimadores propuestos por Prasad y Singh en este trabajo, d_a, con el criterio de minimizar su varianza, se obtiene que el estimador óptimo es precisamente el estimador de regresión.

Proposición 4.3.1 *El estimador óptimo de la clase de estimadores dada por Prasad y Singh es el estimador de regresión* $\widehat{S}^2_{reg_{b_0}}$ *dado en (4.2.2):*

$$\widehat{S}^2_{reg_{b_0}} = s_y^2 + \frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)}{V\left(s_x^2\right)}\left(S_x^2 - s_x^2\right)$$

Demostración.-
En efecto, derivando en la expresión

$$V\left(d_a\right) = V\left(s_y^2\right) + a^2\frac{V\left(s_x^2\right)}{S_x^4} - 2a\frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)}{S_x^2},$$

respecto de a e igualando a cero, se tiene

$$\frac{\partial V\left(d_a\right)}{\partial a} = 2a\frac{V\left(s_x^2\right)}{S_x^4} - 2\frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)}{S_x^2} = 0$$

que proporciona el mínimo para

$$a_0 = \frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)S_x^2}{V\left(s_x^2\right)},$$

y para este mínimo el estimador óptimo es

$$d_{a_0} = s_y^2 - a_0\frac{s_x^2}{S_x^2} + a_0 = s_y^2 - \frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)S_x^2}{V\left(s_x^2\right)}\frac{s_x^2}{S_x^2} + \frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)S_x^2}{V\left(s_x^2\right)} =$$

$$= s_y^2 + \frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)}{V\left(s_x^2\right)}\left(S_x^2 - s_x^2\right) = \widehat{S}^2_{reg_{b_0}},$$

donde $\widehat{S}^2_{reg_{b_0}}$ denota el estimador de regresión óptimo.

Proposición 4.3.2 *Para cada valor de a que define* d_a *es posible encontrar dos valores*

$$b_1 = \frac{a}{S_x^2} \;\; y \;\; b_2 = \frac{2\,\mathrm{Cov}\left(s_y^2, s_x^2\right)}{S_x^2} - \frac{a}{S_x^2},$$

para los que

$$V(d_a) = V\left(\widehat{S}^2_{reg_{b_1}}\right) = V\left(\widehat{S}^2_{reg_{b_2}}\right).$$

Además se tiene

$$d_a = \widehat{S}^2_{reg_{b_1}};\ y \quad d_a = 2\widehat{S}^2_{reg_{b_0}} - \widehat{S}^2_{reg_{b_1}}.$$

Demostración.-
Expresando

$$V(d_a) - V\left(\widehat{S}^2_{reg}\right) =$$

$$= b^2 V\left(s_x^2\right) - 2b\,\mathrm{Cov}\left(s_y^2, s_x^2\right) - a^2\frac{V\left(s_x^2\right)}{S_x^4} + 2a\frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)}{S_x^2} = 0,$$

se tiene el discriminante de la ecuación de segundo grado en b:

$$\Delta = 4\left(\mathrm{Cov}^2\left(s_y^2, s_x^2\right) + a^2\frac{V\left(s_x^2\right)^2}{S_x^4} - 2a\frac{V\left(s_x^2\right)\mathrm{Cov}\left(s_y^2, s_x^2\right)}{S_x^2}\right) =$$

$$= \left(a\frac{V\left(s_x^2\right)}{S_x^2} - \mathrm{Cov}\left(s_y^2, s_x^2\right)\right)^2,$$

y por tanto los valores

$$b_1 = \frac{a}{S_x^2}\ \text{y} \quad b_2 = \frac{2\,\mathrm{Cov}\left(s_y^2, s_x^2\right)}{S_x^2} - \frac{a}{S_x^2}$$

satisfacen el enunciado puesto que

$$b = \frac{\mathrm{Cov}\left(s_y^2, s_x^2\right) \pm\sqrt{\Delta}}{V\left(s_x^2\right)}.$$

De una forma más sencilla:

$$d_a = s_y^2 - a\frac{s_x^2}{S_x^2} + a = s_y^2 + \frac{a}{S_x^2}\left(S_x^2 - s_x^2\right) = \widehat{S}^2_{reg_{b_1}}.$$

Además, notar que el estimador que inicialmente consideran en su trabajo

$$d_1 = s_y^2 - \frac{s_x^2}{S_x^2} + 1,$$

para $a = 1$ no es más que el estimador de regresión $\widehat{S}_{reg}^2$ con $b = \frac{1}{S_x^2}$.

En este sentido afirmamos que la clase de estimadores d_a de la varianza poblacional S_y^2 propuesta por Prasad y Singh puede verse como un caso especial del estimador de regresión.

En este mismo trabajo Prasad y Singh concluyen con un estudio empírico considerando unos datos de Cochran (1977, Cap 6) y que nosotros retomaremos para terminar el estudio teórico del estimador construido con el método repetido de sustitución o método de exponenciación, que presentaremos después del estudio de la técnica de estimación de producto que desarrollamos a continuación.

§4.4 Estimación de producto.

4.4.1 Introducción.

Es conocido que el estimador producto de la media se utiliza si la correlación entre las variables es negativa y que mejora al estimador simple siempre y cuando

$$\rho(y, x) < \frac{-C_x}{2C_y},$$

donde C_x y C_y son los coeficientes de variación de las variables x e y, respectivamente.

Isaki para significar el estimador de razón argumentaba que la correlación alta y positiva entre y_i y x_i (las razones $\frac{y_i}{x_i}$ $i = 1, \ldots, N$ son aproximadamente constantes) podría llevar a una correlación también entre $\left(y_i - y_j\right)^2$ y $\left(x_i - x_j\right)^2$. De esta forma, al poder escribir el parámetro varianza poblacional en la forma

$$S_y^2 = \frac{1}{N(N-1)} \sum_{i \neq j} \left(y_i - y_j\right)^2,$$

tendría sentido intentar mejorar la precisión del estimador de expansión simple s_y^2 por medio de un estimador tipo razón.

¿Será esto también cierto si la correlación es negativa? o equivalentemente, si los productos $y_i x_i$ $i = 1, \ldots, N$ son aproximadamente constantes, ¿podrían serlo también los pares $\left(y_i - y_j\right)^2$ y $\left(x_i - x_j\right)^2$?. En caso afirmativo será razonable definir un estimador producto de la varianza que mejore, bajo ciertas condiciones, al estimador simple.

Consideremos un muestreo aleatorio simple con tamaño muestral lo suficientemente grande para prescindir del factor de corrección por finitud.

4.4.2 Definición del estimador.

De esta forma se define

$$\widehat{S}_p^2 = s_y^2 \frac{s_x^2}{S_x^2}.$$

Las propiedades más importantes de este estimador las estudiamos a continuación.

4.4.3 Propiedades.

Proposición 4.4.1 *El estimador $\widehat{S}_p^2$ es consistente.*

Demostración.- Es evidente, en el sentido de consistencia en poblaciones finitas.

Proposición 4.4.2 *El estimador $\widehat{S}_p^2$ es sesgado y su sesgo exacto viene dada por la expresión:*

$$\text{sesgo}\left(\widehat{S}_p^2\right) = S_y^2 \left[\frac{1}{n}(\theta - 1)\right],$$

que como vemos tiende a cero cuando n aumenta. Además, para $\theta = 1$ el sesgo es cero.

Demostración.-
Consideremos las variables

$$e_0 = \frac{s_y^2 - S_y^2}{S_y^2} \quad ; \quad e_1 = \frac{s_x^2 - S_x^2}{S_x^2}.$$

Mediante ellas expresamos el estimador $\widehat{S}_p^2$ de la forma

$$\widehat{S}_\alpha^2 = S_y^2 (1 + e_0)(1 + e_1) = S_y^2 (1 + e_0 + e_1 + e_0 e_1)$$

Así,

$$\widehat{S}_p^2 - S_y^2 \simeq S_y^2 (e_0 + e_1 + e_0 e_1). \tag{4.4.1}$$

Tomando esperanzas se tiene

$$\text{sesgo}\left(\widehat{S}_p^2\right) = S_y^2 E(e_0 + e_1 + e_0 e_1) = S_y^2 E(e_0 e_1) =$$

$$= S_y^2 \frac{\text{Cov}\left(s_y^2, s_x^2\right)}{S_x^2}.$$

Sustituyendo los valores dados en (4.3.4) (prescindiendo del factor de corrección por finitud) obtendríamos las aproximaciones:

$$\frac{V(s_x^2)}{S_x^4} = \frac{1}{n}(\beta_2(x) - 1),$$

$$\frac{\text{Cov}(s_y^2, s_x^2)}{S_y^2 S_x^2} = \frac{1}{n}(\theta - 1), \tag{4.4.2}$$

a partir de la cuales se llega a la aproximación expuesta en el enunciado de la proposición.

Su precisión la medirá por tanto su error cuadrático medio, siendo una expresión aproximada de este error la que proporciona la siguiente proposición:

Proposición 4.4.3 *Una aproximación del error cuadrático medio del estimador $\widehat{S}_p^2$ viene dada por la expresión:*

$$\text{ECM}\left(\widehat{S}_p^2\right) \simeq \frac{1}{n} S_y^4 [\beta_2(y) + \beta_2(x) + 2\theta - 4]. \tag{4.4.3}$$

Demostración.-
Partiendo de (4.4.1) obtenemos

$$\text{ECM}\left(\widehat{S}_p^2\right) = E\left(\widehat{S}_p^2 - S_y^2\right)^2 \simeq$$

$$\simeq S_y^4 E\left(e_0^2 + e_1^2 + 2e_0e_1\right) = S_y^4\left(E\left(e_0^2\right) + E\left(e_1^2\right) + 2E\left(e_0e_1\right)\right).$$

Procediendo como en el cálculo del sesgo, tenemos

$$\text{ECM}\left(\widehat{S}_p^2\right) \simeq S_y^4\left[\frac{V\left(s_y^2\right)}{S_y^4} + \frac{V\left(s_x^2\right)}{S_x^4} + \frac{2\,\text{Cov}\left(s_x^2, s_y^2\right)}{S_x^2 S_y^2}\right]. \tag{4.4.4}$$

Sustituyendo el valor dado en (4.4.2) llegamos a la expresión de la proposición.

4.4.4 Caso particular de normalidad.

Si los momentos de la distribución (y, x) son los mismos de una normal bivariante hasta el orden cuatro

$$\theta = 1 + 2\rho^2 \quad ; \quad \beta_2(x) = 3$$

y así si las variables son incorreladas el estimador es insesgado y conforme mayor sea la correlación (en valor absoluto) mayor será el sesgo, puesto que

$$\text{sesgo}\left(\widehat{S}_p^2\right) = S_y^2 \frac{2}{n}\rho^2.$$

Una aproximación del error cuadrático medio del estimador, en este caso "normal", $\widehat{S}_p^2$ viene dada por la expresión:

$$\text{ECM}\left(\widehat{S}_p^2\right) \simeq \frac{4}{n} S_y^4\left[1 + \rho^2\right].$$

4.4.5 Comparación con el estimador de expansión simple.

El problema es pues determinar si bajo ciertas condiciones el error cuadrático medio del estimador producto es más pequeño que la varianza $V\left(s_y^2\right)$. Comparando la precisión de ambos estimadores se obtiene el siguiente resultado:

Proposición 4.4.4 *El estimador producto $\widehat{S}_p^2$ es más preciso que el estimador simple s_y^2 si y sólo si*

$$\theta \leq \frac{3-\beta_2(x)}{2}$$

Demostración.-

$$\mathrm{ECM}\left(\widehat{S}_p^2\right) - V\left(s_y^2\right) \simeq \frac{1}{n}S_y^4\left[\beta_2(y)+\beta_2(x)+2\theta-4\right] - \frac{1}{n}S_y^4\left[\beta_2(y)-1\right] =$$

$$= \frac{1}{n}S_y^4\left[\beta_2(x)+2\theta-3\right] \leq 0; \quad \theta \leq \frac{3-\beta_2(x)}{2}.$$

Así pues tendrá sentido definir el estimador producto siempre y cuando haya poblaciones para las cuales sea cierta la condición anterior. De esta forma queda descartado el uso del estimador producto para poblaciones en las que los momentos sean los mismos hasta el orden cuatro de una normal bivariante pues en ellas $\beta_2(x) = 3$ y la condición $\theta < 0$ no puede darse. Además, si (y, x) es "normal bivariante" (en el sentido anterior),

$$\mathrm{ECM}\left(\widehat{S}_p^2\right) - V\left(s_y^2\right) = \frac{2}{n}S_y^4\left[1+2\rho^2\right] > 0.$$

De igual forma, si la variable x verifica $\beta_2(x) > 3$ (leptocúrtica) su utilización como variable auxiliar en la estimación producto de la varianza no mejorará la precisión de la estimación directa obtenida a partir de s_y^2.

Sin embargo, la desigualdad de la proposición anterior puede ser cierta, como muestra el siguiente ejemplo:

x	4.61	4.99	4.40	4.49	5.29	4.96	4.62	4.44	5.09	4.79
y	0.32	4.06	4	1.53	1.97	1.8	0.66	0.85	2.16	10.39

para el que se obtienen los valores:

$$\beta_2(x) = 1.82350; \quad \theta = 0.20129.$$

Por tanto puede haber poblaciones para las cuales la estimación de la varianza mediante la técnica de estimación de producto puede mejorar en precisión a la estimación simple. Además el estimador producto se puede calcular siempre, en contraposición con el estimador de regresión óptimo.

En la última sección comparamos este estimador con el estimador de razón de Isaki y comprobamos que la condición $\theta < 1$ produce una mejora en precisión respecto al estimador de razón.

§4.5 Método repetido de sustitución.

4.5.1 Introducción.

En este apartado se propone un nuevo estimador de la varianza poblacional, en una población finita, que utiliza información auxiliar. Demostramos que este estimador mejora al estimador simple y al estimador de razón de la varianza dado por *Isaki* (1983) y al estimador de producto antes expuesto.

Supongamos que el problema sigue siendo estimar la varianza poblacional. Hemos estudiado ya los estimadores dados por *Prasad* y *Singh* (1990, 1992) y el estimador de *Isaki* (1983). En este apartado proponemos un nuevo estimador de razón insesgado para la varianza, obtenido siguiendo la idea del método de sustitución repetida dado por *Srivastava* (1967) para la media poblacional. Este nuevo estimador va a tener la propiedad de ser siempre más preciso que el estimador de expansión simple, que el estimador de *Isaki*, que el estimador producto y asintóticamente igual de preciso que el estimador de regresión para estimar la varianza de y. Además va a ser sencillo de computar, generaliza al estimador de Isaki, al estimador producto, es cuasi–insesgado y sólo precisa del conocimiento de la correlación entre las variables, en el caso "normal bivariante", para ser computado.

4.5.2 Definición del estimador $\widehat{S}^2_\alpha$.

Definimos el nuevo estimador de razón para la varianza S^2_y de la forma

$$\widehat{S}^2_\alpha = s^2_y \left(\frac{S^2_x}{s^2_x} \right)^\alpha \quad \text{con } \alpha \text{ real.}$$

La justificación de la definición es la siguiente. Consideremos el estimador de razón

$$\widehat{S}^2_I = s^2_y \left(\frac{S^2_x}{s^2_x} \right) = \widehat{S}^2_1.$$

Si este estimador es mejor que s_y^2 (si $\rho(s_y^2, s_x^2) > \frac{1}{2}\frac{CV(s_x^2)}{CV(s_y^2)}$) podemos utilizar $\widehat{S}_I^2$ en vez de s_y^2 en la estimación de razón, por lo que obtendremos el estimador

$$\widehat{S}_2^2 = \widehat{S}_1^2 \frac{S_x^2}{s_x^2} = s_y^2 \frac{S_x^2}{s_x^2}\frac{S_x^2}{s_x^2} = s_y^2 \left(\frac{S_x^2}{s_x^2}\right)^2 .$$

Si este estimador es a su vez mejor que s_y^2 podemos repetir el proceso obteniendo, en la siguiente iteración

$$\widehat{S}_3^2 = s_y^2 \left(\frac{S_x^2}{s_x^2}\right)^3 .$$

Repitiendo el proceso α veces, donde α es un entero, llegamos al estimador

$$\widehat{S}_\alpha^2 = s_y^2 \left(\frac{S_x^2}{s_x^2}\right)^\alpha .$$

Así, aunque en esta justificación α debe ser entero, nosotros consideraremos el estudio del estimador $\widehat{S}_\alpha^2$ con α cualquier número real.

