EUSKAL ESTATISTIKA ERAKUNDA
INSTITUTO VASCO DE ESTADÍSTICA

Cálculo de coeficientes de variación para diferentes estimadores directos e indirectos utilizados en las encuestas económicas de Eustat

Índice

1. Introducción

El objetivo de este documento es presentar algunos estimadores clásicos basados en el diseño y estimadores asistidos por modelos básicos que se utilizan en diferentes encuestas económicas de Eustat. Bajo la filosofía del diseño, la variable objeto de estudio (y en adelante) se trata como una cantidad fija y desconocida. La aleatoriedad viene dada por la forma de selección de la muestra que puede realizarse por diferentes procedimientos (muestreo aleatorio simple, muestreo aleatorio estratificado, etc.). El éxito de los métodos de estimación basados en el diseño radica en lo representativa que sea la muestra de la población objeto de estudio.

2. Estimadores directos

Se supone definida una población $U = (1, ..., N)$ que representa toda la Comunidad Autónoma de Euskadi y se definen estratos combinando, por ejemplo, Territorios Históricos y clasificación CNAE (Código Nacional de Actividades Económicas). En este documento se supone que en cada uno de los estratos se realiza un muestreo aleatorio simple. En este contexto, definiendo los estratos a priori y seleccionando la muestra de forma aleatoria dentro de cada uno de ellos, se pueden utilizar estimadores directos de expansión clásicos basados en el diseño dado que el tamaño muestral n_h en cada estrato es fijo.

En lo que sigue, y sin pérdida de generalidad, se supone que las unidades básicas de muestreo son establecimientos.

Sea,

- y_{hj}: valor que toma la variable y en el establecimiento j ($j = 1, ..., N_h$) del estrato h ($h = 1, ..., H$);
- Y_h: total poblacional (de y) en el estrato h;
- Y: total poblacional (de y) en la CAE;
- N_h: tamaño poblacional del estrato h;
- n_h: tamaño de la muestra en el estrato h;
- $\pi_{hj} = \frac{n_h}{N_h}$: probabilidad de inclusión de la unidad j en el estrato h;
- $\omega_{hj} = \frac{1}{\pi_{hj}}$: peso de muestreo de la unidad j en el estrato h;

- $\bar{y}_h = \frac{\sum_{j=1}^{n_h} y_{hj}}{n_h}$: media muestral en el estrato h;
- $s_{yh}^2 = \frac{\sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2}{n_h - 1}$: varianza muestral en el estrato h;
- $\hat{t}_{yh} = \frac{N_h}{n_h} \sum_{j=1}^{n_h} y_{hj} = N_h \bar{y}_h$: estimador del total en el estrato h;
- $\widehat{\text{var}}(\hat{t}_{yh}) = \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_{yh}^2}{n_h}$: estimador de la varianza del total en el estrato h;
- $\bar{y}_{str} = \frac{\hat{t}_{str.y}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \frac{\sum_{h=1}^{H} \sum_{j=1}^{n_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j=1}^{n_h} w_{hj}}$: media muestral estratificada;
- $\hat{t}_{str.y} = \sum_{h=1}^{H} \hat{t}_{yh} = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \sum_{j=1}^{n_h} w_{hj} y_{hj}$: estimador estratificado del total.

Dado que el muestreo se realiza de modo independiente en cada estrato, la varianza estratificada del total viene dada por

$$\text{var}(\hat{t}_{str.y}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_{yh}^2}{n_h},$$

el error estándar es

$$\text{e.e}(\hat{t}_{str.y}) = \sqrt{\sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_{yh}^2}{n_h}},$$

y el coeficiente de variación

$$\text{c.v}(\hat{t}_{str.y}) = \frac{\text{e.e}(\hat{t}_{str.y})}{\hat{t}_{str.y}}.$$

Propiedades: los estimadores $\bar{y}_{str}$ y $\hat{t}_{str}$ son incondicionalmente insesgados.

