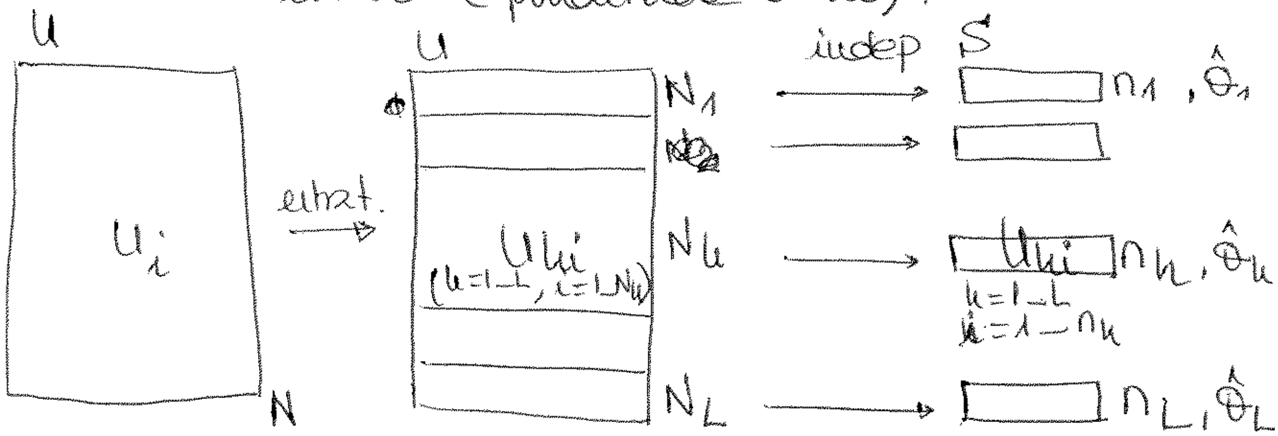
MUEST_T8. ESTIM. de RAZÓN en el m. estratif: SEPARADO Y COMBINADO.
SESGO, VARIANZA y sus ESTIMACIONES.
COMPARACIÓN de PRECISIONES.
REFERENCIA A LOS ESTIM. de REGRESIÓN.

∅ — INTRODUCCIÓN

Muestreo → Cjto de técnicas cuyo objetivo es obtener una muestra representativa de la población con el fin de estimar un parámetro poblac. desconocido cometiendo un error medible y acotable.

Estratificado → Muestreo que se utiliza cuando la variabilidad poblacional es heterogénea entre zonas (ó estratos), de modo que se divide la población en estratos (↑ HETEROG entre, ↑ HOMOG. DENTRO), y de cada estrato se obtiene, de manera indepdte sin necesidad de utilizar el mismo tipo de muestreo en todos los estratos, una muestra para estimar el parámetro poblacional de cada estrato.
El estimador global se obtiene como una suma extendida de los estim. en cada estrato (ponderado o no).



$$\hat\theta_{st} = \sum_{h=1}^{L} w_h \cdot \hat\theta_h \quad / \; w_h = \begin{cases} 1 & \text{para } \theta = X \\ \frac{N_h}{N} & \text{para } \theta = \bar{X} \end{cases}$$

Estim. razón → Mt. de estimación que utiliza, además de la característica a estudio, X, información de una variable auxiliar Y (*) (misma variable en fechas anteriores u otra variable altamente correlacionada con la var. a estudio).

Se utiliza para:

- estimar estructuras complejas:
$$R = \frac{X}{Y} = \frac{\overline{X}}{\overline{Y}} \longrightarrow \hat{R} = \frac{\overline{x}}{\overline{y}}$$
- obtener estim. más precisas de X utilizando la información de Y:
$$\hat{X}_R = \hat{R} \cdot Y \quad ; \quad \hat{\overline{X}} = \hat{R} \cdot \overline{Y}$$

(*) de la que se conocen sus valores poblacionales y muestrales en todos los estratos, $\overline{Y}_h$, $\overline{y}_h$ (para asegurar la estim. separada)

1_ESTIM. de RAZÓN en el MUESTR. ESTRATIFICADO

En el muestreo estratificado también pueden considerarse estimadores de la razón.

Estim. simple o separada:

Para cada estrato se estima la razón poblacional. El estim. de razón se obtiene como suma ponderada de los estim. en cada estrato.