4.5.3 Propiedades.

Proposición 4.5.1 *El estimador $\widehat{S}_\alpha^2$ es consistente.*

Demostración.- Es evidente, en el sentido de consistencia en poblaciones finitas.

Proposición 4.5.2 *El estimador $\widehat{S}_\alpha^2$ es sesgado y una aproximación de su sesgo viene dada por la expresión:*

$$\text{sesgo}\left(\widehat{S}_\alpha^2\right) \simeq \frac{S_y^2}{n}\left(\alpha^2\left(\beta_2(x) - 1\right) - \alpha\left(\theta - 1\right)\right).$$

Demostración.-
Consideremos las variables

$$e_0 = \frac{s_y^2 - S_y^2}{S_y^2} \quad ; \quad e_1 = \frac{s_x^2 - S_x^2}{S_x^2}.$$

Mediante ellas expresamos el estimador $\widehat{S}_\alpha^2$ de la forma

$$\widehat{S}_\alpha^2 = S_y^2 (1+e_0)(1+e_1)^{-\alpha} =$$

(siempre que $\mid e_1 \mid < 1$ y $\mid \alpha e_1 \mid < 1$)

$$= S_y^2 (1+e_0)\left(1 - \alpha e_1 + \alpha^2 e_1^2 - \cdots\right) \simeq S_y^2 (1+e_0)\left(1 - \alpha e_1 + \alpha^2 e_1^2\right) \simeq$$

$$\simeq S_y^2 \left(1 + e_0 - \alpha e_1 + \alpha^2 e_1^2 - \alpha e_0 e_1 + \alpha^2 e_0 e_1^2\right).$$

Así,

$$\widehat{S}_\alpha^2 - S_y^2 \simeq S_y^2 \left(e_0 - \alpha e_1 + \alpha^2 e_1^2 - \alpha e_0 e_1 + \alpha^2 e_0 e_1^2\right). \qquad (4.5.1)$$

Tomando esperanzas y reteniendo sólo los términos de grado dos en e_0 y e_1, se tiene

$$\text{sesgo}\left(\widehat{S}_\alpha^2\right) \simeq S_y^2 E\left(e_0 - \alpha e_1 + \alpha^2 e_1^2 - \alpha e_0 e_1\right) = S_y^2 \left(\alpha^2 E\left(e_1^2\right) - \alpha E\left(e_0 e_1\right)\right) =$$

$$= S_y^2 \left(\alpha^2 \frac{V\left(s_x^2\right)}{S_x^4} - \alpha \frac{\text{Cov}\left(s_y^2, s_x^2\right)}{S_y^2 S_x^2}\right).$$

Sustituyendo los valores dados en (4.4.2) en la aproximación del sesgo nos lleva a la expresión de la proposición.

Su precisión la medirá por tanto su error cuadrático medio, siendo una expresión aproximada de este error la que proporciona la siguiente proposición:

Proposición 4.5.3 *Una aproximación del error cuadrático medio del estimador* $\widehat{S}_\alpha^2$ *viene dada por la expresión:*

$$\text{ECM}\left(\widehat{S}_\alpha^2\right) \simeq \frac{S_y^4}{n}\left[\beta_2(y) - 1 + \alpha^2\left(\beta_2(x) - 1\right) - 2\alpha\left(\theta - 1\right)\right]. \qquad (4.5.2)$$

Demostración.-

Partiendo de (4.5.1) obtenemos

$$\text{ECM}\left(\widehat{S}_\alpha^2\right) = E\left(\widehat{S}_\alpha^2 - S_y^2\right)^2 \simeq$$

$$\simeq S_y^4 E\left(e_0^2+\alpha^2 e_1^2-2\alpha e_0 e_1\right)=S_y^4\left(E\left(e_0^2\right)+\alpha^2 E\left(e_1^2\right)-2\alpha E\left(e_0 e_1\right)\right).$$

Procediendo como en el cálculo aproximado del sesgo, tenemos

$$\mathrm{ECM}\left(\widehat{S}_\alpha^2\right)\simeq S_y^4\left[\frac{V\left(s_y^2\right)}{S_y^4}+\alpha^2\frac{V\left(s_x^2\right)}{S_x^4}-\frac{2\alpha\,\mathrm{Cov}\left(s_x^2,s_y^2\right)}{S_x^2 S_y^2}\right]. \tag{4.5.3}$$

Finalmente, sustituyendo los valores dados en (4.4.2) se llega a la aproximación expuesta en el enunciado de la proposición.

4.5.4 El estimador óptimo.

El problema ahora sería seleccionar de entre toda la clase de estimadores $\widehat{S}_\alpha^2$ aquél que sea óptimo en el sentido de mayor precisión, es decir, determinar el valor α que minimice $\mathrm{ECM}\left(\widehat{S}_\alpha^2\right)$.

Proposición 4.5.4 *El estimador óptimo que minimiza el error cuadrático medio del estimador $\widehat{S}_\alpha^2$ se obtiene para*

$$\alpha_1=\frac{\theta-1}{\beta_2(x)-1}.$$

Para este valor α_1, el estimador

$$\widehat{S}_{\alpha_1}^2=s_y^2\left(\frac{S_x^2}{s_x^2}\right)^{\frac{\theta-1}{\beta_2(x)-1}} \tag{4.5.4}$$

es cuasi-insesgado y su varianza aproximada viene dada por la expresión:

$$V\left(\widehat{S}_{\alpha_1}^2\right)\simeq V\left(s_y^2\right)\left(1-\rho^2\left(s_x^2,s_y^2\right)\right).$$

Demostración.-
Derivando en α la expresión (4.5.3) e igualando a cero

$$\frac{\partial\,\mathrm{ECM}\left(\widehat{S}_\alpha^2\right)}{\partial\alpha}=S_y^4\left(2\alpha\frac{V\left(s_x^2\right)}{S_x^4}-2\frac{\mathrm{Cov}\left(s_x^2,s_y^2\right)}{S_x^2 S_y^2}\right)=0,$$

se obtiene

$$\alpha_1 = \frac{\mathrm{Cov}\left(s_x^2, s_y^2\right)}{V\left(s_x^2\right)} \frac{S_x^2}{S_y^2} = \frac{\theta - 1}{\beta_2(x) - 1}.$$

Para este valor α_1, el estimador obtenido $\widehat{S}_{\alpha_1}^2$ es cuasi–insesgado, puesto que

$$\mathrm{sesgo}\left(\widehat{S}_{\alpha_1}^2\right) \simeq \frac{S_y^2}{n}\left(\alpha_1^2\left(\beta_2(x) - 1\right) - \alpha_1\left(\theta - 1\right)\right) =$$

$$= \frac{S_y^2}{n}\left(\frac{(\theta-1)^2}{(\beta_2(x)-1)^2}\left(\beta_2(x) - 1\right) - \frac{\theta - 1}{\beta_2(x) - 1}\left(\theta - 1\right)\right) = 0.$$

Además, para este valor de α la varianza aproximada del estimador es

$$V\left(\widehat{S}_{\alpha_1}^2\right) \simeq S_y^4\left[\frac{V\left(s_y^2\right)}{S_y^4} + \alpha_1^2\frac{V\left(s_x^2\right)}{S_x^4} - \frac{2\alpha_1\,\mathrm{Cov}\left(s_x^2, s_y^2\right)}{S_x^2 S_y^2}\right] =$$

$$= S_y^4\left[\frac{V\left(s_y^2\right)}{S_y^4} + \frac{\mathrm{Cov}^2\left(s_x^2, s_y^2\right)}{V\left(s_x^2\right)^2}\,\frac{S_x^4}{S_y^4}\frac{V\left(s_x^2\right)}{S_x^4} - \frac{2\frac{\mathrm{Cov}\left(s_x^2, s_y^2\right)}{V\left(s_x^2\right)}\,\frac{S_x^2}{S_y^2}\,\mathrm{Cov}\left(s_x^2, s_y^2\right)}{S_x^2 S_y^2}\right] =$$

$$= S_y^4\left[\frac{V\left(s_y^2\right)}{S_y^4} + \frac{\mathrm{Cov}^2\left(s_x^2, s_y^2\right)}{V\left(s_x^2\right)S_y^4} - \frac{2\,\mathrm{Cov}^2\left(s_x^2, s_y^2\right)}{S_y^4 V\left(s_x^2\right)}\right] =$$

$$= V\left(s_y^2\right)\left[1 - \frac{\mathrm{Cov}^2\left(s_x^2, s_y^2\right)}{V\left(s_x^2\right)V\left(s_y^2\right)}\right] = V\left(s_y^2\right)\left[1 - \rho^2\left(s_x^2, s_y^2\right)\right].$$

Por tanto este estimador es siempre más preciso que el estimador de expansión simple, s_y^2.

4.5.5 Caso particular de normalidad.

Proposición 4.5.5 *Si los momentos de la distribución (y, x) son los mismos de una normal bivariante hasta el orden cuatro, el valor de α que maximiza la precisión y hace cuasi–insesgado al estimador $\widehat{S}_\alpha^2$ es*

$$\alpha_1 = \rho^2,$$

donde ρ es el coeficiente de correlación entre las variables y y x. Para este valor de α_1, el estimador adopta la forma

$$\widehat{S}^2_{\rho^2} = s^2_y \left(\frac{S^2_x}{s^2_x}\right)^{\rho^2} \tag{4.5.5}$$

y su varianza aproximada viene dada por la expresión:

$$V\left(\widehat{S}^2_{\rho^2}\right) = \frac{2S^4_y}{n}\left(1-\rho^4\right).$$

Demostración.-
Teniendo en cuenta que en una población "normal bivariante"

$$\theta = 1 + 2\rho^2 \quad ; \quad \beta_2(x) = 3$$

se llega a que el valor óptimo de α es

$$\alpha_1 = \frac{\theta - 1}{\beta_2(x) - 1} = \rho^2, \tag{4.5.6}$$

y para este valor

$$V\left(\widehat{S}^2_{\rho^2}\right) = V\left(s^2_y\right)\left(1 - \rho^2\left(s^2_x, s^2_y\right)\right) =$$

$$= \frac{2}{n}S^4_y\left(1 - \frac{S^4_y S^4_x \left(\frac{2}{n}\rho^2\right)^2}{S^2_x \frac{2}{n} S^2_y \frac{2}{n}}\right) = \frac{1}{n}S^4_y\left(2 - \frac{(2\rho^2)^2}{2}\right) = \frac{2S^4_y}{n}\left(1-\rho^4\right),$$

dándose la igualdad en la precisión de este estimador y el de expansión simple en caso de independencia de las variables.

Comentario 1.- Se observa que sólo depende de ρ^2, y no de S^2_y.

Comentario 2. En la práctica es común disponer de información relativa a la correlación entre las variables. Utilizando esta información podemos determinar el estimador $\widehat{S}^2_{\rho^2}$ obteniendo así un aumento en precisión respecto al estimador simple (s^2_y) y al estimador de razón usual $(\widehat{S}^2_I)$. Además en este

caso "normal" el estimador producto, como caso particular del estimador $\widehat{S}^2_{\alpha}$, no mejora al estimador de expansión simple puesto que $\alpha_1 = \rho^2 > 0$.

Comentario 3.- Notar también que de (4.5.6) deducimos que para el caso de "normalidad" el valor óptimo está comprendido entre 0 y 1 y de acuerdo con la justificación del método, no se realizará más de una iteración.

4.5.6 Comparación con el estimador de regresión.

El estimador de regresión para el parámetro varianza se define

$$\widehat{S}^2_{reg} = s_y^2 + b\left(S_x^2 - s_x^2\right)$$

siendo b una constante fijada.

Este estimador es insesgado como vimos al estudiar sus propiedades y su varianza aproximada viene dada por la expresión

$$V\left(\widehat{S}^2_{reg}\right) \simeq \frac{1}{n}\left(S_y^4\left(\beta_2(y)-1\right) + b^2 S_x^4\left(\beta_2(x)-1\right) - 2bS_y^2 S_x^2(\theta - 1)\right)$$

El valor de b que porporciona el estimador más preciso es, como vimos en (4.2.2)

$$b_0 = \frac{\mathrm{Cov}\left(s_y^2, s_x^2\right)}{V\left(s_x^2\right)} = \frac{S_y^2(\theta-1)}{S_x^2\left(\beta_2(x)-1\right)}$$

y para este valor el estimador obtenido $\widehat{S}^2_{reg_{b_0}}$ es óptimo en el sentido de tener menor varianza. Este mínimo viene dado por

$$V\left(\widehat{S}^2_{reg_{b_0}}\right) \simeq \frac{1}{n} S_y^4\left(\left(\beta_2(y)-1\right) - \frac{(\theta-1)^2}{\beta_2(x)-1}\right),$$

que se puede escribir de la forma

$$V\left(\widehat{S}^2_{reg_{b_0}}\right) \simeq V\left(s_y^2\right)\left(1 - \rho^2\left(s_x^2, s_y^2\right)\right),$$

expresión que coincide con la obtenida para el estimador que hemos propuesto, $\widehat{S}^2_{\alpha_1}$. Por tanto este estimador tiene las mismas propiedades de precisión que el estimador de regresión.

En caso "normalidad" el estimador de regresión óptimo lo proporciona

$$b_0 = \frac{S_y^2}{S_x^2}\rho^2,$$

con lo que el estimador de regresión óptimo tiene la expresión

$$\widehat{S}^2_{reg_{b_0}} = s_y^2 + \frac{S_y^2}{S_x^2}\rho^2\left(S_x^2 - s_x^2\right),$$

y por tanto, para calcular el estimador óptimo es preciso conocer ρ^2 y S_y^2.

En el caso del estimador que presentamos, hacer notar que además de ser igual de preciso que el estimador de regresión es más sencillo de computar ya que su expresión es

$$\widehat{S}^2_{\rho^2} = s_y^2\left(\frac{S_x^2}{s_x^2}\right)^{\rho^2},$$

que sólo depende de ρ^2.

4.5.7 Estudio empírico.

Por último, como ilustración, presentamos los datos dados por *Cochran* (1977) utilizados posteriormente por *Prasad* y *Singh* (1992) para ilustrar su método.

Los datos corresponden a

y = número de casos de parálisis de polio en el grupo placebo

x = número de casos de parálisis de polio en el grupo no inoculado

Los cálculos intermedios dan los valores siguientes:

$$S_y^2 = 9.9528 \quad ; \quad S_x^2 = 10.8451 \quad ; \quad \theta = 3.2825$$

$$\beta_2(y) = 4.2998 \quad ; \quad \beta_2(x) = 5.3661$$

$$N = 34 \quad ; \quad n = 10$$

A partir de ellos se obtienen las precisiones de los estimadores $\widehat{S}_I^2$ (4.3.1), d_1 (4.3.7), $\widehat{S}^2_{\alpha_1}$ (4.5.4) y $\widehat{S}^2_{reg_{b_0}}$ (4.2.2):

$$V\left(s_y^2\right) = 24.7695 \quad ; \quad \text{ECM}\left(\widehat{S}_I^2\right) = 23.4293$$

$$V(d_1) = 21.7045 \quad ; \quad V\left(\widehat{S}^2_{\alpha_1}\right) = 15.9798 = V\left(\widehat{S}^2_{reg_{b_0}}\right)$$

y la eficiencia relativa de los diferentes estimadores de razón respecto al estimador directo s_y^2 es

$$\text{E. R.}\left(\widehat{S}^2_I, s_y^2\right) = 1.0572$$

$$\text{E. R.}\left(d_1, s_y^2\right) = 1.1412$$

$$\text{E. R.}\left(\widehat{S}^2_{\alpha_1}, s_y^2\right) = 1.5508$$

Por tanto es claro que el estimador propuesto, $\widehat{S}^2_{\alpha_1}$, es notoriamente superior en precisión a los otros estimadores s_y^2, $\widehat{S}^2_I$ y d_1, además de ser cuasi–insesgado.