2.1. Estimador de Horvitz-Thompson

El estimador Horvitz-Thompson del total poblacional $Y_h = \sum_{j=1}^{N_h} y_{hj}$ de la variable y en el estrato h viene dado por

$$\hat{t}_{yh.\text{HT}} = \sum_{j=1}^{n_h} w_{hj} y_{hj},$$

donde $w_{hj} = 1/\pi_{hj}$ son los pesos muestrales de la j-ésima unidad en el estrato h y π_{hj} es su probabilidad de inclusión (o fracción de muestreo). En un muestreo aleatorio simple con n_h unidades seleccionadas del total N_h, con $w_{hj} = N_h/n_h$, $j = 1, ..., n_h$, un estimador insesgado de la varianza del estimador Horvitz-Thompson viene dadao por la siguiente expresión:

$$\widehat{\text{var}}(\hat{t}_{yh.\text{HT}}) = N_h^2\Big(1 - \frac{n_h}{N_h}\Big)\frac{1}{n_h}\widehat{\text{var}}(y_{hj}) = N_h^2\Big(1 - \frac{n_h}{N_h}\Big)\frac{1}{n_h}\frac{\sum_{j=1}^{n_h}(y_{hj} - \bar{y}_h)^2}{n_h - 1}.$$

El estimador de Horvitz-Thompson es un estimador directo y no hace uso de ningún tipo de información auxiliar, es decir, utiliza únicamente para su cálculo la información obtenida en la muestra y los pesos de muestreo. Cuando el tamaño muestral es pequeño no es un estimador adecuado aunque sea insesgado bajo el diseño ya que es un estimador muy inestable y su varianza puede ser muy grande en estos casos.

2.2. Estimador de regresión generalizado (GREG)

El estimador de regresión generalizado es un estimador que utiliza información auxiliar de la variable, por ejemplo, x para estimar la variable y. Se diferencia del estimador de regresión habitual en que introduce pesos en la estimación de los coeficientes del modelo (normalmente los pesos de muestreo). Este tipo de estimadores utilizan los modelos de regresión como un medio para conseguir estimadores consistentes desde el punto de vista del diseño. Requieren que el muestreo sea aleatorio. Han sido propuestos fundamentalmente por Särndal, Swensson y Wrettman (1989). El estimador de regresión generalizado del total Y_h en el estrato h viene dado por

$$\hat{t}_{yh.\text{GREG}} = \sum_{j=1}^{N_h}\hat{y}_{hj} + \sum_{j=1}^{n_h}\omega_{hj}(y_{hj} - \hat{y}_{hj}), \tag{1}$$

donde $\hat{y}_{hj}$, $j = 1, ..., N_h$ son los valores predichos por un modelo dado en el estrato h. El término $\sum_{j=1}^{n_h}\omega_{hj}(y_{hj}-\hat{y}_{hj})$ puede interpretarse como un ajuste de regresión dado el estimador proporcionado por el modelo. El efecto es que produce una importante reducción de su varianza, especialmente cuando la relación entre y y x es muy fuerte. Si el modelo elegido es un modelo de regresión lineal, $y_{hj} = x'_{hj}\beta_h + \epsilon_{hj}$, con $\text{var}(\epsilon_{hj}) = \sigma_h^2$ y $x_{hj} = (1, x_{hj1}, ..., x_{hjk})'$, entonces $\hat{y}_{hj} = x'_{hj}\hat{\beta}_{h.\text{GREG}}$, donde

$$\hat{\beta}_{h.\text{GREG}} = \left(\sum_{j=1}^{n_h}\omega_{hj}x'_{hj}x_{hj}c_{hj}\right)^{-1}\sum_{j=1}^{n_h}\omega_{hj}x'_{hj}y_{hj}c_{hj},$$

y c_{hj} son constantes especificadas. La expresión (1) puede escribirse también como:

$$\hat{t}_{yh.\mathrm{GREG}} = \hat{t}_{yh.\mathrm{HT}} + (X_h - \hat{t}_{xh.\mathrm{HT}})' \hat{\beta}_{h.\mathrm{GREG}},$$