$\forall h = 1 \ldots L$, $\hat{R}_h = \frac{\bar{x}_h}{\bar{y}_h}$ ← $R_h = \frac{\bar{X}_h}{\bar{Y}_h}$?

$\hat{R}_S = \sum_{h=1}^{L} \frac{Y_h}{Y} \hat{R}_h \cancel{= \sum_{h=1}^{L} \frac{N_h}{N_0} \hat{R}_h}$ ← $R = \frac{X}{Y} = \frac{\bar{X}}{\bar{Y}}$

Total poblacional, X → $\hat{X}_{Rh} = \frac{\bar{x}_h}{\bar{y}_h} \cdot Y_h = \hat{R}_h \cdot Y_h = \hat{R}_S \cdot Y$

(en m.e. $\hat{X}_{st} = \sum \hat{X}_h$) $\hat{X}_{RS} = \sum_{h=1}^{L} \hat{X}_{Rh} = \sum_{h=1}^{L} \hat{R}_h \cdot Y_h$ ✓

Media poblacional, $\bar{X}$ → $\hat{\bar{X}}_{Rh} = \frac{\bar{x}_h}{\bar{y}_h} \cdot \bar{Y}_h = \hat{R}_h \cdot \bar{Y}_h$

(en m.e. $\hat{\bar{X}}_{st} = \sum W_h \hat{\bar{X}}_h$) $\hat{\bar{X}}_{RS} = \sum_{h=1}^{L} W_h \cdot \hat{\bar{X}}_{Rh} = \sum_{h=1}^{L} \frac{\cancel{N_h}}{N} \cdot \hat{R}_h \cdot \frac{Y_h}{\cancel{N_h}} = \frac{1}{N} \sum \hat{R}_h \cdot Y_h$
$= \frac{1}{N} \sum_h \hat{X}_{Rh}$ ✓

Ventajas:

- Obtiene $\hat{R}_h$ para cada estrato
- Tiene en cuenta que R_h puede variar en cada estrato.

Inconvenientes:

- Requiere conocer Y_h en cada estrato
- Se acumulan los sesgos de los estratos para el sesgo total.

PREGUNTA: Si $N_h = \frac{Y_h}{Y}$ ≠ nº individ. estrato h, ¿se está aplicando afij. valor en el sentido de que la importancia del estrato no depende de su tamaño respecto al total sino de el valor del total de la var. auxiliar respecto al valor del total de Y

¡SOCORRO!

IMPOSIBLE

Estim. combinada:

Se realizan estim. para los parámetros poblacionales directamente, mediante razón de estimadores estratificados de la variable en estudio y de la variable auxiliar.

$$\hat{R}_c = \frac{\hat{X}_{st}}{\hat{Y}_{st}} = \frac{\overline{x}_{st}}{\overline{y}_{st}} = \frac{\sum W_h \cdot \overline{x}_h}{\sum W_h \cdot \overline{y}_h}$$

en estim. razón

Total Poblacional: $X \longrightarrow \hat{X}_{RC} = \hat{R}_c \cdot Y$

Media poblacional: $\overline{X} \longrightarrow \hat{\overline{X}}_{RC} = \hat{R}_c \cdot \overline{Y}$

Ventajas:

- No requiere conocer Y_h, basta conocer Y.
- No acumula el sesgo de los estratos

Inconvenientes:

- De forma implícita supone ~~de~~ constante la razón en todos los estratos.
- No permite disponer de información a subnivel de estratos.

En la práctica, se utiliza el estimador combinado cuando el tamaño de los estratos es pequeño ($L\uparrow \Rightarrow$ sesgo$\uparrow$).
En general, $\hat{R}_c$ se utiliza siempre que la estimación separada presenta demasiado sesgo ó cuando las razones en los estratos, R_h, son ctes.

2_ SESGOS, VARIANZAS y SUS ESTIMACIONES

2.1. SESGOS:

a) Estim. separada

- Para cada estrato, $\hat{R}_h$ es sesgado $\longrightarrow B(\hat{R}_h) = -\dfrac{\mathrm{cov}[\hat{R}_h, \bar{y}_h]}{\bar{Y}_h}$