4.5.8 Mejoras.

El estimador propuesto presenta las siguientes mejoras:

- es cuasi–insesgado,
- es fácil de computar,
- generaliza al estimador de Isaki y al estimador producto,
- es más preciso que el estimador de expansión simple,
- es más preciso que el estimador de Isaki,
- es más preciso que el estimador producto,
- es asintóticamente igual de preciso que el estimador de regresión,
- el conocimiento acerca de la correlación entre las variables, en el caso de poblaciones en las que los momentos sean los mismos de una normal bivariante hasta el orden cuatro, permite determinar el estimador, cosa que no es posible con el estimador de regresión, consiguiendo un aumento en precisión respecto a los estimadores simple, de Isaki y de producto.

§4.6 Comparación entre estimadores indirectos de la varianza poblacional.

El objetivo de esta sección es comparar entre sí los métodos indirectos de estimación de la varianza poblacional y respecto a la estimación directa, dando condiciones bajo las cuales unos estimadores son preferibles a otros, intentando así sintetizar las técnicas de estimación expuestas en esta segunda parte del proyecto.

A continuación vamos a hacer una comparación entre las precisiones de los distintos estimadores considerados y determinar bajo qué condiciones un estimador es preferible a los demás, en función de los momentos de orden cuatro de la distribución $(y,\, x)$, más concretamente en función de $\beta_2(x)$, $\beta_2(y)$ y θ.

Haremos el estudio para una población cualquiera y al final particularizaremos al caso de poblaciones en las que los momentos de la distribución (y, x) sean los mismos hasta el orden cuatro de una normal bivariante. Según hemos visto a lo largo de los capítulos cuatro y cinco, la precisión aproximada de los estimadores

$$\widehat{S}_I^2, \quad \widehat{S}_{reg_{b_0}}^2, \quad \widehat{S}_{\alpha_1}^2, \quad \widehat{S}_p^2,$$

viene dada, expresada en función de $\beta_2(y)$, $\beta_2(x)$ y θ por

$$\mathrm{ECM}\left(\widehat{S}_I^2\right) \simeq \frac{S_y^4}{n}\left[\beta_2(y) + \beta_2(x) - 2\theta\right],$$

$$V\left(\widehat{S}_{reg_{b_0}}^2\right) \simeq V\left(\widehat{S}_{\alpha_1}^2\right) \simeq \frac{S_y^4}{n}\left[(\beta_2(y) - 1) - \frac{(\theta - 1)^2}{\beta_2(x) - 1}\right],$$

$$\mathrm{ECM}\left(\widehat{S}_p^2\right) \simeq \frac{S_y^4}{n}\left[(\beta_2(y) - 1) + (\beta_2(x) - 1) + 2\,(\theta - 1)\right],$$

respectivamente. Además

$$V\left(s_y^2\right) = \frac{S_y^4}{n}\left(\beta_2(y) - 1\right).$$

Entonces se pueden deducir los siguientes resultados, en los que el término mejor alude a la precisión de los estimadores:

Proposición 4.6.1 *El estimador de razón $\widehat{S}_I^2$ es mejor que s_y^2 si y sólo si*

$$\frac{\beta_2(x)+1}{2} \leq \theta$$

Demostración.-

$$\mathrm{ECM}\left(\widehat{S}_I^2\right) - V\left(s_y^2\right) \simeq$$

$$\simeq \frac{S_y^4}{n}\left[\beta_2(y)+\beta_2(x)-2\theta-(\beta_2(y)-1)\right] =$$

$$= \frac{1}{n} S_y^4 \left[\beta_2(x) - 2\theta + 1\right] < 0$$

si se verifica el enunciado de la proposición.

Proposición 4.6.2 *El estimador producto $\widehat{S}_p^2$ es mejor que s_y^2 si y sólo si*

$$\theta \leq \frac{3-\beta_2(x)}{2}$$

Demostración.-

$$\mathrm{ECM}\left(\widehat{S}_p^2\right) V\left(s_y^2\right) \simeq$$

$$\simeq \frac{S_y^4}{n}\left[(\beta_2(y)-1)+(\beta_2(x)-1)+2(\theta-1)-(\beta_2(y)-1)\right] =$$

$$= \frac{1}{n} S_y^4 \left[\beta_2(x) + 2\theta - 3\right] < 0$$

siempre que se verifique la condición de la proposición.

Proposición 4.6.3 *El estimador $\widehat{S}_p^2$ es mejor que $\widehat{S}_I^2$ si y sólo si $\theta < 1$.*

Demostración.-

$$\mathrm{ECM}\left(\widehat{S}_p^2\right) - \mathrm{ECM}\left(\widehat{S}_I^2\right) =$$

$$= \frac{S_y^4}{n}\left[(\beta_2(y)+\beta_2(x)+2\theta-4)-(\beta_2(y)+\beta_2(x)-2\theta)\right] =$$

$$= \frac{S_y^4}{n} 4\left[\theta - 1\right] < 0,$$

si y sólo si $\theta < 1$.

Proposición 4.6.4 *Los estimadores* $\widehat{S}^2_{\alpha_1}$ *y* $\widehat{S}^2_{reg_{b0}}$ *son mejores que* s_y^2 *siempre, independientemente de los valores de* θ *y* $\beta_2(x)$.

Demostración.-

$$V\left(\widehat{S}^2_{\alpha_1}\right) - V\left(s_y^2\right) \simeq$$

$$\simeq \frac{S_y^4}{n}\left[(\beta_2(y) - 1) - \frac{(\theta-1)^2}{\beta_2(x)-1} - \beta_2(y) + 1\right] = \frac{S_y^4}{n}\left[-\frac{(\theta-1)^2}{\beta_2(x)-1}\right] < 0$$

siempre.

Proposición 4.6.5 *Los estimadores* $\widehat{S}^2_{\alpha_1}$ *y* $\widehat{S}^2_{reg_{b_0}}$ *son mejores que* $\widehat{S}_I^2$ *siempre.*

Demostración.-

$$V\left(\widehat{S}^2_{\alpha_1}\right) - \mathrm{ECM}\left(\widehat{S}_I^2\right) \simeq$$

$$\simeq \frac{S_y^4}{n}\left[(\beta_2(y) - 1) - \frac{(\theta-1)^2}{\beta_2(x)-1} - \beta_2(y) - \beta_2(x) + 2\theta\right] =$$

$$= \frac{S_y^4}{n}\left[2\theta - \frac{(\theta-1)^2}{\beta_2(x)-1} - \beta_2(x) - 1\right] < 0$$

siempre. Para comprobarlo, operando

$$2\theta - \frac{(\theta-1)^2}{\beta_2(x)-1} - \beta_2(x) - 1 = \frac{-\theta^2 + 2\beta_2(x)\theta - \beta_2(x)^2 - 2\beta_2(x)}{\beta_2(x)-1},$$

y por tanto el signo de la fracción es el del polinomio de segundo grado del numerador que es siempre negativo puesto que su discriminante es $-8\beta_2(x) < 0$.

Proposición 4.6.6 *El estimador de regresión* $\widehat{S}^2_{reg_{b_0}}$ *es mejor que el estimador producto* $\widehat{S}^2_p$ *a no ser que* $\theta = 2 - \beta_2(x)$ *en cuyo caso son iguales.*

Demostración.-

$$V\left(\widehat{S}^2_{reg_{b_0}}\right) - \mathrm{ECM}\left(\widehat{S}^2_p\right) \simeq$$

$$\simeq \frac{S_y^4}{n}\left[(\beta_2(y)-1) - \frac{(\theta-1)^2}{\beta_2(x)-1} - \beta_2(y) + 1 - \beta_2(x) + 1 - 2\theta + 2\right] < 0$$

siempre y para $\theta = 2-\beta_2(x)$ la diferencia es cero. Para comprobarlo, operando

$$-\frac{(\theta-1)^2}{\beta_2(x)-1} - \beta_2(x) - 2\theta + 3 = \frac{-\theta^2 + 2\theta(2-\beta_2(x)) + 4\beta_2(x) - 4 - \beta_2(x)^2}{\beta_2(x)-1} =$$

$$= \frac{-\theta^2 + 2\theta(2-\beta_2(x)) + 4\beta_2(x) - 4 - \beta_2(x)^2}{\beta_2(x)-1} = \frac{-(\theta - (2-\beta_2(x))^2}{\beta_2(x)-1} < 0.$$

Las figuras siguientes ilustran, para los diferentes valores de los parámetros $\beta_2(x)$ y θ, los niveles de precisión de cada uno de los estimadores considerados.

Figura 1. $(\beta_2(x) \geq 3)$

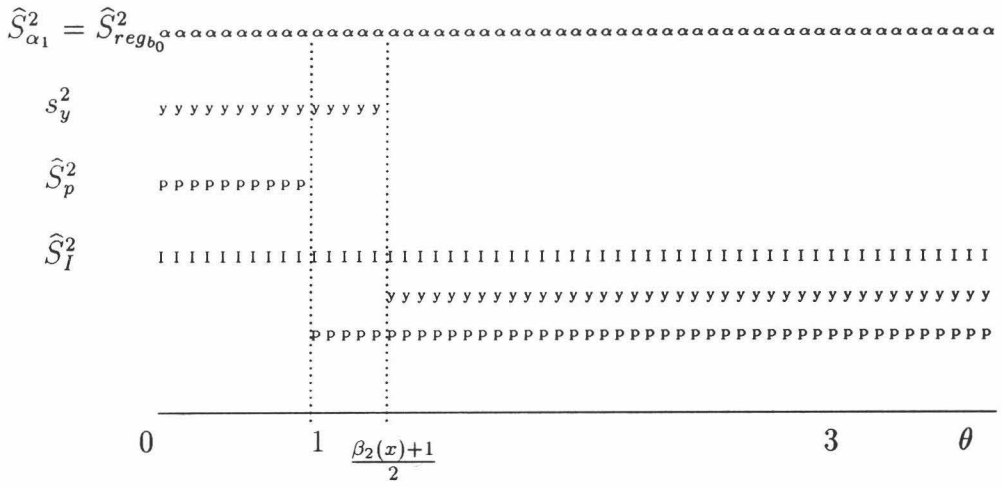
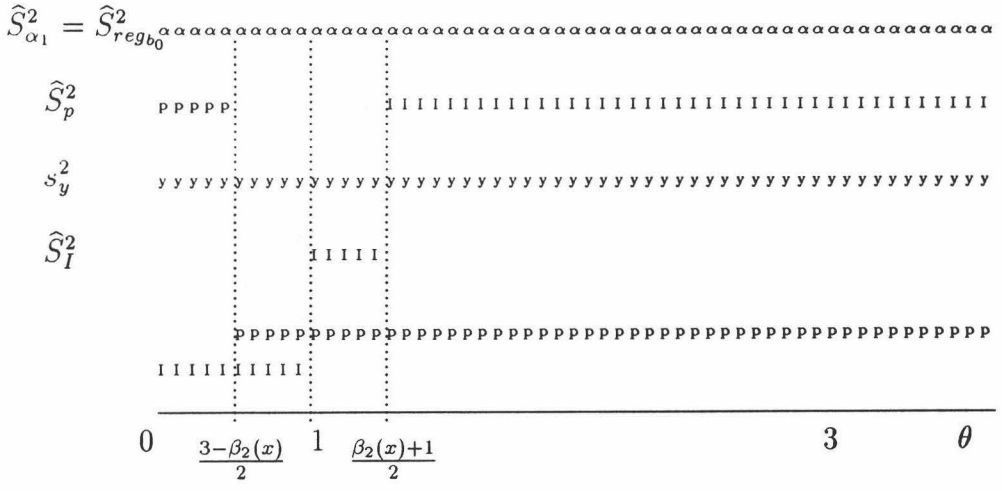


Figura 2. $(\beta_2(x) < 3)$



En ambas gráficas se observa que en cualquier caso el estimador propuesto obtenido por el método repetido de sustitución es, asintóticamente, igual de preciso que el estimador de regresión y ambos más precisos que los estimadores de razón, producto y simple.

Si el coeficiente de curtosis de la variable auxiliar x es mayor o igual que tres, la técnica de estimación de producto no mejora en precisión a la simple, pero si puede mejorar a la de razón.

Si el coeficiente de curtosis de la variable auxiliar x es menor que tres (mesocúrtica), la técnica de estimación de producto puede mejorar simúltaneamente en precisión a la de razón y a la que no utiliza información auxiliar.

En todos los casos, la técnica de estimación de producto presenta la ventaja frente a la de regresión o al método repetido de sustitución de proporcionar un estimador que se puede calcular exactamente siempre.

Capítulo 5

Estimadores múltiples.

§5.1 Introducción.

En una población finita, los métodos de estimación más conocidos y simples consideran estimadores que sólo utilizan los valores del carácter de estudio, y, observados sobre cada una de las unidades de la muestra. Sin embargo, frecuentemente se presenta el problema de utilizar la información de una variable auxiliar x con valores observables para todas las unidades de la población, ya estudiados anteriormente.

Cuando la información auxiliar de la que se dispone es múltiple x_1, x_2, ..., x_k es importante obtener estimadores que utilicen la información que proporcionan todas estas variables.

En este capítulo pretendemos estudiar las técnicas de estimación múltiple de la varianza de una variable y en una población finita conocidas para intentar mejorarlas y proponer otras técnicas de estimación nuevas.

Entre estas técnicas *Isaki* (1983) propuso una de estimación de razón múltiple al modo del estimador de *Olkin* (1958), proporcionando un estimador de razón como combinación lineal de estimadores de razón de la varianza univariantes y determinando la combinación lineal óptima con el criterio de minimizar el error cuadrático medio del estimador construido de esta forma. De este estudio y de algunas mejoras nos ocupamos en la primera sección de este capítulo.

En el mismo trabajo *Isaki* (1983) propone una técnica de estimación de regresión múltiple obteniendo un estimador para la varianza poblacional, estudiado por nosotros en la segunda sección de este capítulo.

Las dos técnicas anteriores se aplican bajo un modelo poblacional muy restrictivo, bajo el cual calculan los estimadores óptimos. Nosotros calculamos el estimador múltiple óptimo de razón en una población cualquiera.

Finalmente, en la última sección presentamos un nuevo estimador tipo razón para la varianza poblacional que solventa varios de los problemas que presentan los estimadores múltiples de razón y regresión dados por *Isaki*, desarrollado sin restringir el tipo de población.

§5.2 Estimadores de razón.

5.2.1 Introducción.

Se está interesado en estimar la varianza de una población finita y se dispone de información suplementaria proporcionada por varias variables auxiliares. Un modo de utilizar esta información mediante estimadores tipo razón lo propuso *Isaki* (1983), en muestreo con probabilidades iguales con reemplazo, en el caso en que la población tenga los mismos momentos hasta el orden cuatro de una normal multivariante $(y, x_1, \ldots, x_n)$ y suponiendo también que los coeficientes de correlación entre variables son todos iguales.

Estudiamos en esta sección esta técnica de estimación generalizando al caso de una población cualquiera, $(y, x_1, \ldots, x_n)$. Posteriormente la adaptamos a un muestreo con probabilidades iguales sin remplazo.

5.2.2 El estimador de *Isaki*.

Supongamos que se extrae una muestra de tamaño n de una población finita de tamaño N con probabilidades iguales y con reemplazo. Se pretende estimar la varianza de una variable y, S_y^2. Se conocen las varianzas poblacionales $S_{x_i}^2$ de las variables auxiliares x_i $(i = 1, \ldots, k)$ y se observan en la muestra las varianzas muestrales $s_{x_i}^2$ de las variables auxiliares x_i $(i = 1, \ldots, k)$ y de la variable principal.