donde $\hat{t}_{yh.\mathrm{HT}} = \sum_{j=1}^{n_h} \omega_{hj} y_{hj}$, es el estimador Horvitz-Thompson de Y_h, y $\hat{t}_{xh.\mathrm{HT}} = \sum_{j=1}^{n_h} \omega_{hj} x_{hj}$ es el estimador Horvitz-Thompson de X_h donde $X_h = \sum_{j=1}^{N_h} x_{hj}$. Efectivamente ambas expresiones coinciden ya que,

$$\begin{aligned} \hat{t}_{yh.\mathrm{GREG}} &= \textstyle\sum_{j=1}^{N_h} x'_{hj} \hat{\beta}_{h.\mathrm{GREG}} + \sum_{j=1}^{n_h} \omega_{hj} (y_{hj} - x'_{hj}) \hat{\beta}_{h.\mathrm{GREG}} \\ &= \hat{t}_{yh.\mathrm{HT}} + (X_h - \hat{t}_{xh.\mathrm{HT}})' \hat{\beta}_{h.\mathrm{GREG}}. \end{aligned}$$

También se puede expresar el estimador de regresión generalizado como una ponderación lineal sobre los y_j de modo que

$$\hat{t}_{yh.\mathrm{GREG}} = \sum_{j=1}^{n_h} \omega^*_{hj} y_{hj} = \sum_{j=1}^{n_h} \omega_{hj} g_{hj} y_{hj},$$

donde los pesos $\omega^*_{hj} = \omega_{hj} g_{hj}$ con $\omega_{hj} = 1/\pi_{hj}$,

$$g_{hj} = 1 + \Big(\sum_{j=1}^{N_h} x_{hj} - \sum_{j=1}^{n_h} \omega_{hj} x_{hj} \Big)' T_h^{-1} x_{hj} c_{hj} = 1 + (X_h - \hat{t}_{xh.\mathrm{HT}})' T_h^{-1} x_{hj} c_{hj},$$

y

$$T_h = \sum_{j=1}^{n_h} \omega_{hj} x_{hj} x'_{hj} c_{hj}.$$

El valor de g_{hj} está próximo a la unidad para la mayoría de los casos. Cuanto mayor es la muestra mayor proximidad debemos encontrar a la unidad. Es relativamente raro encontrar g_{hj} que sean mayores que 4 o menores que 0. Los pesos ω^*_j se llaman pesos calibrados ya que estos pesos aplicados a x_j reproducen exactamente la población total de x_j, es decir

$$\sum_{j=1}^{n_h} \omega^*_{hj} x_{hj} = \sum_{j=1}^{N_h} x_{hj} = X_h.$$

La varianza del estimador GREG viene dada por

$$\mathrm{var}(\hat{t}_{yh.\mathrm{GREG}}) = \sum_{j=1}^{N_h} \sum_{k=1}^{N_h} \Big(\frac{\omega_j \omega_k}{\omega_{jk}} - 1 \Big) \epsilon_j \epsilon_k,$$

donde $\epsilon_j = y_j - x_j'\beta_{h.\text{GREG}}$ y se estima mediante la expresión:

$$\hat{\text{var}}(\hat{t}_{yh.\text{GREG}}) = \sum_{j=1}^{n_h}\sum_{k=1}^{n_h}(\omega_{hj}\omega_{hj} - \omega_{hjk})(g_{hj}\hat{\epsilon}_j),(g_{kh}\hat{\epsilon}_j)$$

donde $\hat{\epsilon}_j = y_j - x_j'\hat{\beta}_{h.\text{GREG}}$. En el caso de muestreo aleatorio simple, esta expresión toma la forma:

$$\widehat{\text{var}}(\hat{t}_{yh.\text{GREG}}) = N_h^2(1 - \frac{n_h}{N_h})\frac{1}{n_h}\widehat{\text{var}}(g_h\hat{\epsilon}), \quad (2)$$

donde $g_h = (g_1, ..., g_{n_h})$ y $\hat{\epsilon} = (\hat{\epsilon}_1, ..., \hat{\epsilon}_{n_h})$.