Dem:

$$\mathrm{cov}(\hat{R}_h, \bar{y}_h) = E[\hat{R}_h \cdot \bar{y}_h] - E[\hat{R}_h]E[\bar{y}_h] =$$
$$= E\left[\frac{\bar{x}_h}{\bar{y}_h}\cdot \bar{y}_h\right] - (R_h + B(\hat{R}_h))\cdot \bar{Y}_h =$$
$$= \bar{X}_h - \frac{\bar{X}_h}{\bar{Y}_h}\cdot \bar{Y}_h + B(\hat{R}_h)\cdot \bar{Y}_h =$$
$$= 0 + B(\hat{R}_h)\cdot \bar{Y}_h$$

- El estim. separado de razón también es sesgado.

$$E[\hat{R}_S] = E\left[\sum_h \frac{Y_h}{Y}\cdot \hat{R}_h\right] = \sum \frac{Y_h}{Y}\underbrace{E[\hat{R}_h]}_{\neq R_h} \neq R.$$

$$B[\hat{R}_S] = E[\hat{R}_S] - R = \sum \frac{Y_h}{Y} E[\hat{R}_h] - R =$$
$$= \sum_{h=1}^{L} \frac{Y_h}{Y}\left(R_h + B(\hat{R}_h)\right) - R =$$
$$= \sum_{h=1}^{L} \frac{Y_h}{Y}\cdot\frac{X_h}{Y_h} + \sum_{h=1}^{L} \frac{Y_h}{Y} B(\hat{R}_h) - R =$$
$$= \underbrace{\sum_{h=1}^{L} X_h / Y}_{= R} + \sum \frac{Y_h}{Y} B(\hat{R}_h) - \cancel{R} \quad = \sum_{h=1}^{L} \frac{Y_h}{Y} B(\hat{R}_h)$$

- Para que el estimador sea insesgado, en cada estrato la variable en estudio y la var. auxiliar deben estar incorrelacionadas:

$$\mathrm{cov}(\hat{R}_h, \bar{y}_h) = 0 \Rightarrow \underset{h=1\ldots L}{E[\hat{R}_h] = R_h} \Rightarrow E[\hat{R}_S] = R$$

- Para cada estrato, el valor del sesgo puede acotarse:

$$B(\hat{R}_h) = -\frac{cov(\hat{R}_h, \bar{y}_h)}{\bar{Y}_h} = -\rho_{\hat{R}_h, \bar{y}_h} \cdot \sigma_{\hat{R}_h} \cdot \sigma_{\bar{y}_h} / \bar{Y}_h \leq \sigma_{\hat{R}_h} \cdot \frac{\sigma_{\bar{y}_h}}{\bar{Y}_h}$$

$$\left| \frac{B(\hat{R}_h)}{\sigma_{\hat{R}_h}} \right| \leq C_{\bar{y}_h} \implies B(\hat{R}_S) \leq \sqrt{L} \cdot C_{\bar{y}_h} / \text{máx}$$

- suponiendo que todos los sesgos tienen el mismo signo, para que no se compensen, se puede acotar el valor del sesgo del estimador de razón separado:

$$B(\hat{R}_S) = \sum N_h B(\hat{R}_h) \implies B(\hat{R}_S) \leq L \cdot B(\hat{R}_h)$$

En mi opinión:

$$\left. \begin{array}{l} B(\hat{R}_S) \leq L \cdot B(\hat{R}_h) \quad / \hat{R}_h = \text{máx}\{\hat{R}_h\} \\ \sigma_{\hat{R}_S} \leq \sqrt{L} \cdot \sigma_{\hat{R}_h} \quad / \sigma_{\hat{R}_h} = \text{máx}\{\hat{R}_h\} \end{array} \right| \implies \left| \frac{B(\hat{R}_S)}{\sigma_{\hat{R}_S}} \right| \leq \sqrt{L} \cdot C_{\bar{y}_h}$$

(PENSAR)

Por tanto, si el nº de estratos es muy grande, el sesgo también puede serlo y su contribución al ECM puede ser importante. En la práctica, el sesgo suele ser ~~menor~~ mucho menor que la cota superior.

Sin embargo, para muestras de los estratos sufic. grandes el sesgo pasa a ser despreciable.