Supongamos además que $(y, x_1, \ldots, x_k)$ sigue una distribución cuyos momentos son iguales a los de una normal multivariante hasta el orden cuatro, tal que

$$\rho_{yx_i} = \rho_{x_i x_j} = \rho, \quad \forall i, j, \tag{5.2.1}$$

y que $-k^{-1} < \rho < 1$.

A partir de los estimadores de razón univariantes

$$\widehat{S}^2_{R_i} = s_y^2 \frac{S^2_{x_i}}{s^2_{x_i}}, \quad i = 1, \ldots, k,$$

Isaki propone un estimador de razón múltiple siguiendo la idea de *Olkin* (1958) de la forma:

$$\widehat{S}^2_{yM} = \sum_{i=1}^{k} w_i \widehat{S}^2_{R_i} = \sum_{i=1}^{k} w_i s_y^2 \frac{S^2_{x_i}}{s^2_{x_i}},$$

donde los pesos w_i verifican $\sum_{i=1}^{k} w_i = 1$.

Propiedades.

Este estimador es sesgado y una aproximación de su error cuadrático medio, bajo el modelo dado en (5.2.1) viene dada por la expresión

$$\mathrm{ECM}\left(\widehat{S}^2_{yM}\right) \simeq \frac{S_y^4}{n} 2\left(1-\rho^2\right)\left(2\sum_{i=1}^{k} w_i^2 + \sum_{i \neq j} w_i w_j\right).$$

Omitimos la comprobación pues la haremos en el siguiente apartado en un caso más general, sin suponer el modelo dado por *Isaki.*

Determinación del estimador óptimo.

Isaki determina los pesos que hacen mínimo el error cuadrático medio del estimador, minimizando

$$\mathrm{ECM}\left(\widehat{S}^2_{yM}\right),$$

con la condición

$$\sum_{i=1}^{k} w_i = 1,$$

obteniendo que el óptimo se alcanza para

$$w_1^0 = \cdots = w_k^0 = \frac{1}{k},$$

con un error cuadrático medio mínimo aproximado de

$$\mathrm{ECM}_0\left(\widehat{S}^2_{yM}\right) \simeq \frac{S_y^4}{n} 2\left(1-\rho^2\right)\frac{k+1}{k}.$$

Para comprobarlo, construimos la función

$$F\left(w_i\right) = 2\sum_{i=1}^{k} w_i^2 + \sum_{i\neq j} w_i w_j + \lambda\left(\sum_{i=1}^{k} w_i - 1\right),$$

y derivamos

$$\frac{\partial F}{\partial w_i} = 4\sum_{i=1}^{k} w_i + \sum_{j\neq i} w_j + \lambda = 0, \quad i = 1,\ldots,k,$$

o equivalentemente

$$1 - w_i = -\lambda - 4; \quad w_i = 5 + \lambda,$$

y donde imponiendo

$$\sum_{i=1}^{k} w_i = 1; \quad \sum_{i=1}^{k}(5+\lambda) = k(5+\lambda) = 1,$$

se obtiene

$$\lambda = \frac{1-5k}{k},$$

con lo que

$$w_i = 5 + \frac{1-5k}{k} = \frac{1}{k}, \quad i = 1,\ldots,k.$$

Comparación con el estimador directo.

Bajo el modelo (5.2.1), se tiene

$$V\left(s_y^2\right) \simeq \frac{S_y^4}{n} 2.$$

Así, el estimador $\widehat{S}^2_{yM}$ es más preciso que el estimador s_y^2 si y sólo si

$$\mathrm{ECM}_0\left(\widehat{S}^2_{yM}\right) - V\left(s^2_y\right) \simeq \frac{S^4_y}{n} 2\left(1-\rho^2\right)\frac{k+1}{k} - \frac{S^4_y}{n} 2 < 0,$$

si y sólo si

$$\left(1-\rho^2\right)\frac{k+1}{k} - 1 < 0,$$

si y sólo si

$$\rho^2 > \frac{1}{k+1}.$$

Notar que para $k = 1$, la condición es la obtenida al comparar la precisión del estimador de razón univariante con la variable auxiliar x_1 con el estimador de expansión simple, ya vista en el capítulo anterior:

$$\rho^2 > \frac{1}{2}.$$

5.2.3 Extensión al caso de una población cualquiera.

Definición del estimador.

Supongamos que se extrae una muestra de tamaño n de una población finita de tamaño N con probabilidades iguales y con reemplazo. Se pretende estimar la varianza de una variable y, S^2_y. Se conocen las varianzas poblacionales $S^2_{x_i}$ de las variables auxiliares x_i $(i = 1, \ldots, k)$ y se observan en la muestra las varianzas muestrales $s^2_{x_i}$ de las variables auxiliares x_i $(i = 1, \ldots, k)$ y de la variable princial.

A partir de los estimadores de razón univariantes

$$\widehat{S}^2_{R_i} = s^2_y \frac{S^2_{x_i}}{s^2_{x_i}}, \quad i = 1, \ldots, k,$$

proponemos un estimador de razón múltiple siguiendo la idea de *Olkin* (1958) de la forma:

$$\widehat{S}^2_{yG} = \sum_{i=1}^{k} w_i \widehat{S}^2_{R_i} = \sum_{i=1}^{k} w_i s^2_y \frac{S^2_{x_i}}{s^2_{x_i}},$$

donde los pesos w_i verifican $\sum_{i=1}^{k} w_i = 1$.

Sesgo.

El estimador anterior es sesgado y una aproximación de su sesgo viene dada en la siguiente proposición:

Proposición 5.2.1 *Una aproximación del sesgo del estimador $\widehat{S}^2_{yG}$ viene dada por la expresión:*

$$\text{sesgo}\left(\widehat{S}^2_{yG}\right) \simeq \frac{S_y^2}{n}\left(\sum_{i=1}^{k} w_i\left(\beta_2(x_i)-1\right) - \sum_{i=1}^{k} w_i\left(\theta_{yx_i}-1\right),\right).$$

Demostración.-
En efecto, llamando

$$e_0 = \frac{s_y^2 - S_y^2}{S_y^2}; \quad e_i = \frac{s_{x_i}^2 - S_{x_i}^2}{S_{x_i}^2}, \quad i = 1, \ldots, k,$$

obtenemos

$$\widehat{S}^2_{R_i} = S_y^2 \frac{1+e_0}{1+e_i} = S_y^2\left(1+e_0\right)\left(1+e_i\right)^{-1} =$$

(siempre que $\mid e_i \mid < 1$)

$$= S_y^2\left(1+e_0\right)\left(1-e_i+e_i^2-\cdots\right) = S_y^2\left(1+e_0-e_i-e_0e_i+e_i^2+\cdots\right).$$

De esta forma, aproximamos la diferencia

$$\widehat{S}^2_{R_i} - S_y^2 \simeq S_y^2\left(e_0-e_i-e_0e_i+e_i^2\right)$$

con lo que

$$\text{sesgo}\left(\widehat{S}^2_{R_i}\right) \simeq S_y^2\left(-E\left(e_0e_i\right)+E\left(e_i^2\right)\right).$$

Ahora bien

$$\text{sesgo}\left(\widehat{S}^2_{yG}\right) = E\left(\widehat{S}^2_{yG} - S^2_y\right) = E\left(\sum_{i=1}^{k} w_i \widehat{S}^2_{R_i} - S^2_y\right) =$$

$$= E\left(\sum_{i=1}^{k} w_i \widehat{S}^2_{R_i} - \sum_{i=1}^{k} w_i S^2_y\right) = \sum_{i=1}^{k} w_i E\left(\widehat{S}^2_{R_i} - S^2_y\right) = \sum_{i=1}^{k} w_i \,\text{sesgo}\left(\widehat{S}^2_{R_i}\right) \simeq$$

$$\simeq S^2_y\left(-\sum_{i=1}^{k} w_i E\left(e_0 e_i\right) + \sum_{i=1}^{k} w_i E\left(e_i^2\right)\right) =$$

$$= S^2_y\left(-\sum_{i=1}^{k} w_i \frac{\text{Cov}\left(s^2_y, s^2_{x_i}\right)}{S^2_y S^2_{x_i}} + \sum_{i=1}^{k} w_i \frac{V\left(s^2_{x_i}\right)}{S^2_{x_i}}\right).$$

Entonces, en muestreo con probabilidades iguales con reemplazo, se tiene

$$\frac{\text{Cov}\left(s^2_y, s^2_{x_i}\right)}{S^2_y S^2_{x_i}} \simeq \frac{1}{n}\left(\theta_{yx_i} - 1\right); \quad \frac{V\left(s^2_{x_i}\right)}{S^2_{x_i}} \simeq \frac{1}{n}\left(\beta_2(x_i) - 1\right), \tag{5.2.2}$$

con lo que, sustituyendo se llega a la expresión aproximada para el sesgo antes expuesta.

Precisión.

El estimador $\widehat{S}^2_{yG}$ es sesgado por lo que su precisión la medirá su error cuadrático medio. Una expresión aproximada de éste la proporciona la siguiente proposición:

Proposición 5.2.2 *Una aproximación del error cuadrático medio del estimador* $\widehat{S}^2_{yG}$ *es*

$$\text{ECM}\left(\widehat{S}^2_{yG}\right) \simeq \frac{S^4_y}{n}\left(\beta_2(y) + \sum_{i=1}^{k} w_i^2 \beta_2(x_i) + \sum_{i \neq j} w_i w_j \theta_{x_i x_j} - 2\sum_{i=1}^{k} w_i \theta_{yx_i}\right).$$

Demostración.-
Para comprobarlo, partimos de la expresión

$$\mathrm{ECM}\left(\widehat{S}^2_{yG}\right) = \sum_{i=1}^{k} w_i^2\, \mathrm{ECM}\left(\widehat{S}^2_{R_i}\right) + \sum_{i\neq j} w_i w_j E\left(\widehat{S}^2_{R_i} - S_y^2\right)\left(\widehat{S}^2_{R_j} - S_y^2\right).$$

Calculamos aproximadamente la expresión

$$E\left(\widehat{S}^2_{R_i} - S_y^2\right)\left(\widehat{S}^2_{R_j} - S_y^2\right),$$

mediante las variables e_0 y e_i, $i = 1, \ldots, k$, como en el caso del sesgo:

$$\widehat{S}^2_{R_i} - S_y^2 \simeq S_y^2\left(e_0 - e_i - e_0 e_i\right),$$

con lo que

$$\left(\widehat{S}^2_{R_i} - S_y^2\right)\left(\widehat{S}^2_{R_j} - S_y^2\right) \simeq S_y^4\left(e_0^2 - e_0 e_i - e_0 e_j + e_i e_j\right),$$

y de esta forma

$$E\left(\widehat{S}^2_{R_i} - S_y^2\right)\left(\widehat{S}^2_{R_j} - S_y^2\right) \simeq S_y^4\left(E\left(e_0^2\right) - E\left(e_0 e_i\right) - E\left(e_0 e_j\right) + E\left(e_i e_j\right)\right) =$$

$$\simeq S_y^4\left(\frac{V\left(s_y^2\right)}{S_y^4} - \frac{\mathrm{Cov}\left(s_y^2, s_{x_i}^2\right)}{S_y^2 S_{x_i}^2} - \frac{\mathrm{Cov}\left(s_y^2, s_{x_j}^2\right)}{S_y^2 S_{x_j}^2} + \frac{\mathrm{Cov}\left(s_{x_i}^2, s_{x_j}^2\right)}{S_{x_i}^2 S_{x_j}^2}\right),$$

y sustituyendo

$$\frac{V\left(s_y^2\right)}{S_y^2} \simeq \frac{1}{n}\left(\beta_2(y) - 1\right), \quad \frac{\mathrm{Cov}\left(s_{x_i}^2, s_{x_j}^2\right)}{S_{x_i}^2 S_{x_j}^2} \simeq \frac{1}{n}\left(\theta_{x_i x_j} - 1\right),$$

y los valores dados en (5.2.2), se tiene

$$E\left(\widehat{S}^2_{R_i} - S_y^2\right)\left(\widehat{S}^2_{R_j} - S_y^2\right) \simeq \frac{S_y^4}{n}\left(\beta_2(y) - \theta_{yx_i} - \theta_{yx_j} + \theta_{x_i x_j}\right).$$

Sustituyendo ahora en la expresión del error cuadrático medio del estimador $\widehat{S}^2_{yG}$, teniendo en cuenta la aproximación del error cuadrático medio del estimador de razón univariante dada en capítulo cuatro (4.3.6):

$$\mathrm{ECM}\left(\widehat{S}^2_{R_i}\right) \simeq \frac{1}{n} S_y^4 \left(\beta_2(y) + \beta_2(x_i) - 2\theta_{yx_i}\right),$$

se llega a que

$$\mathrm{ECM}\left(\widehat{S}^2_{yG}\right) = \sum_{i=1}^{k} w_i^2 \,\mathrm{ECM}\left(\widehat{S}^2_{R_i}\right) + \sum_{i\neq j} w_i w_j E\left(\widehat{S}^2_{R_i} - S_y^2\right)\left(\widehat{S}^2_{R_j} - S_y^2\right) \simeq$$

$$\simeq \sum_{i=1}^{k} w_i^2 \left(\beta_2(y) + \beta_2(x_i) - 2\theta_{yx_i}\right) + \sum_{i\neq j} w_i w_j \left(\beta_2(y) - \theta_{yx_i} - \theta_{yx_j} + \theta_{x_i x_j}\right) \simeq$$

$$\simeq \frac{S_y^4}{n}\left[\beta_2(y)\left(\sum_{i=1}^{k} w_i^2 + \sum_{i\neq j} w_i w_j\right) + \sum_{i=1}^{k} w_i^2 \beta_2(x_i) + \sum_{i\neq j} w_i w_j \theta_{x_i x_j} - \right.$$

$$\left. - \left(2\sum_{i=1}^{k} w_i^2 \theta_{yx_i} + \sum_{i\neq j} w_i w_j \theta_{yx_i} + \sum_{i\neq j} w_i w_j \theta_{yx_j}\right)\right] \simeq$$

$$\simeq \frac{S_y^4}{n}\left[\beta_2(y) + \sum_{i=1}^{k} w_i^2 \beta_2(x_i) + \sum_{i\neq j} w_i w_j \theta_{x_i x_j} - \right.$$

$$\left. - \left(2\sum_{i=1}^{k} w_i^2 \theta_{yx_i} + \sum_{i=1}^{k} w_i \left(1 - w_i\right)\theta_{yx_i} + \sum_{j=1}^{k} w_j \left(1 - w_j\right)\theta_{yx_j}\right)\right],$$

$$\simeq \frac{S_y^4}{n}\left[\beta_2(y) + \sum_{i=1}^{k} w_i^2 \beta_2(x_i) + \sum_{i\neq j} w_i w_j \theta_{x_i x_j} - 2\sum_{i=1}^{k} w_i \theta_{yx_i}\right].$$

Determinación del estimador óptimo.

Para determinar los pesos que hacen el error cuadrático medio del estimador mínimo, reescribimos su expresión aproximada en la forma:

$$\mathrm{ECM}\left(\widehat{S}^2_{yG}\right) \simeq \frac{S_y^4}{n} w' A w,$$

donde $w = w_{(k\times 1)} = (w_1, \ldots, w_k)'$, $A = A_{(k\times k)} = (a_{ij})$ con

$$a_{ij} = \beta_2(y) - \theta_{yx_i} - \theta_{yx_j} + \theta_{x_ix_j},$$

y haciendo notar que para $i = j$

$$a_{ii} = \beta_2(y) - \theta_{yx_i} - \theta_{yx_i} + \theta_{x_ix_i} = \beta_2(y) + \beta_2(x_i) - 2\theta_{yx_i}.$$

Proposición 5.2.3 *Los pesos que minimizan el error cuadrático medio del estimador* $\widehat{S}^2_{yG}$ *son*

$$w^0 = \frac{A^{-1}e}{e'A^{-1}e},$$

y el error cuadrático medio mínimo es, aproximadamente,

$$\mathrm{ECM}_0\left(\widehat{S}^2_{yG}\right) \simeq \frac{S_y^4}{n}\frac{1}{e'A^{-1}e}.$$

donde $e = e_{(k\times 1)} = (1, \ldots, 1)'$.