2.2.1. Estimador directo de razón

Cuando hay una única variable auxiliar, la regresión pasa por el origen y el modelo de regresión es heterocedástico, con pesos $c_{hj} = x_{hj}$, el estimador GREG es un estimador directo de razón. Los valores g_{hj} arriba definidos son en este caso constantes para todas las observaciones $j = 1, ..., n_h$ y vienen dados por

$$g_h = 1 + \frac{X_h - \hat{t}_{x.\text{HT}}}{\hat{t}_{x.\text{HT}}},$$

donde $\hat{t}_{x.\text{HT}} = \sum_{j=1}^{n_h}\omega_{hj}x_{hj}$ es un número y no un vector. Además,

$$\hat{\beta}_{h.\text{D}} = \frac{\sum_{j=1}^{n_h}\omega_{hj}y_{hj}}{\sum_{j=1}^{n_h}\omega_{hj}x_{hj}},$$

y entonces $\hat{t}_{yh.\text{GREG}} = \hat{t}_{yh.\text{D}} = X'\hat{\beta}_{h.\text{D}} = (\sum_{j=1}^{N_h}x_{hj})'\hat{\beta}_{h.\text{D}}$. En estadísticas oficiales es frecuente expresar este estimador como

$$\hat{t}_{yh.\text{D}} = \frac{\sum_{j=1}^{N_h}x_{hj}}{\sum_{j=1}^{n_h}\omega_{hj}x_{hj}}\sum_{j=1}^{n_h}\omega_{hj}y_{hj} = \frac{X_h}{\hat{t}_{xh.\text{HT}}}\hat{t}_{yh.\text{HT}} = (\text{FE})\hat{t}_{yh.\text{HT}},$$

donde FE es el factor de elevación que no depende de la variable a estimar. Obsérvese que este factor de elevación coincide con los g_h del estimador GREG. Si el dominio fuera pequeño, de modo que haya pocas observaciones de la muestra que caen en ese dominio, el estimador es muy inestable. Se trata de un estimador directo que utiliza solamente información de su propio dominio. Su varianza es de orden $O(1/n_h)$, por tanto, bastante grande. Se obtiene como caso particular de la expresión (2), de la que se deduce

$$\widehat{\text{var}}(\hat{t}_{yh.\text{D}}) \approx N_h^2(1 - \frac{n_h}{N_h})\frac{1}{n_h}(\frac{X_h}{\hat{t}_{x.\text{HT}}})^2\widehat{\text{var}}_h(\hat{\epsilon}),$$

donde $\widehat{\text{var}}_h(\hat{\epsilon}) = \frac{\sum_{j=1}^{n_h}(\hat{\epsilon}_{hj}-\hat{\epsilon}_j)^2}{n_h-1}$ es la varianza muestral de los residuos del modelo $y_{hj} = \beta_h x_{hj} + \epsilon_{hj}$, desde $j = 1, ..., n_h$ con $\text{var}(\epsilon_{hj}) = \sigma^2 x_{hj}$. Es decir, los residuos se obtienen directamente al calcular $\hat{\epsilon}_{hj} = y_{hj} - \hat{y}_{hj} = y_{hj} - x'_{hj}\hat{\beta}_{h.\text{D}}$. Como el sesgo se considera prácticamente nulo, el error cuadrático medio de este estimador se aproxima por su varianza, es decir $\text{MSE}(\hat{t}_{yh.\text{D}}) \approx \text{var}(\hat{t}_{yh.\text{D}})$. En este caso el estimador de su coeficiente de variación se estima mediante la expresión

$$\widehat{\text{c.v.}}(\hat{t}_{yh.\text{D}}) = \frac{\widehat{\text{e.e.}}(\hat{t}_{yh.\text{D}})}{\hat{t}_{yh.\text{D}}},$$

donde $\widehat{\text{e.e.}}(\hat{t}_{yh.\text{D}}) = \sqrt{\widehat{\text{var}}(\hat{t}_{yh.\text{D}})}$.