- En la práctica, la expresión del sesgo no se puede obtener, por lo que se utiliza una aproximación:

Para cada estrato,

$$B(\hat{R}_h) \simeq \frac{R_h V(\bar{y}_h) - Cov(\bar{x}_h, \bar{y}_h)}{\bar{Y}_h^2} = R_h (C^2_{\bar{y}_h} - C_{\bar{x}_h \bar{y}_h})$$

que se puede estimar con:

$$\widehat{B(\hat{R}_h)} \simeq \hat{R}_h (\hat{C}^2_{\bar{y}_h} - \hat{C}_{\bar{x}_h \bar{y}_h})$$

Para $\hat{R}_S$, se sustituye directamente en la expresión

$$B(\hat{R}_S) = \sum_{h=1}^{L} \frac{Y_h}{Y} B(\hat{R}_h).$$

- Si utilizamos $\hat{R}_S$ para estimar el total poblacional:

$$B(\hat{X}_{RS}) = B(\hat{R}_S \cdot Y) = Y B(\hat{R}_S) = Y \sum N_u B(\hat{R}_u) = Y \sum \frac{Y_u}{Y} B(\hat{R}_u) =$$
$$= \sum Y_u B(\hat{R}_u).$$

$\rightarrow$ insesgado si $\hat{R}_u$ lo es

Aproximación del sesgo:

$$(SR) \rightarrow B(\hat{X}_{RS}) \simeq \sum Y_u \cdot \frac{1-f_u}{n_u} \cdot \frac{(R_u S^2_{y_u} - S_{x_u y_u})}{\overline{Y_u}^2}$$

$$(CR) \rightarrow B(\hat{X}_{RS}) \simeq \sum Y_u \cdot \frac{1}{n_u} \cdot \frac{(R_u \sigma^2_{y_u} - \sigma_{x_u y_u})}{\overline{Y_u}^2}$$

que puede estimarse por:

$$(SR) \rightarrow \hat{B}(\hat{X}_{RS}) \simeq \sum Y_u \cdot \frac{(1-f_u)}{n_u} \frac{(\hat{R}_u \hat{S}^2_{y_u} - \hat{S}_{x_u y_u})}{\overline{y_u}^2}$$

$$(CR) \rightarrow \hat{B}(\hat{X}_{RS}) \simeq \sum Y_u \frac{1}{n_u} \frac{(\hat{R}_u \hat{S}^2_{y_u} - \hat{S}_{x_u y_u})}{\overline{y_u}^2}$$

- Si utilizamos $\hat{R}_S$ para estimar la media poblacional:

$$\hat{\overline{X}}_{RS} = \sum W_u \hat{\overline{X}}_{Ru} = \ldots = \frac{1}{N} \hat{X}_{RS}$$

$$B(\hat{\overline{X}}_{RS}) = \frac{1}{N} B(\hat{X}_{RS})$$

$$\hat{B}(\hat{\overline{X}}_{RS}) \simeq \frac{1}{N} \text{aprox } B(\hat{X}_{RS})$$

etc.

b) Estimador combinado para el total :

El estimador de razón combinado es en general sesgado, y en consecuencia tb. lo es $\hat{X}_{RC}$:

$$\mathrm{cov}(\hat{R}_c, \bar{y}_{st}) = E[\hat{R}_c \cdot \bar{y}_{st}] - E[\hat{R}_c]E[\bar{y}_{st}] =$$
$$= E\left[\frac{\bar{x}_{st}}{\cancel{\bar{y}_{st}}} \cdot \cancel{\bar{y}_{st}}\right] - E[\hat{R}_c] \cdot \bar{Y} =$$
$$= \bar{X} - E[\hat{R}_c]\bar{Y} = \bar{Y}\left[R - E[\hat{R}_c]\right] \Rightarrow$$

$$B(\hat{R}_c) = -\frac{\mathrm{cov}(\hat{R}_c, \bar{y}_{st})}{\bar{Y}}$$

$$\hat{X}_{RC} = \hat{R}_c \cdot Y$$

$$E[\hat{X}_{RC}] = E[\hat{R}_c \cdot Y] = Y \cdot E[\hat{R}_c]$$

luego

$$B(\hat{X}_{RC}) = E[\hat{X}_{RC}] - X = Y E[\hat{R}_c] - X = Y\left[E[\hat{R}_c] - R\right] =$$
$$= Y \cdot B(\hat{R}_c) = -N \cdot \mathrm{cov}(\hat{R}_c, \bar{y}_{st}) .$$

que puede acotarse ;

$$\mathrm{cov}(\hat{R}_c, \bar{y}_{st}) = \rho_{\hat{R}_c, \bar{y}_{st}} \cdot \sigma_{\hat{R}_c} \cdot \sigma_{\bar{y}_{st}} = \bar{Y}(R - E[\hat{R}_c]) \Rightarrow$$

$$\frac{|B(\hat{R}_c)|}{\sigma_{\hat{R}_c}} = |\rho_{\hat{R}_c, \bar{y}_{st}}| \cdot \underbrace{\frac{\sigma_{\bar{y}_{st}}}{\bar{Y}}}_{C_{\bar{y}_{st}}} \leq C_{\bar{y}_{st}}$$

El estimador combinado no acumula los sesgos de los estratos (si son del mismo signo), por lo que se recomienda el estim. combinado.