Demostración.-
Minimizar el error cuadrático medio del estimador $\widehat{S}^2_{yG}$ equivale a calcular

$$\text{mín } w'Aw = \frac{1}{e'A^{-1}e},$$

siempre que A^{-1} exista.

De esta forma, los pesos que minimizan el error cuadrático medio son

$$w^0 = \frac{A^{-1}e}{e'A^{-1}e},$$

y el error cuadrático medio mínimo es, aproximadamente,

$$\mathrm{ECM}_0\left(\widehat{S}^2_{yG}\right) \simeq \frac{S_y^4}{n}\frac{1}{e'A^{-1}e}.$$

Caso particular del modelo propuesto por *Isaki.*

En el caso en que la población tenga los mismos momentos hasta el orden cuatro de una normal multivariante $(y, x_1, \ldots, x_n)$ y suponiendo también que los coeficientes de correlación entre variables sean todos iguales:

$$\rho(y, x_i) = \rho\left(x_i, x_j\right) = \rho, \quad \forall i, j,$$

la aproximación del sesgo, teniendo en cuenta que bajo este modelo

$$\left(\theta_{yx_i} - 1\right) = 2\rho^2, \quad (\beta_2(x_i) - 1) = 2 \quad i = 1, \ldots, k,$$

conduce a

$$\text{sesgo}\left(\widehat{S}^2_{yG}\right) \simeq \frac{S_y^2}{n}\left(\sum_{i=1}^{k} 2w_i - \sum_{i=1}^{k} 2w_i\rho^2\right) = \frac{S_y^2}{n}\left(2 - 2\rho^2\right) = \frac{2S_y^2}{n}\left(1 - \rho^2\right).$$

De la misma forma, el error cuadrático medio es aproximadamente y bajo el modelo

$$\text{ECM}\left(\widehat{S}^2_{yG}\right) \simeq \frac{S_y^4}{n}\left(\beta_2(y) + \sum_{i=1}^{k} w_i^2\beta_2(x_i) + \sum_{i\neq j} w_i w_j \theta_{x_i x_j} - 2\sum_{i=1}^{k} w_i \theta_{yx_i}\right) =$$

$$= \frac{S_y^4}{n}\left(3 + 3\sum_{i=1}^{k} w_i^2 + 2\rho^2 \sum_{i\neq j} w_i w_j + \sum_{i\neq j} w_i w_j - 2\sum_{i=1}^{k} w_i - 4\rho^2 \sum_{i=1}^{k} w_i\right) =$$

$$= \frac{S_y^4}{n}\left(4 - 4\rho^2 + \sum_{i\neq j} w_i w_j\left(-2 + 2\rho^2\right)\right) = \frac{S_y^4}{n}\left(2 - 2\rho^2\right)\left(2 - \sum_{i\neq j} w_i w_j\right).$$

Finalmente, teniendo en cuenta de nuevo que

$$1 = \sum_{i=1}^{k} w_i^2 + \sum_{i\neq j} w_i w_j,$$

se obtiene

$$\mathrm{ECM}\left(\widehat{S}^2_{yG}\right) \simeq \frac{S_y^4}{n} 2\left(1-\rho^2\right)\left(2\sum_{i=1}^{k} w_i^2 + \sum_{i\neq j} w_i w_j\right).$$

Para la determinación de los pesos que proporcionan el estimador óptimo hay que notar que, bajo el modelo (5.2.1) se tiene

$$a_{ij} = \begin{cases} 2\left(1-\rho^2\right) & \text{si} \quad i \neq j \\ 4\left(1-\rho^2\right) & \text{si} \quad i = j, \end{cases}$$

y por tanto

$$A = 2\left(1-\rho^2\right)\begin{pmatrix} 2 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 2 \end{pmatrix} = 2\left(1-\rho^2\right) B.$$

Tenemos

$$B^{-1} = \begin{pmatrix} 2 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 2 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{k}{k+1} & \frac{-1}{k+1} & \cdots & \frac{-1}{k+1} \\ \frac{-1}{k+1} & \frac{k}{k+1} & \cdots & \frac{-1}{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-1}{k+1} & \frac{-1}{k+1} & \cdots & \frac{k}{k+1} \end{pmatrix}$$

y por tanto $B^{-1}e = \left(\frac{1}{k+1}, \ldots, \frac{1}{k+1}\right)'$ y $e'B^{-1}e = \frac{k}{k+1}$. De esta forma, los pesos que minimizan el error cuadrático medio son

$$w^0 = \frac{A^{-1}e}{e'A^{-1}e} = \frac{\frac{1}{2(1-\rho^2)}B^{-1}e}{\frac{1}{2(1-\rho^2)}e'B^{-1}e} = \frac{k+1}{k}\left(\frac{1}{k+1}, \ldots, \frac{1}{k+1}\right)' = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right)'.$$

y el error cuadrático medio mínimo es, aproximadamente,

$$\mathrm{ECM}_0\left(\widehat{S}^2_{yG}\right) \simeq \frac{S_y^4}{n}\frac{1}{e'A^{-1}e} = 2\left(1-\rho^2\right)\frac{k+1}{k}.$$

5.2.4 Extensión a SRSWOR.

Definición del estimador.

Supongamos que se extrae una muestra de tamaño n de una población finita de tamaño N con probabilidades iguales y sin reemplazo, que se se conocen las varianzas poblacionales $S^2_{x_i}$ de las variables auxiliares x_i $(i = 1, \ldots, k)$ y se observan en la muestra las varianzas muestrales $s^2_{x_i}$ de las variables auxiliares x_i $(i = 1, \ldots, k)$ y s^2_y.

Para estimar la varianza de la variable y, S^2_y, proponemos en este diseño muestral el estimador

$$\widehat{S}^2_{yO} = \sum_{i=1}^{k} w_i \widehat{S}^2_{R_i} = \sum_{i=1}^{k} w_i s^2_y \frac{S^2_{x_i}}{s^2_{x_i}},$$

donde los pesos w_i verifican $\sum_{i=1}^{k} w_i = 1$.

Sesgo.

El estimador anterior es sesgado y una aproximación de su sesgo viene dada en la siguiente proposición:

Proposición 5.2.4 *Una aproximación del sesgo del estimador $\widehat{S}^2_{yO}$ viene dada por la expresión:*

$$\text{sesgo}\left(\widehat{S}^2_{yO}\right) \simeq S^2_y \frac{(N-n)}{(N-2)n} \sum_{i=1}^{k} w_i \left(K\left(\beta_2(x_i) - \theta_{yx_i}\right)\right).$$

Demostración.-

Procediendo como en el caso del estimador $\widehat{S}^2_{yG}$ se llega a la expresión aproximada

$$\text{sesgo}\left(\widehat{S}^2_{yO}\right) \simeq S^2_y \left(-\sum_{i=1}^{k} w_i \frac{\text{Cov}\left(s^2_y, s^2_{x_i}\right)}{S^2_y S^2_{x_i}} + \sum_{i=1}^{k} w_i \frac{V\left(s^2_{x_i}\right)}{S^2_{x_i}}\right).$$

Recordando las expresiones dadas en (4.3.3) y (4.3.4), podemos escribir

$$\text{sesgo}\left(\widehat{S}^2_{yO}\right) \simeq S^2_y \frac{N-n}{(N-2)n}\left(-\sum_{i=1}^{k} w_i\left(K\theta_{yx_i} + K_1\rho^2_{yx_i} - K_2\right) + \right.$$

$$\left. + \sum_{i=1}^{k} w_i\left(K\beta_2(x_i) - M\right)\right),$$

y operando

$$\text{sesgo}\left(\widehat{S}^2_{yO}\right) \simeq S^2_y \frac{(N-n)}{(N-2)n}\sum_{i=1}^{k} w_i\left(K\left(\beta_2(x_i) - \theta_{yx_i}\right) - M - K_1\rho^2_{yxi} + K_2\right),$$

expresión que se aproxima a la que queríamos comprobar cuando la correlación entre las variables es alta, pues en tal caso

$$M + K_1\rho^2_{yx_i} - K_2 \simeq 0.$$

Precisión.

Proposición 5.2.5 *Una aproximación del error cuadrático medio del estimador* $\widehat{S}^2_{yO}$ *es*

$$\text{ECM}\left(\widehat{S}^2_{yO}\right) \simeq S^4_y \frac{N-n}{(N-2)n} w'Cw$$

donde $w = w_{(k\times 1)} = (w_1, \ldots, w_k)'$, $C = C_{(k\times k)} = (c_{ij})$ *con*

$$c_{ij} = K\left(\beta_2(y) - \theta_{yx_i} - \theta_{yx_j} + \theta_{x_ix_j}\right) + K_1\left(-\rho^2_{yx_i} - \rho^2_{yx_j} + \rho^2_{x_ix_j} + 1\right).$$

Demostración.-
Procediendo como en el diseño SRSWR llegamos a la expresión

$$\text{ECM}\left(\widehat{S}^2_{yO}\right) \simeq S^4_y\left[\sum_{i=1}^{k} w_i^2\left\{\frac{V\left(s^2_y\right)}{S^4_y} + \frac{V\left(s^2_{x_i}\right)}{S^4_{x_i}} - 2\frac{\text{Cov}\left(s^2_{x_i}, s^2_y\right)}{S^2_{x_i}S^2_y}\right\} + \right.$$

$$+\sum_{i\neq j} w_i w_j \left\{ \frac{V\left(s_y^2\right)}{S_y^4} - \frac{\mathrm{Cov}\left(s_y^2, s_{x_i}^2\right)}{S_y^2 S_{x_i}^2} - \frac{\mathrm{Cov}\left(s_y^2, s_{x_j}^2\right)}{S_y^2 S_{x_j}^2} + \frac{\mathrm{Cov}\left(s_{x_i}^2, s_{x_j}^2\right)}{S_{x_i}^2 S_{x_j}^2} \right\} \Bigg].$$

De las expresiones dadas en (4.3.3) y (4.3.4) obtenemos

$$\mathrm{ECM}\left(\widehat{S}_{yO}^2\right) \simeq S_y^4 \left[\sum_{i=1}^{k} w_i^2 \left\{ (K\beta_2(y) - M) + (K\beta_2(x_i) - M) - \right.\right.$$

$$\left. -2\left(K\theta_{yx_i} + K_1\rho_{yx_i}^2 - K_2\right)\right\} + \sum_{i\neq j} w_i w_j \left\{ (K\beta_2(y) - M) - \right.$$

$$\left. - \left(K\theta_{yx_i} + K_1\rho_{yx_i}^2 - K_2\right) - \left(K\theta_{yx_j} + K_1\rho_{yx_j}^2 - K_2\right) + \right.$$

$$\left. \left. + \left(K\theta_{x_i x_j} + K_1\rho_{x_i x_j}^2 - K_2\right)\right\}\right] =$$

$$= S_y^4 \frac{N-n}{(N-2)n} \left[\sum_{i=1}^{k} w_i^2 \left\{ K\left(\beta_2(y) + \beta_2(x_i) - 2\theta_{yx_i}\right) + 2K_1\left(1 - \rho_{yx_i}^2\right)\right\} + \right.$$

$$\left. + \sum_{i\neq j} w_i w_j \left\{ K\left(\beta_2(y) - \theta_{yx_i} - \theta_{yx_j} + \theta_{x_i x_j}\right) + K_1\left(\rho_{x_i x_j}^2 - \rho_{yx_i}^2 - \rho_{yx_j}^2 + 1\right)\right\}\right] =$$

$$= S_y^4 \frac{N-n}{(N-2)n} \sum_{i\neq j} w_i w_j \left\{ K\left(\beta_2(y) - \theta_{yx_i} - \theta_{yx_j} + \theta_{x_i x_j}\right) + \right.$$

$$\left. + K_1\left(\rho_{x_i x_j}^2 - \rho_{yx_i}^2 - \rho_{yx_j}^2 + 1\right)\right\} = S_y^4 \frac{N-n}{(N-2)n} w' C w.$$

Determinación del estimador óptimo.

Proposición 5.2.6 *Los pesos que minimizan el error cuadrático medio del estimador* $\widehat{S}_{yO}^2$ *son*

$$w^0 = \frac{C^{-1}e}{e'C^{-1}e},$$

y el error cuadrático medio mínimo es, aproximadamente,

$$\mathrm{ECM}_0\left(\widehat{S}^2_{yO}\right) \simeq S_y^4 \frac{N-n}{(N-2)n} \frac{1}{e'C^{-1}e}.$$

donde $e = e_{(k\times 1)} = (1,\ldots,1)'$.

Demostración.-

Minimizar el error cuadrático medio del estimador $\widehat{S}^2_{yO}$ equivale a calcular

$$\text{mín } w'Cw = \frac{1}{e'C^{-1}e},$$

siempre que C^{-1} exista. Como en el caso del diseño SRSWR, aplicando la desigualdad extendida de Cauchy–Swartz, el error cuadrático medio mínimo es, aproximadamente,

$$\mathrm{ECM}_0\left(\widehat{S}^2_{yO}\right) \simeq S_y^4 \frac{N-n}{(N-2)n} \frac{1}{e'C^{-1}e},$$

y se alcanza para los pesos

$$w^0 = \frac{C^{-1}e}{e'C^{-1}e}.$$

Casos particulares.

Si el tamaño de la población, N, es suficientemente grande,

$$K \simeq 1; \quad K_1 \simeq 0; \quad \text{y} \quad \frac{N-n}{N-2} \simeq 1,$$

por lo que las expresiones aproximadas de los errores cuadráticos medios de los estimadores $\widehat{S}^2_{yO}$ y $\widehat{S}^2_{yG}$ coinciden. Por tanto, el inconveniente que supone trabajar con los coeficientes que aparecen en el caso del muestreo sin reemplazo puede solventarse, si el tamaño de la población es suficientemente grande, utilizando los resultados obtenidos para el muestreo con reemplazo, puesto que son equivalentes.

Destacar también que si además la población es "normal multivariante" y se cumple el modelo de Isaki

$$\text{mín ECM}\left(\widehat{S}^2_{yO}\right) \simeq \text{mín ECM}\left(\widehat{S}^2_{yG}\right) \simeq \frac{S_y^4}{n} 2\left(1-\rho^2\right)\left(2\sum_{i=1}^{k} w_i^2 + \sum_{i\neq j} w_i w_j\right),$$

y los pesos óptimos son

$$w^0 = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right)',$$

puesto que en ese caso $c_{ij} \simeq a_{ij}, \forall i, j$.

De esta forma, el estimador $\widehat{S}^2_{yO}$ es más preciso que el estimador simple s_y^2 si y sólo si

$$\frac{1}{k+1} < \rho^2.$$

§5.3 El estimador de regresión.

5.3.1 El estimador de Isaki.

A continuación estudiamos la extensión múltiple del estimador de regresión dado en el capítulo cuatro al caso de varias variables auxiliares para estimar la varianza poblacional de la variable objeto de estudio y.

Definición.

Isaki define el estimador múltiple de regresión para la varianza poblacional de y de la forma

$$\widehat{S}^2_{MR} = s_y^2 + \sum_{i=1}^{k} B_i\left(S^2_{x_i} - s^2_{x_i}\right). \tag{5.3.1}$$

Precisión.