3. Estimadores indirectos

Los estimadores indirectos utilizan para estimar el total o media de una variable en un dominio dado información de otros dominios.

3.1. Estimador indirecto de razón o estimador sintético

Supongamos que se utiliza información auxiliar procedente de otros dominios de forma que

$$\hat{\beta} = \frac{\sum_{h=1}^{H}\sum_{j=1}^{n_h}\omega_{hj}y_{hj}}{\sum_{h=1}^{H}\sum_{j=1}^{n_h}\omega_{hj}x_{hj}}.$$

Obsérvese que $\hat{\beta}$ utiliza la información muestral global, tanto para los valores de y como de x y es, por lo tanto, más estable. El estimador toma 'información prestada' del resto de dominios. La varianza de $\hat{\beta}$ es de orden $O(1/n)$, luego bastante menor. El estimador del total en el estrato h recibe el nombre de estimador sintético y viene dado por:

$$\hat{t}_{yh.\text{SYN}} = X_h\hat{\beta} = X_h\frac{\sum_{h=1}^{H}\sum_{j=1}^{n_h}\omega_{hj}y_{hj}}{\sum_{h=1}^{H}\sum_{j=1}^{n_h}\omega_{hj}x_{hj}}.$$

Este estimador puede ser considerablemente sesgado pero su varianza es pequeña cuando el tamaño de la muestra global $n = \sum_{h=1}^{H} n_h$ es grande. En particular este estimador puede ser insesgado cuando los $\hat{\beta}_{h.\text{D}}$ de los diferentes estratos son semejantes entre sí, esto es, son semejantes a las del dominio que las contiene. Bajo estas hipótesis, y no otras, es recomendable el uso del estimador sintético, ya que entonces es une stimador estable y puede llegar a ser prácticamente insesgado. Su varianza viene dada por

$$\widehat{\text{var}}(\hat{t}_{yh.\text{SYN}}) \approx N^2(1-\frac{n}{N})\frac{1}{n}(\frac{X_h}{\sum_{h=1}^{H}\sum_{j=1}^{n_h}\omega_{hj}x_{hj}})^2\widehat{\text{var}}(\epsilon),$$

donde $\widehat{\text{var}}(\epsilon)$ es la varianza muestral de los residuos del modelo heterocedástico $y_j = \beta x_j + \epsilon_j$ con $\text{var}(\epsilon) = \sigma^2 x_j$, $j = 1, ..., N$, a nivel poblacional (es decir, se calculan los residuos en toda la muestra, no sólo en el estrato de estudio).

Särndal and Hidiroglou (1989) proporcionan una aproximación del sesgo del estimador sintético según la cual $E(\hat{t}_{yh.\text{SYN}}) - t_{yh.\text{SYN}} \approx -\sum_{j=1}^{N}\epsilon_j$ donde $\epsilon_j = y_j - x_j'\hat{\beta}$. Luego el estimador será aproximadamente insesgado si se verifica que $\sum_{j=1}^{N}\epsilon_k = 0$. Esta condición no se satisface normalmente. Si el modelo no ajusta bien en el dominio de interés, la suma de residuales puede estar lejos de cero, indicando un sesgo considerable. En caso contrario, podemos esperar un sesgo limitado. Por ello, es desesable estimar el error cuadrático medio como medida de precisión del estimador. Viene dado por

$$\text{MSE}(\hat{t}_{yh.\text{SYN}}) = \text{var}(\hat{t}_{yh.\text{SYN}}) + (\text{sesgo}_{yh})^2,$$

y se estima mediante la expresión:

$$\widehat{\text{MSE}}(\hat{t}_{yh.\text{SYN}}) = \widehat{\text{var}}(\hat{t}_{yh.\text{SYN}}) + (\sum_{j=1}^{n_h}\hat{\epsilon}_j)^2,$$

donde $\hat{\epsilon}_j = y_j - x_j'\hat{\beta}$, $j = 1, ..., n$ son los residuos obtenidos a partir del modelo estimado con todos los datos muestrales, aunque en cada estrato solamente se suman los específicos de ese estrato.