El estimador $\widehat{S}^2_{MR}$ es insesgado y su precisión la proporciona su varianza que tiene por expresión exacta

$$V\left(\widehat{S}^2_{MR}\right) = V\left(s_y^2\right) - 2\sum_{i=1}^{k} B_i \operatorname{Cov}\left(s_y^2, s^2_{x_i}\right) + \sum_{i=1}^{k} B_i^2 V\left(s^2_{x_i}\right) +$$

$$+\sum_{i\neq j}^{k} B_i B_j \operatorname{Cov}\left(s_{x_i}^2, s_{x_j}^2\right). \tag{5.3.2}$$

5.3.2 Determinación del estimador óptimo.

Para determinar el estimador óptimo hay que minimizar su varianza, dada en (5.3.2). La derivada parcial de ésta con respecto a B_i viene dada por la expresión

$$\frac{\partial V\left(\widehat{S}_{MR}^2\right)}{\partial B_i} = 2B_i V\left(s_{x_i}^2\right) - 2\operatorname{Cov}\left(s_y^2, s_{x_i}^2\right) + \sum_{j\neq i} B_j \operatorname{Cov}\left(s_{x_i}^2, s_{x_j}^2\right) =$$

$$= 2B_i V\left(s_{x_i}^2\right) - 2\operatorname{Cov}\left(s_y^2, s_{x_i}^2\right) + 2\sum_{\substack{j=1\\ j\neq i}} B_j \operatorname{Cov}\left(s_{x_i}^2, s_{x_j}^2\right), \quad i = 1, \ldots, k.$$

De esta forma los pesos que minimizan la varianza, B_i^0, $i = 1, \ldots, k$, son las soluciones del sistema

$$\frac{\partial V\left(\widehat{S}_{MR}^2\right)}{\partial B_i} = 0, \quad i = 1, \ldots, k. \tag{5.3.3}$$

Para su resolución y puesto que depende de parámetros desconocidos como $V\left(s_{x_i}^2\right)$, $\operatorname{Cov}\left(s_y^2, s_{x_i}^2\right)$, $\operatorname{Cov}\left(s_{x_i}^2, s_{x_j}^2\right)$, Isaki propone que éstos sean sustituídos, en el caso de que la población $(y, x_1, \ldots, x_k)$ sea "normal multivariante" (en el sentido antes expuesto), por estimadores consistentes como $\left(s_{x_i}^2\right)^2$, $s_{yx_i}^2$, $s_{x_i x_j}^2$. De esta forma, escribiendo $A = A_{(k\times k)} = a_{ij}$ con

$$a_{ij} = \begin{cases} V\left(s_{x_i}^2\right) & \text{si} \quad i = j \\ \operatorname{Cov}\left(s_{x_i}^2, s_{x_j}^2\right) & \text{si} \quad i \neq j, \end{cases}$$

$\widehat{A} = \widehat{A}_{(k\times k)} = \widehat{a}_{ij}$ con

$$\widehat{a}_{ij} = \begin{cases} \left(s_{x_i}^2\right)^2 & \text{si} \quad i = j \\ s_{x_i x_j}^2 & \text{si} \quad i \neq j, \end{cases}$$

$C = C_{(k\times 1)} = (c_1, \ldots, c_k)'$ con $c_i = \text{Cov}\left(s_y^2, s_{x_i}^2\right)$, $i = 1, \ldots, k$ y $\widehat{C} = \widehat{C}_{(k\times 1)} = (\widehat{c}_1, \ldots, \widehat{c}_k)'$ con $\widehat{c}_i = s_{yx_i}^2$, $i = 1, \ldots, k$, el sistema original se escribe 0

$$AB = C,$$

donde $B = B_{(k\times 1)} = (B_1, \ldots, B_k)'$, y el sistema con las sustituciones se escribe como

$$\widehat{A}B = \widehat{C}.$$

Así, si $\widehat{A}^{-1}$ existe, y suponiendo que $\widehat{B}^0$ sean las soluciones del sitema aproximado, usando un resultado debido a *Fuller* (1976), se tiene

$$\widehat{B}^0 - \widehat{B} = O_p\left(n^{-\frac{1}{2}}\right).$$

De esta forma, *Isaki* define el estimador

$$\widehat{S}_{MRE}^2 = s_y^2 + \sum_{i=1}^{k} \widehat{B}_i \left(S_{x_i}^2 - s_{x_i}^2\right), \tag{5.3.4}$$

y comprueba que, bajo el modelo

$$\rho_{yx_i} = \rho_{x_ix_j} = \rho, \quad \forall i, j,$$

la varianza del estimador de regresión múltiple con los B_i óptimos, $V\left(\widehat{S}_{MR}^2\right)$ y la varianza del estimador de regresión múltiple con los $\widehat{B}_i^0$ óptimos, $V\left(\widehat{S}_{MRE}^2\right)$, son asintóticamente iguales.

Además, bajo este modelo, $V\left(\widehat{S}_{MR}^2\right)$ es mínima para

$$B_i^0 = \frac{1}{1 + (k-1)\rho^2}\rho^2 R_i, \quad i = 1, \ldots, k,$$

donde $R_i = \dfrac{S_y^2}{S_{x_i}^2}$.

En efecto, puesto que bajo el modelo

$$V\left(s_{x_i}^2\right) = \frac{2}{n}S_{x_i}^4,$$

$$\text{Cov}\left(s_y^2, s_{x_i}^2\right) = \frac{2}{n}S_y^2 S_{x_i}^2 \rho^2$$

y

$$\text{Cov}\left(s_{x_i}^2, s_{x_j}^2\right) = \frac{2}{n} S_{x_i}^2 S_{x_j}^2 \rho^2,$$

el sistema (5.3.3) se escribe

$$B_i S_{x_i}^4 + S_{x_i}^2 \rho^2 \sum_{\substack{j=1 \\ j \neq i}} B_j S_{x_j}^2 - S_y^2 S_{x_i}^2 \rho^2 = 0, \quad i = 1, \ldots, k$$

$$B_i S_{x_i}^2 = \rho^2 S_y^2 - \rho^2 \sum_{\substack{j=1 \\ j \neq i}} B_j S_{x_j}^2, \quad i = 1, \ldots, k.$$

Sumando

$$\sum_{i=1}^{k} B_i S_{x_i}^2 = k\rho^2 S_y^2 - \rho^2 (k-1) \sum_{i=1}^{k} B_i S_{x_i}^2,$$

de donde

$$\sum_{i=1}^{k} B_i S_{x_i}^2 = \frac{k\rho^2 S_y^2}{1 + \rho^2 (k-1)},$$

y sustituyendo

$$B_i S_{x_i}^2 = \rho^2 S_y^2 - \rho^2 \left(\sum_{i=1}^{k} B_i S_{x_i}^2 - B_i S_{x_i}^2 \right),$$

de donde

$$B_i S_{x_i}^2 \left(1 - \rho^2\right) = \rho^2 S_y^2 - \rho^2 \frac{k\rho^2 S_y^2}{1 + \rho^2 (k-1)} =$$

$$= \rho^2 S_y^2 \left(1 - \frac{k\rho^2}{1 + \rho^2 (k-1)} \right) = \rho^2 S_y^2 \frac{1 - \rho^2}{1 + \rho^2 (k-1)},$$

y finalmente

$$B_i S_{x_i}^2 = \frac{\rho^2 S_y^2}{1 + \rho^2 (k-1)}, \quad B_i = \frac{\rho^2}{1 + \rho^2 (k-1)} R_i.$$

La varianza mínima es, salvo términos de orden $O_p\left(n^{-\frac{3}{2}}\right)$,

$$V\left(\widehat{S}^2_{MR}\right) = \frac{1}{n-1}2\rho^2\left(1 - \frac{1}{1+(k+1)\rho^2}k\rho^4\right).$$

Aquí es donde se presenta uno de los inconvenientes del estimador múltiple de regresión propuesto por *Isaki*, puesto que para el cálculo del estimador óptimo es necesario conocer S_y^2, inconveniente que *Isaki* solventa definiendo el estimador aproximado $\widehat{S}^2_{MRE}$.

En la siguiente sección presentamos un nuevo estimador que no va a tener este inconveniente.

§5.4 Método de exponenciación.

5.4.1 Introducción.

En la presente sección estudiamos como se aplica el método de exponenciación en el caso de disponer de varias variables auxiliares para estimar la varianza poblacional. Proponemos un nuevo estimador de razón múltiple cuasi–insesgado, más preciso que el estimador de razón múltiple propuesto por Isaki e igual de preciso que el estimador de regresión múltiple estudiado en la sección anterior. Además este estimador salva uno de los grandes inconvenientes que presentaba el estimador de regresión múltiple pues para la determinación del estimador óptimo y su cálculo, bajo el modelo en el que desarrolla Isaki el estimador de regresión, pues no precisa tener conocimiento sobre S_y^2, como ocurre con el estimador de regresión, sino sólo sobre la correlación entre las variables disponibles, hecho que puede ser usual. Proponemos por tanto un estimador múltiple tipo razón, que tiene igual precisión que el estimador de regresión. Además para el modelo considerado por Isaki, este nuevo estimador es mucho más sencillo de computar que el de regresión.

5.4.2 Definición del estimador.

Sea una población finita $U = (U_1, \ldots, U_N)$. Sea y la variable principal y supongamos que se extrae una muestra aleatoria con reemplazo. Si asumimos que disponemos de información auxiliar de varias variables $x_1, \ldots, x_k$ (cuyas varianzas $S^2_{x_i}$ son conocidas, $i = 1, \ldots, k$), definimos el estimador de la varianza poblacional de y

$$\widehat{S}^2_{\alpha_1,\ldots,\alpha_k} = s_y^2 \left(\frac{S^2_{x_1}}{s^2_{x_1}}\right)^{\alpha_1} \cdots \left(\frac{S^2_{x_k}}{s^2_{x_k}}\right)^{\alpha_k},$$

donde $\alpha_1, \ldots, \alpha_k$ son números reales..

Se observa que:

- si $k = 1$, el estimador $\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}$ coincide con el estimador propuesto en el capítulo cuatro obtenido con el método repetido de sustitución,
- si $\alpha_1 = \cdots = \alpha_k = 0$ el estimador coincide con el estimador de expansión simple, s_y^2,
- si $\alpha_i = 1$ y $\alpha_j = 0$ $\forall i \neq j$ el estimador coincide con el estimador de razón univariante que se puede constuir a partir de la variable x_i,
- si $\alpha_i = -1$ y $\alpha_j = 0$ $\forall i \neq j$ el estimador coincide con el estimador de producto univariante que se puede constuir a partir de la variable x_i.

5.4.3 Propiedades.

Proposición 5.4.1 *El estimador $\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}$ es consistente.*

Demostración.
Obvia, en el sentido de consistencia en poblaciones finitas.

Proposición 5.4.2 *El estimador $\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}$ es sesgado y una aproximación de su sesgo la proporciona la expresión*

$$\text{sesgo}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right) \simeq \frac{S_y^2}{n}\left(\sum_{i=1}^{k} \alpha_i^2 \left(\beta_2(x_i) - 1\right) - \sum_{i=1}^{k} \alpha_i \left(\theta_{yx_i} - 1\right) + \right.$$

$$\left. + \sum_{i \neq j} \alpha_i \alpha_j \left(\theta_{x_i x_j} - 1\right)\right).$$

Demostración.-
Considerando las variables

$$e_0 = \frac{s_y^2 - S_y^2}{S_y^2} \quad ; \quad e_i = \frac{s_{x_i}^2 - S_{x_i}^2}{S_{x_i}^2}, \quad i = 1, \ldots, k,$$

expresamos el estimador $\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}$ de la forma

$$\widehat{S}^2_{\alpha_1,\ldots,\alpha_k} = S_y^2 (1 + e_0) \prod_{i=1}^{k} (1 + e_i)^{-\alpha_i} =$$

(siempre que $\mid e_i \mid < 1$, $\mid \alpha_i e_i \mid < 1 \quad i = 1, \ldots, k$)

$$= S_y^2 (1 + e_0) \prod_{i=1}^{k} \left(1 - \alpha_i e_i + \alpha_i^2 e_i^2 - \cdots\right) \simeq S_y^2 (1 + e_0) \prod_{i=1}^{k} \left(1 - \alpha_i e_i + \alpha_i^2 e_i^2\right) \simeq$$

(reteniendo sólo hasta el orden dos)

$$\simeq S_y^2 \left(1 + e_0 + \sum_{i=1}^{k} \alpha_i^2 e_i^2 - \sum_{i=1}^{k} \alpha_i e_i - \sum_{i=1}^{k} \alpha_i e_0 e_i + \sum_{i \neq j} \alpha_i \alpha_j e_i e_j\right).$$

Así,

$$\widehat{S}^2_{\alpha_1,\ldots,\alpha_k} - S_y^2 \simeq S_y^2 \left(e_0 + \sum_{i=1}^{k} \alpha_i^2 e_i^2 - \sum_{i=1}^{k} \alpha_i e_i - \sum_{i=1}^{k} \alpha_i e_0 e_i + \sum_{i \neq j} \alpha_i \alpha_j e_i e_j\right). \tag{5.4.1}$$

Tomando esperanzas

$$\text{sesgo}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right) \simeq S_y^2 \left(\sum_{i=1}^{k} \alpha_i^2 E\left(e_i^2\right) - \sum_{i=1}^{k} \alpha_i E\left(e_0 e_i\right) + \sum_{i \neq j} \alpha_i \alpha_j E\left(e_i e_j\right)\right).$$

Sustituyendo los valores:

$$E\left(e_i^2\right) = \frac{V(s_{x_i}^2)}{S_{x_i}^4} = \frac{1}{n}\left(\beta_2(x_i) - 1\right), \quad i = 1, \ldots, k,$$

$$E\left(e_0^2\right) = \frac{V(s_y^2)}{S_y^4} = \frac{1}{n}\left(\beta_2(y) - 1\right),$$

$$E\left(e_0 e_i\right) = \frac{\mathrm{Cov}(s_y^2, s_{x_i}^2)}{S_y^2 S_{x_i}^2} = \frac{1}{n}\left(\theta_{yxi} - 1\right), \quad i = 1, \ldots, k,$$

$$E\left(e_i e_j\right) = \frac{\mathrm{Cov}(s_{x_j}^2, s_{x_i}^2)}{S_{x_j}^2 S_{x_i}^2} = \frac{1}{n}\left(\theta_{x_j x i} - 1\right), \; i \neq j = 1, \ldots, k,$$

en la expresión de la aproximación del sesgo, nos lleva a la expresión

$$\text{sesgo}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right) \simeq \frac{S_y^2}{n}\left(\sum_{i=1}^{k} \alpha_i^2 \left(\beta_2(x_i) - 1\right) - \sum_{i=1}^{k} \alpha_i \left(\theta_{yx_i} - 1\right) + \right.$$

$$\left. + \sum_{i \neq j} \alpha_i \alpha_j \left(\theta_{x_i x_j} - 1\right)\right).$$

Su precisión la medirá por tanto su error cuadrático medio, siendo una expresión aproximada de este error la que proporciona la siguiente proposición:

Proposición 5.4.3 *Una aproximación del error cuadrático medio del estimador* $\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}$ *viene dada por la expresión:*

$$\text{ECM}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right) \simeq S_y^4 \frac{1}{n}\left[\beta_2(y) - 1 + \sum_{i=1}^{k} \alpha_i^2 \left(\beta_2(x_i) - 1\right) - 2\sum_{i=1}^{k} \alpha_i \left(\theta_{yx_i} - 1\right) + \right.$$

$$\left. + \sum_{i \neq j} \alpha_i \alpha_j \left(\theta_{x_i x_j} - 1\right)\right]. \tag{5.4.2}$$

Demostración.-
Partiendo de (5.4.1) obtenemos

$$\text{ECM}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right) = E\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k} - S_y^2\right)^2 \simeq$$

$$\simeq S_y^4 E\left(e_0^2 + \sum_{i=1}^{k} \alpha_i^2 e_i^2 - 2\sum_{i=1}^{k} \alpha_i e_0 e_i + \sum_{i \neq j} \alpha_i \alpha_j e_i e_j\right) =$$

$$= S_y^4 \left(E\left(e_0^2\right) + \sum_{i=1}^{k} \alpha_i^2 E\left(e_i^2\right) - 2\sum_{i=1}^{k} \alpha_i E\left(e_0 e_i\right) + \sum_{i \neq j} \alpha_i \alpha_j E\left(e_i e_j\right) \right).$$

Procediendo como en el cálculo aproximado del sesgo, tenemos

$$\mathrm{ECM}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right) \simeq S_y^4 \left[\frac{V\left(s_y^2\right)}{S_y^4} + \sum_{i=1}^{k} \alpha_i^2 \frac{V\left(s_{x_i}^2\right)}{S_{x_i}^4} - 2\sum_{i=1}^{k} \frac{\alpha_i \,\mathrm{Cov}\left(s_{x_i}^2, s_y^2\right)}{S_{x_i}^2 S_y^2} + \right.$$

$$\left. + \sum_{i \neq j} \alpha_i \alpha_j \frac{\mathrm{Cov}(s_{x_j}^2, s_{x_i}^2)}{S_{x_j}^2 S_{x_i}^2} \right] = \frac{S_y^4}{n} \left[\beta_2(y) - 1 + \sum_{i=1}^{k} \alpha_i^2 (\beta_2(x_i) - 1) - \right.$$

$$\left. -2\sum_{i=1}^{k} \alpha_i (\theta_{yx_i} - 1) + \sum_{i \neq j} \alpha_i \alpha_j (\theta_{x_i x_j} - 1) \right].$$

5.4.4 El estimador múltiple óptimo: $\widehat{S}^2_{\alpha_{opt}}$.

Nos planteamos el problema de seleccionar de entre todos los estimadores de la forma $\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}$ aquél que tenga mayor precisión, o lo que es igual elegir $\alpha_1, \ldots, \alpha_k$ de forma que minimicen el error cuadrático medio.