El estimador de su coeficiente de variación viene dado por

$$\widehat{\text{c.v.}}(\hat{t}_{yh.\text{SYN}}) = \frac{\widehat{\text{rmse}}(\hat{t}_{yh.\text{SYN}})}{\hat{t}_{yh.\text{SYN}}},$$

donde, $\widehat{\text{rmse}}(\hat{t}_{yh.\text{SYN}}) = \sqrt{\widehat{\text{MSE}}(\hat{t}_{yh.\text{SYN}})}$.

4. Estimadores compuestos

El estimador compuesto se construye para compensar el sesgo del estimador indirecto, frente a la inestabilidad del estimador directo. Viene dado por

$$\hat{t}_{yh.\text{C}} = \phi_h \hat{t}_{yh.\text{D}} + (1-\phi_h)\hat{t}_{yh.\text{I}},$$

donde $0 \leq \phi_h \leq 1$, $\hat{t}_{yh.\mathrm{D}}$ es un estimador directo y $\hat{t}_{yh.\mathrm{I}}$ es un estimador indirecto. Muchos de los estimadores propuestos en la literatura bajo diseños o bajo modelos son estimadores compuestos. Entre los estimadores compuestos basados en el diseño podemos tomar por ejemplo $\hat{t}_{yh.\mathrm{D}}$ como estimador directo de razón y $\hat{t}_{yh.\mathrm{I}}$ como estimador indirecto de razón o estimador sintético.

Pfefferman (2002) propone el estimador compuesto:

$$\hat{t}_{yh.\mathrm{C}} = \phi_h \hat{t}_{yh.\mathrm{D}} + (1-\phi_h)\hat{t}_{yh.\mathrm{SYN}}, \qquad \phi_h = \frac{n_h}{N_h}. \tag{3}$$

Esta elección de ϕ_h es especialmente adecuada para poblaciones que no sean muy grandes, ya que en otro caso el cociente n_h/N_h no favorecería necesariamente al estimador directo cuando n_h crece. Con estos pesos, el peso del estimador directo o indirecto es mayor según sea su representación muestral. Esto es, a mayor fracción muestral, mayor contribución del estimador directo. Cuando la muestra está poco representada en la población, es el estimador indirecto el que tiene más peso en el estimador compuesto. Puede ocurrir también que $n_h = 1$ y $N_h = 1$ en cuyo caso el estimador compuesto sería igual al directo.

El error cuadrático medio del estimador compuesto (3) viene dado por

$$\begin{aligned}\mathrm{MSE}(\hat{t}_{yc.\mathrm{C}}) \quad \approx \quad & \phi_h^2 \mathrm{MSE}(\hat{t}_{yd.\mathrm{D}}) + (1-\phi_h)^2 \mathrm{MSE}(\hat{t}_{yd.\mathrm{SYN}}) \\ & +2\phi_h(1-\phi_h)E[(\hat{t}_{yh.\mathrm{D}} - Y_h)(\hat{t}_{yh.\mathrm{SYN}} - Y_h)].\end{aligned}$$

Su estimación no es fácil, ya que puede ocurrir que el tercer término de este sumatorio, es decir el de la covarianza no sea pequeño. Una posible aproximación viene dada por