Proposición 5.4.4 *El valor de* $\alpha = \alpha_{(k\times 1)} = (\alpha_1, \ldots, \alpha_k)'$ *que minimiza el error cuadrático medio del estimador* $\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}$ *es*

$$\alpha_{opt} = A^{-1} A_0,$$

donde $A = A_{(k\times k)} = \left(a_{ij}\right)$ *con* $a_{ij} = \left(\theta_{x_i x_j} - 1\right)$ *y* $A_0 = A_{0\ (k\times 1)} = (a_{0_i})$ *con* $a_{0_i} = \left(\theta_{yx_i} - 1\right)$.

Además el error cuadrático medio mínimo aproximado es

$$\mathrm{ECM}\left(\widehat{S}^2_{\alpha_{opt}}\right) \simeq S_y^4 \frac{1}{n} \left[\beta_2(y) - 1 - A_0' A^{-1} A_0\right],$$

y el estimador múltiple $\widehat{S}^2_{\alpha_{opt}}$ *es cuasi-insesgado.*

Escribiendo el error cuadrático medio en forma matricial, con la notación de la proposición, y teniendo en cuenta que $\theta_{x_i x_i} = \beta_2(x_i)$,

$$\text{ECM}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right) \simeq S_y^4 \frac{1}{n}\left[\beta_2(y) - 1 + \alpha' A \alpha - 2\alpha' A_0\right].$$

Si A tiene inversa, el error cuadrático medio mínimo tiene por expresión aproximada

$$\text{ECM}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right) \simeq S_y^4 \frac{1}{n}\left[\beta_2(y) - 1 - A_0' A^{-1} A_0\right].$$

El valor de α que hace mínimo el error cuadrático medio es

$$\alpha_{opt} = A^{-1} A_0,$$

y para él la aproximación del sesgo del estimador es cero puesto que, reescribiendo la expresión aproximada del sesgo en forma matricial

$$\text{sesgo}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right) \simeq \frac{S_y^2}{n}\left(\sum_{i=1}^{k} \alpha_i^2 \left(\beta_2(x_i) - 1\right) - \sum_{i=1}^{k} \alpha_i \left(\theta_{y x_i} - 1\right) + \right.$$

$$\left. + \sum_{i \neq j} \alpha_i \alpha_j \left(\theta_{x_i x_j} - 1\right)\right) = S_y^2 \frac{1}{n}\left[\alpha' A \alpha - 2\alpha' A_0\right],$$

y para $\alpha = \alpha_{opt}$, se tiene

$$\text{sesgo}\left(\widehat{S}^2_{\alpha_{opt}}\right) \simeq S_y^4 \frac{1}{n}\left[\alpha_{opt}' A \alpha_{opt} - 2\alpha_{opt}' A_0\right] =$$

$$= S_y^4 \frac{1}{n}\left[A_0' A^{-1} A A^{-1} A_0 - 2 A_0' A^{-1} A_0\right] = 0.$$

Para $k = 1$ el valor óptimo de α es

$$\alpha_{opt} = (\theta - 1)/(\beta_2(x) - 1).$$

que como se puede ver en el capítulo cuatro proporciona un estimador igual de preciso que el estimador de regresión y cuasi–insesgado.

5.4.5 Comparación con el estimador de regresión múltiple.

Proposición 5.4.5 *El estimador múltiple* $\widehat{S}^2_{\alpha_{opt}}$ *y el estimador de regresión óptimo son igual de precisos, para muestras grandes.*

Demostración.-

El estimador de regresión múltiple propuesto por *Isaki* (1983) viene dado por

$$\widehat{S}^2_{MR} = s_y^2 + \sum_{i=1}^{k} B_i \left(S^2_{x_i} - s^2_{x_i} \right),$$

y su varianza por la expresión

$$V\left(\widehat{S}^2_{MR}\right) =$$

$$= V\left(s_y^2\right) + \sum_{i=1}^{k} B_i^2 V\left(s^2_{x_i}\right) - 2\sum_{i=1}^{k} B_i \operatorname{Cov}\left(s_y^2, s^2_{x_i}\right) + \sum_{i\neq j} B_i B_j \operatorname{Cov}\left(s^2_{x_i}, s^2_{x_j}\right). \tag{5.4.3}$$

Este estimador es insesgado y los coeficientes que proporcionan el estimador de regresión óptimo vienen dados por las soluciones del sistema:

$$B_i V\left(s^2_{x_i}\right) - \operatorname{Cov}\left(s_y^2, s^2_{x_i}\right) + \sum_{i\neq j} B_j \operatorname{Cov}\left(s^2_{x_i}, s^2_{x_j}\right) = 0, \quad i = 1, \ldots, k. \tag{5.4.4}$$

Consideremos ahora el estimador propuesto $\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}$. Obtuvimos

$$\operatorname{ECM}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right) \simeq \left[V\left(s_y^2\right) + \sum_{i=1}^{k} \alpha_i^2 V\left(s^2_{x_i}\right) \frac{S_y^4}{S^4_{x_i}} - 2\sum_{i=1}^{k} \alpha_i \operatorname{Cov}\left(s^2_{x_i}, s_y^2\right) \frac{S_y^2}{S^2_{x_i}} + \right.$$

$$\left. + \sum_{i\neq j} \alpha_i \alpha_j \frac{\operatorname{Cov}(s^2_{x_j}, s^2_{x_i})}{S^2_{x_j} S^2_{x_i}} S_y^4 \right].$$

Llamando $\gamma_i = \alpha_i R_i, \quad i = 1, \ldots, k$ con $R_i = \dfrac{S_y^2}{S^2_{x_i}}$, se tiene

$$\text{ECM}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right) \simeq \left[V\left(s_y^2\right) + \sum_{i=1}^{k} \gamma_i^2 V\left(s_{x_i}^2\right) - 2\sum_{i=1}^{k} \gamma_i \,\text{Cov}\left(s_{x_i}^2, s_y^2\right) + \right.$$

$$\left. + \sum_{i \neq j} \gamma_i \gamma_j \,\text{Cov}(s_{x_j}^2, s_{x_i}^2)\right].$$

Así pues los valores que minimizan ECM $\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right)$ son las soluciones del sistema

$$\gamma_i V\left(s_{x_i}^2\right) - \text{Cov}\left(s_y^2, s_{x_i}^2\right) + \sum_{i \neq j} \gamma_j \,\text{Cov}\left(s_{x_i}^2, s_{x_j}^2\right) = 0, \quad i = 1, \ldots, k,$$

que coincide con el sistema (5.4.4).

Los γ_i óptimos que minimizan ECM $\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right)$, γ_i^0, y los B_i óptimos que minimizan $V\left(\widehat{S}^2_{MR}\right)$, B_i^0, son iguales:

$$B_i^0 = \gamma_i^0 = \alpha_i^0 R_i, \quad i = 1, \ldots, k \tag{5.4.5}$$

De esta forma, de (5.4.3) y omitiendo los términos de orden $O_p\left(n^{-\frac{1}{2}}\right)$,

$$\text{mín } V\left(\widehat{S}^2_{MR}\right) = \text{mín MSE}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right),$$

y el estimador propuesto, $\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}$, es asintóticamente equivalente al estimador de regresión múltiple, $\widehat{S}^2_{MR}$.

5.4.6 El estimador $\widehat{S}^2_{\alpha_{opt}}$ bajo el modelo propuesto por Isaki.

Si la distribución de $(y, x_1, \ldots, x_k)$ tiene los mismos momentos hasta el orden cuatro de una normal multivariante,

$$\rho_{x_i x_j} = \rho_{y x_i} = \rho, \quad \forall i, j \tag{5.4.1}$$

y $-\frac{1}{k} < \rho < 1$, $V\left(\widehat{S}^2_{MR}\right)$ es mínima para

$$B_i^0 = \frac{\rho^2 R_i}{1 + (k-1)\rho^2} \quad i = 1, 2, \ldots, k.$$

De (5.4.5), ECM $\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right)$ es mínimo para

$$\alpha(k,\rho^2) = \alpha_1^0 = \cdots = \alpha_k^0 = \frac{\rho^2}{1+(k-1)\rho^2}.$$

Así, bajo el modelo (5.4.1) no se necesita conocer la varianza poblacional S_y^2 para calcular $\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}$ puesto que este estimador sólo depende de la correlación entre variables. En este caso

$$\widehat{S}^2_{\alpha_1,\ldots,\alpha_k} = s_y^2 \left(\frac{S_{x_1}^2}{s_{x_1}^2} \cdots \frac{S_{x_k}^2}{s_{x_k}^2}\right)^{\alpha(k,\rho^2)}.$$

Además, omitiendo términos de orden $O_p\left(n^{-\frac{1}{2}}\right)$, tenemos

$$\text{ECM}\left(\widehat{S}^2_{\alpha_1,\ldots,\alpha_k}\right) = V\left(\widehat{S}^2_{MR}\right) = \frac{2S_y^4}{n-1}\left[1 - \frac{k\rho^4}{1+(k-1)\rho^2}\right].$$

Es común que en la práctica se tenga alguna información acerca de la correlación entre las variables y en ese caso se puede calcular de forma sencilla el estimador propuesto y medir su precisión.

5.4.7 Ventajas.

El estimador propuesto presenta las siguientes mejoras:

- es cuasi–insesgado,
- resuelve el problema de la determinación del estimador óptimo en una población cualquiera, en contraposición con el estimador de regresión en el cual sólo se calcula el óptimo en el caso de que la población tenga los mismos momentos hasta el orden cuatro de una población normal multivariante,
- generaliza al estimador $\widehat{S}^2_\alpha$ obtenido por el método repetido de sustitución con cualquier variable auxiliar, al estimador de razón que se puede construir con cualquier variable auxiliar, al estimador de producto que se puede construir con cualquier variable auxiliar y al estimador de expansión simple,

- es asintóticamente igual de preciso que el estimador de regresión y por tanto asintóticamente más preciso que el estimador de razón múltiple y que el estimador directo,

- el estimador óptimo bajo el modelo en que Isaki estudia el estimador de regresión óptimo sólo precisa del conocimiento de la correlación entre las variables, conocimiento del que puede ser habitual disponer en la práctica, mejorando al estimador de regresión múltiple que precisa además del conocimiento de la varianza poblacional de y, S_y^2, para ser computado.

Bibliografía

[1] Ayachit, G. R. (1953), *Some aspects of large–scale sample surveys with particular reference to the ratio method of estimation*, M. Sc. Thesis, Bombay University, Bombay.

[2] Azorín, F. y Sánchez–Crespo, J. L. (1986), Métodos y Aplicaciones del Muestreo, *Alianza Universidad Textos*. Madrid.

[3] Barnet, V. (1982), Elements of Sampling Theory, *Hodder and Stoughton*, London.

[4] Cassel, C. M., Särndal, C. E. y Wretman, J. H. (1977). Foundations of Inference in Survey Sampling, *Wiley*, Nueva York.

[5] Chaubey, Y. P. (1991), *A study of ratio and product estimators under a super population model*, Commun. Statist. Theory and Meth. **20**(5 y 6), 1731–1746.

[6] Chaubey, Y. P., Dwivedi, T. D. y Singh, M. P. (1984), *An efficiency comparison of the product and ratio estimator*, Commun. Statist. Ser. A, **13**, 699–709.

[7] Chaudhuri, A. y Adhikari, A. K. (1989), *On efficiency of the ratio estimator*, Metrika **36**.

[8] Chaudhuri, A. y Vos, J. W. E. (1988), Unified theory and strategies of survey sampling, *North–Holland*, Amsterdam.

[9] Cochran, W. G. (1977), Sampling Techniques (3rd ed.), *John Wiley*, New York.

[10] Cochran, W. G. (1978), *Laplace' s ratio estimator.* In: H. A. David, ed., *Contributions to Survey Sampling and Applied Statistics*, Academic Press, New York.

[11] Cramér, H. (1946). Mathematical Methods of Statistics, *Princeton University Press.* Princeton.

[12] Dalabehera, M. y Sahoo, L. N. (1993), *Unbiased estimators using harmonic mean of the auxiliary variable*, Paper presented at the 47th Annual Conference of the Indian Society of Agricultural Statistics, Tirupati, India.

[13] Deming, W. E. (1950), Some Theory of Sampling, *Wiley*, New York.

[14] Deshpande, M. N. (1984), *A note on Rao, Hartley and Cochran's method*, J. Indian Soc. Agric. Statist., **36**, 114–116.

[15] Durbin, J. (1967), *Desing of Multi–stage Surveys for the Estimation of Sampling Error*, Applied Statistics **16**, 152–164.

[16] Durbin, J. (1953), *Some results in sampling theory when the units are selected with unequal probabilities*, Journal of the Royal Sta. Society, Series A, **15**, 262–269.

[17] Durbin, J. (1959), *A note on the application of Quenouille' s method of bias reduction to the estimator of ratio' s*, Biometrika **46**, 477–480.

[18] Fieller (1932), *The distribution of the index in a normal bivariate population*, Biometrika **24**.

[19] Fuller, W. A. (1970), *Sampling with Random Stratum Boundaries*, Journal of the Royal Statistical Society, ser B **32**, 203–226.

[20] Fuller, W. A. (1976). Introduction to Statistical Time Series, *John Wiley.* New York.

[21] Fuller, W. A. (1975), *Regression Analysis for Sample Survey*, Sankhyā, Series C **37**, 117–132.

[22] Goodman, L. A. and Hartley, H. O. (1958), *The precision of unbiased ratio–type estimators*, J. Amer. Statist. Assoc. **53**, 491–508.

[23] Grosbras (1987), Methodes Statistiques des Sondages, *Económica*, Paris.

[24] Gupta, P. C. y Adhvaryu, D. (1982), *On some unbiased product–type strategies*, J. Indian Soc. Agric. Statist., **34**, 48–54.

[25] Hájek, K. J. (1960), *Limiting distributions in simple random sampling from a finite population*, Pub. Math. Inst. Hungarian Acad. Sci., **5**.

[26] Hansen, H. O. y Hurwitz. W. N. (1943), *On the theory of sampling from finite populations*, Ann. Math. Statist. **14**.