$$\begin{aligned}& E[(\hat{t}_{yh.\mathrm{D}} - Y_h)(\hat{t}_{yh.\mathrm{SYN}} - Y_h)] = \\ & = E[\hat{t}_{yh.\mathrm{D}}\hat{t}_{yh.\mathrm{SYN}}] - Y_h E[\hat{t}_{yh.\mathrm{D}}] - Y_h E[\hat{t}_{yh.\mathrm{SYN}}] + Y_h^2 \\ & \approx E[\hat{t}_{yh.\mathrm{D}}\hat{t}_{yh.\mathrm{SYN}}] - Y_h^2 - Y_h(Y_h + \mathrm{sesgo}_{\mathrm{h.SYN}}) + Y_h^2 \\ & = E[\hat{t}_{yh.\mathrm{D}}\hat{t}_{yh.\mathrm{SYN}}] - Y_h^2 - Y_h(\mathrm{sesgo}_{\mathrm{h.SYN}}) \\ & \approx E[\hat{t}^2_{yh.\mathrm{D}}] - Y_h(\mathrm{sesgo}_{h.\mathrm{SYN}}) \\ & = \mathrm{MSE}(\hat{t}_{yh.\mathrm{D}}) - Y_h(\mathrm{sesgo}_{h.\mathrm{SYN}}).\end{aligned}$$

Obsérvese que $E[\hat{t}_{yh.\mathrm{D}}\hat{t}_{yh.\mathrm{SYN}}]$ puede aproximarse con $E[\hat{t}^2_{yh.\mathrm{D}}]$ ya que la covarianza es distinta de cero solamente para los términos comunes de ambos estimadores, es decir para los términos que intervienen en el cálculo del estimador directo.

Sustituyendo $\mathrm{MSE}(\hat{t}_{yh.\mathrm{D}})$ por su estimador, Y_h por su estimador sintético $\hat{t}_{yh.\mathrm{SYN}}$ y $\mathrm{sesgo}_{h.\mathrm{SYN}}$ por su estimador, se deriva el estimador del error cuadrático medio como

$$
\begin{aligned}
\widehat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{C}}) &\approx \phi_h^2\widehat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{D}}) + (1-\phi_h)^2\widehat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{SYN}}) \\
&+ 2\phi_h(1-\phi_h)[\widehat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{D}}) - \hat{t}_{yh.\mathrm{SYN}}(\mathrm{sesgo}_{h-\mathrm{SYN}})].
\end{aligned}
$$

El término $\widehat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{D}}) - \hat{t}_{yh.\mathrm{SYN}}(\mathrm{sesgo}_{h.\mathrm{SYN}})$ puede llegar a ser negativo, en cuyo caso se puede aproximarse por cero y entonces $\widehat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{C}}) \approx \phi_h^2\widehat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{D}})+(1-\phi_h)^2\widehat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{SYN}})$. Cuando $n = 0$ se toma $\widehat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{C}}) = \widehat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{SYN}})$ ya que entonces el estimador compuesto es igual al estimador sintético, independientemente de lo que valga el estimador directo, que se habrá obtenido por agregación a niveles superiores, ya que en la propia CNAE no hay muestra.

El estimador de su coeficiente de variación viene dado por

$$
\widehat{\mathrm{c.v}}(\hat{t}_{yh.\mathrm{C}}) = \frac{\widehat{\mathrm{rmse}}(\hat{t}_{yh.\mathrm{C}})}{\hat{t}_{yh.\mathrm{C}}},
$$

donde $\widehat{\mathrm{rmse}}(\hat{t}_{yh.\mathrm{C}}) = \sqrt{\widehat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{C}})}$.

Referencias

[1] Prasad, N.G.N and Rao, J.N.K. (1990). The Estimation of Mean Squared Error of Small Area Estimators. *Journal of the American Statistical Association*, **85**, 163-171.

[2] Pfeffermann, D. (2002). Snall Area Estimations-New Developments and Directions. *International Statistical Review*, **70**, 125-143.

[3] Rao, J.N.K. *Small Area Estimation*. Wiley, New York, 2003.

[4] Särndal, C.E., Hidirogloy, M.A. (1989). Small Domain Estimation: A Conditional Analysis. *Journal of the American Statistical Association*, **84**, 266-275.

[5] Särndal, C.E., Swensson, B. and Wretman, J.H. (1989). The Weighted Technique for Estimating the Variance of the General Regression Estimatorof the Finite Population Total. *Biometrica*, **76**, 3, 527-537.

[6] Särndal, C.E., Swensson, B. and Wretman, J.H. *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992.