[27] Hansen, H. O., Hurwitz, W. N. y Gurney, M. (1946), *Problems and methods of the sample surveys of business*, J. Amer. Statist. Assoc. **41**.

[28] Hansen, H. O., Hurwitz, W. N., y Madow, W. G. (1953), Sample Survey Methods and Theory, Vol 1 and 2 *John Wiley*. New York and London,

[29] Hartley, H. O., Rao, J. N. K., y Kiefer, G. (1969), *Variance Estimation with one Unit per Stratum*, Journal of the American Statistical Association **64**, 841–851.

[30] Hartley, H. O. y Ross, A. (1954), *Unbiased ratio estimators*, Nature **174**, 270–271.

[31] Hung, H. M. (1989). *On the utility of transformed auxiliary variables in finite population inference.* Commun. Statist. Theory and Meth. **18**(1), 171–188.

[32] Hwang y Tsay (1988), *On the asymtotic distribution of the ratio and regression estimators*, Sankhyā, Series B, **50**.

[33] Iachan, R., Singh, H. P. y Upadhyaya, L. N. (1987), *On unbiased product estimators*, Gujarat Statist. Rev., **14**, 32–50.

[34] Isaki, C. T. (1983), *Variance Estimation Using Auxiliary Information*, Journal of American Statistical Association **78**, n. 381, 117–123.

[35] Jones, H. L. (1965) *The jackknife method*, Proc. IBM Sci. Comp. Symp. Statist. 185–201.

[36] Kaur. P. (1985), *An efficient regression type estimator in a sample surveys*, Biometrical Journal **27**, 107–110.

[37] Kendall, M. y Stuart, A. (1977), The Advanced Theory of Statistics, Vol 1 y 2, *Griffin*, London.

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

[39] Kish, L., Namboodiri, N. K. y Pillai, R. K. (1962), *The ratio bias in surveys*, J. Amer. Statist. Assoc. **57**, 863–876..

[40] Krishnaiah, P.R. y Rao, C. R. (1988), Handbook of Statistics, Vol 6, North–Holland.

[41] Lahiri, D. B. (1951), *A method of sample selection providing unbiased ratio estimates*, Bull. Internat. Statist. Rev., **33**(2), 133–140.

[42] Menéndez. E. y Ferrales. J. (1989), *El estimador de razón generalizado*, Trabajos de Estadística **4**(1).

[43] Mickey, M. R. (1959), *Some finite population unbiased ratio and regression estimators*, J. Amer. Statist. Assoc., **54**, 594–612.

[44] Midzuno, H. (1952), *On the sampling system with probability proportional to the sum of the sizes*, Ann. Inst. Statist. Math., **3**, 99–108.

[45] Mukerjee, R. y Sengupta, S. (1989). *Optimal estimation of finite population total under a general correlated model*, Biometrika **76**, 789–794.

[46] Naik y Gupta (1991), *A general class of estimators for estimating population mean using auxiliary information*, Metrika **38**.

[47] Murthy, M. N. (1964), *Product method of estimation*, Sankhyā, A, **26**, 69–74.

[48] Murthy, M. N. y Nanjamma, N. S. (1959), *Almost unbiased ratio estimators based on interpenetrating sub sample estimates*, Sankhyā, **21**, 381–392.

[49] Nanjamma, N. S., Murthy, M. N. y Sethi, V. K. (1959), *Some sampling systems providing unbiased ratio estimators*, Sankhyā, **21**, 299–314.

[50] Nieto de Pascual, J., (1961), *Unbiased ratio estimators in stratified sampling*, J. Amer. Statist. Assoc. **56**.

[51] Ogus. J. L. y Clark, D. F. (1971), *The Annual Survey of Manufactures: A Report on Methodology*, U. S. Bureau of the Census Tecnical Paper **24**. U. S. Goverment Printing Office, Washington, DC.

[52] Olkin, I. (1958), *Multivariate ratio estimation for finite population*, Biometrika **45**, 154–165.

[53] Pandey, G. S. y Dubey (1989), *On almost unbiased ratio estimators*, Metron **47**, 1–4.

[54] Pandey, G. S. (1980), *Product–cum–power estimators*, Calcutta Statistical Assoc. Bull. **29**, 103–108.

[55] Pascual, J. N. (1961), *Unbiased ratio estimators in stratified sampling*, J. Amer. Statist. Assoc., **56**, 70–87.

[56] Prasad, B. (1986), *Some unbiased estimators versus mean per unit and ratio estimators in finite population sample surveys*, Commun. Statist. Theory and Meth. **15**(12), 3647–3657.

[57] Prasad, B. (1989), *Some improved ratio type estimators of population mean and ratio in finite population sample surveys*, Commun. Statist. Theory and Meth. **18**(1), 379–392.

[58] Prasad, B. y Singh, H. P. (1990), *Some improved ratio–type estimators of finite population variance in sample surveys*, Commun. Statist. Theory and Meth., **19**(3), 1127–1139.

[59] Prasad, B. y Singh, H. P. (1992), *Unbiased estimators of finite population variance using auxiliary information in sample surveys*, Commun. Statist. Theory and Meth., **21**(5), 1367–1376.

[60] Quenouille, M. H. (1956), *Notes on bias in estimation*, Biometrika **43**, 353–360.

[61] Raj, D. (1954), *Ratio estimation in sampling with equal and unequal probabilities*, J. Indian Soc. Agric. Statist., **6**, 127–138.

[62] Raj, D. (1968), Sampling Theory, *MacGraw–Hill*, New York

[63] Ramachandran, V. y Pillai, S. S. (1976), *Multivariate unbiased ratio–type estimation for finite sampling*, J. Indian Soc. Agric. Statist., **28**, 71–80.

[64] Rao, J. N. K. (1964), *Unbiased ratio and regression estimators in multi–stage sampling*, Journal of the Indian Society of Agricultural Statistics **14**, 175–188.

[65] Rao, J. N. K. (1967), *The precision of Mickey' s unbiased ratio estimator*, Biometrika **54**, 321–324.

[66] Rao, J. N. K. (1969), Ratio and regression estimators. New Developments in Survey Sampling, *N. L. Johnson and H. Smith, Jr. (eds), Wiley*, New York, 213–234.

[67] Rao, J. N. K. y Beagle, L. D. (1967), *A monte Carlo Study of some ratio estimators*, Sankhyā, series B **29**, 47–56.

[68] Rao, J. N. K. y Ramachandran, V. (1974), *Comparison of the separate and combined ratio estimates*, Sankhyā, Series C **36**, 151–156.

[69] Rao, J. N. K. y Webster, J. T. (1966), *On two methods of bias reduction in the estimation of ratios*, Biometrika **53**, 571–577.

[70] Rao, P.S.R.S. (1969), *Comparison of four ratio–type estimates under a model*, J. Amer. Statist. Assoc. **64**, 574–580.

[71] Rao, P.S.R.S. (1975a), *On the two–phase ratio estimator in finite population*, J. Amer. Statist. Assoc. **70**, 839–845.

[72] Rao, P.S.R.S. (1975b), *Hartley–Ross type estimator with two phase sampling*, Sankhyā Serie C, **37**, 140–146.

[73] Rao, P.S.R.S. (1981), *Efficiencies of the nine two–phase ratio estimators for the mean*, J. Amer. Statist. Assoc. **76**, 434–442.

[74] Rao, P. S. R. S. and Mudholkar, G. S. (1967), *Generalised multivariate estimators for the mean of finite populations*, J. Amer. Statist. Assoc. **62**, 1008–1012.

[75] Rao, T. J. (1965), *A note on estimation of ratios by Quenouille' s method*, Biometrika **52**, 647–649.

[76] Rao, T. J. (1967), Contributions to the theory of sampling strategies, *Ph. D. thesis, ISI.*, Calcuta.

[77] Rao, T. J. (1981a), *On a on unbiasedness in ratio estimation*, Journal of Statistical Planning and Inference **5**, 335–340.

[78] Rao, T. J. (1981b), *A note on unbiasedness in ratio estimation*, Journal of Statistical Planning and Inference **5**, 335–340.

[79] Rao, T. J. (1983), *A new class of unbiased product estimators*, Tech. Report **15**, Indian Statistical Institute, Calcutta.

[80] Rao, T. J. (1987), *On certain unbiased product estimators*, Commun. Statist. Theory Methods, **16**, 963–978.

[81] Rao, T. J. (1991), *On certain methods of improving ratio and regression estimators*, Commun. Statist. Theory and Meth. **20**(10), 3325–3340.

[82] Rao y Pereira (1967), *On double ratio estimators*, Sankhyā, Series A.

[83] Ray, S. K. and Sahai, A. (1980), *Efficient families of ratio and product–type estimators*, Biometrika **67**, 211–215.

[84] Ray, S. K. y Singh, R. K. (1981) *A product type estimator in double sampling*, Biom. Jour. **23**, 7.

[85] Reddy, V. N. (1974), *On a trasformed ratio method of estimation*, Sankhyā, Series C **36**, 59–70.

[86] Robson, D. S. (1957), *Application of multivariante polykeys to the theory of unbiased ratio–type estimation*, J. Amer. Statist. Assoc., **52**, 511–522.

[87] Royall, R. M. (1970), *On finite population sampling theory under certain linear regression models*, Biometrika **57**, 377–387.

[88] Royall, R. M. y Cumberland, W. G. (1981), *An Empirical Study of the Ratio Estimator and estimator of its variance*, J. Amer. Statist. Assoc. **76**(373), 66–77.

[89] Rueda, M., Ruiz, M. y Arcos, A. (1992), *Estimadores Condensados de razón*, Official Journal of the Chilean Statistical Society **9**, 15–27.

[90] Rueda, M. (1993), Aportaciones a la teoría de estimadores de razón, *Tesis Doctoral, Universidad de Granada.* Granada.

[91] Ruiz, M. (1991), *Comparación de estimadores óptimos de razón, producto y regresión*, Trabajos de Estadística vol. 6 **1**.

[92] Ruiz, M. y Santos, J. (1989), *Unbiased mean–of–the–ratio estimators*, Statistica **49**, 617–622.

[93] Ruiz, M. y Santos, J. (1990), *Sampling desing providing unbiased new product estimators*, Statistica **50**, 285–288.

[94] Sahoo, L. N. y Swain, A. K. P. C. (1980), *Unbiased ratio–cum–product estimator*, Sankhyā, C **42**, 56–62.

[95] Sahoo, L. N. (1983), *On a method of bias reduction in ratio estimation*, J. Statist. Res. **17**, 1–6.

[96] Sahoo, L.N. (1986), *On a class of unbiased estimators using multi–auxiliary information*, J. Indian Soc. Agric. Statist., **37**, 379–382.

[97] Sahoo, L. N. (1987), *On a class of almost unbiased estimators for population ratio*, Statistics, **18**, 119–121.

[98] Sahoo, L. N. y Mishra, G. (1989), *Classes of transformed ratio and product estimators in two stage sampling*, Statistica **49**, number 2, 295–298.

[99] Sahoo, L. N. y Swain, A. K. P. C. (1983), *Unbiased ratio–cum–product estimator using multi–auxiliary information*, Gujarat Statist. Rev., **10**, 11–16.

[100] Scott, A. J. y Wu, C. F. (1981), *On the asymtotic distribution of ratio and regression estimators*, J. Amer. Statist. Assoc. **76**, 98–102.

[101] Shah, D. N. y Shah, S. M. (1979), *Unbiased product–type estimators*, Gujarat Statist. Rev. **6**, 34–43.

[102] Shapiro, G. M. y Bateman, D. V. (1978), *A Better Alternative to the Collapse Stratum Variance Estimate*, Proceeding of the Social Statistics Section, American Statistical Association, 451–456.

[103] Singh, J., Pandey, B. N. e Hirano, K. (1973), *On the Utilization of a Known Coefficient of Kurtosis in the Estimation Procedure of Variance*, Ann. Inst. Statist. Math. **25**, 51–55.

[104] Singh, H. P. y Biradar, R. S. (1992), *Almost Unbiased Ratio–cum–product Estimators for the Finite Population Mean*, Test, Vol. 1 **1**, 19–29.

[105] Singh, H. P., Iachan, R. y Upadhaya, L. N. (1985), *Almost unbiased ratio and product estimators based on interpenetrating sub samples*, Commun. Statist. Theory Methods **14**, 963–978.

[106] Singh, M., Kumar, P. and Chandak, R. R. (1983), *Use of multi–auxiliary variables as a condensed auxiliary variable in selecting a sample*, Commun. Statist. Theory and Meth. **12**(14).

[107] Singh, M. P. (1965), *On the estimation of ratio and product of the population parameters*, Sankhyā, Series B, **27**, 321–328.

[108] Singh, M. P. (1966), *Efficient use of systematic sampling in ratio and product estimation*, Metrika **10**, 199–205.

[109] Singh, M. P. (1967a) *Ratio–cum–product method of estimation.* Metrika, 12, 34–42.

[110] Singh, M. P. (1967b), *Multivariate product method of estimation for finite population*, Journal of the Indian Society of Agricultural Statistics **109**, 1–10.

[111] Singh, P. y Srivastava, A. K. (1980), *Sampling schemes providing unbiased regression estimators*, Biometrika **67**, 205–209.

[112] Srivastava, S. K. (1965), *An estimate of the mean of a finite population using several auxiliary variables*, J. Indian Statist. Assoc. **3**, 189–194.

[113] Srivastava, S. K. (1967), *An Estimator Using Auxiliary Information in Sample Surveys*, Cal. Stat. Assoc. Bull. **16**, 121–132.

[114] Srivastava, S. K. (1980), *A class of estimators using auxiliary information in sample surveys*, Canad. J. Statist. **8**, 253–254.

[115] Srivastava, V. K. Shukla, N. D. y Bhatnagar, S. (1981), *Unbiased product estimators*, Metrika **28**, 191–196.

[116] Srivastava, S. K., Srivastava, A. k. y Khare, (1989), *Chain ratio type estimator for ratio of two population means using auxiliary characters*, Commun. Statist. Theory and Meth. **18**(10).

[117] Srivenkataramana, T. (1980), *A dual to ratio estimator in sample surveys*, Biometrika **67**, 1.

[118] Srivenkataramana, T. y Tracy, D. S. (1979), *On ratio and product methods of estimation in sampling*, Statist. Neerlandica, **33**, 37–49.

[119] Sukhatme, B. V. y David, I. P. (1973), *A Note on Koop' s Approach for Finding the Bias of the Ratio Estimate*, J. Amer. Statist. Assoc. **68**(342), 405–407.

[120] Sukhatme, B. V. y David, I. P. (1974), *On the Bias and Mean Square Error of the Ratio Estimator*, J. Amer. Statist. Assoc. **69**(346), 464–466.

[121] Sukhatme, P. V. and Sukhatme, B. V. (1989), *Improvement of ratio type estimators*, Biometrika **5**.

[122] Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. and Asok, C. (1984). Sampling Theory of Surveys with Applications, *Iowa State University Press*. Iowa.

[123] Swain, A. K. P. C. (1964), *The use of systematic sampling in ratio estimate*, Journal of the Indian Statistical Association **2**, 160–164.

[124] Tankou, V. y Dharmadhikari, S. (1989), *Improvement of ratio–type estimators*, Biometrika **5**, 793–800.

[125] Tin, M. (1965), *Comparison of some ratio estimators*, J. Amer. Statist. Assoc., **60**, 294–307.

[126] Vos, J. W. E. (1980), *Mixing of direct, ratio and product method estimators*, Statist. Neerlandica, **34**, 209-218.

[127] Williams, W. H. (1961), *Generating unbiased ratio and regression estimators*, Biometrics, **17**, 267–274.

[128] Williams, W. H. (1962), *On two methods of unbiased estimation with auxiliary variates*, J. Amer. Statist. Assoc. **57**, 184–186.

[129] Williams, W. H. (1963), *The precision of some unbiased regression estimators*, Biometrics, **19**, 352–361.