En caso contrario, si los sesgos de los estratos tienen ≠ signo, puede ocurrir que el estimador separado sea preferible al estimador combinado.

MIRAR ATRÁS ⇒

Expresión del sesgo del estimador combinado para los diferentes tipos de muestreo:

$$B(\hat{X}_{RC}) = B(\hat{R}_c)\cdot Y$$

$$SR \longrightarrow B(\hat{X}_{RC}) = B(\hat{R}_c)\cdot Y = -N\cdot \mathrm{cov}(\hat{R}_c, \bar{y}_{st}) \overset{\text{aprox.}}{\cong}$$

$$= N^2 \sum_{h=1}^{L} \frac{W_h^2}{Y} \frac{(1-f_h)}{n_h} (RS^2_{y_h} - S_{xy_h})$$

que puede estimarse como:

$$\hat{B}(\hat{X}_{RC}) = N^2 \sum_h \frac{W_h^2}{Y} \frac{(1-f_h)}{n_h} (\hat{R}\hat{S}^2_{y_h} - \hat{S}_{xy_h})$$

$$CR \longrightarrow B(\hat{X}_{RC}) = N^2 \sum_{h=1}^{L} \frac{W_h^2}{Y} \cdot \frac{1}{n_h} (R\sigma^2_{y_h} - \sigma_{xy_h})$$

que puede estimarse como:

$$\hat{B}(\hat{X}_{RC}) = N^2 \sum_{h=1}^{L} \frac{W_h^2}{Y} \cdot \frac{1}{n_h} (\hat{R}\hat{\sigma}^2_{y_h} - \hat{\sigma}_{xy_h})$$

Para la media poblacional:

$$B(\hat{X}_{RC}) = B(\hat{R}_c)\,\bar{Y}$$

$$SR \longrightarrow B(\hat{X}_{RC}) = \sum_{h=1}^{L} \frac{W_h^2}{\bar{Y}} \frac{(1-f_h)}{n_h} (RS^2_{y_h} - S_{xy_h})$$

que puede estimarse como:

$$\hat{B}(\hat{X}_{RC}) = \sum_{h=1}^{L} \frac{W_h^2}{\bar{Y}} \cdot \frac{(1-f_h)}{n_h} (\hat{R}S^2_{y_h} - \hat{S}_{xy_h})$$

$$CR \longrightarrow B(\hat{X}_{RC}) = \sum_{h=1}^{L} \frac{W_h^2}{\bar{Y}} \cdot \frac{1}{n_h} (R\sigma^2_{y_h} - \sigma_{xy_h})$$

$$\hat{B}(\hat{X}_{RC}) = \sum_{h=1}^{L} \frac{W_h^2}{\bar{Y}} \cdot \frac{1}{n_h} (\hat{R}\cdot\hat{\sigma}^2_{y_h} - \hat{\sigma}_{xy_h})$$

VARIANZAS:

a) Estimador separado del total:

$$\hat{X}_{RS} = \sum_{h=1}^{L} \hat{R}_h Y_h$$

muestreo indep. en los estratos

n_h sufic. grande $\hat{R}_h$ insesgado de R_h

$$V[\hat{X}_{RS}] = \sum_{h=1}^{L} V(\hat{R}_h \cdot Y_h) = \sum_{h=1}^{L} Y_h^2 V(\hat{R}_h) \simeq \sum_{h=1}^{L} \frac{Y_h^2 \cdot R_h^2}{\cancel{X_h^2}} (C^2_{\bar{x}_h} + C^2_{\bar{y}_h} - 2C_{\bar{x}_h\bar{y}_h})$$

$$(SR) \rightarrow V[\hat{X}_{RS}] \simeq \sum_{h=1}^{L} Y_h^2 \cdot \frac{1}{\bar{Y}_h^2} \cdot \frac{1-f_h}{n_h} \left(S^2_{x_h} + R_h^2 S^2_{y_h} - 2R_h S_{xy_h}\right) =$$

$$= \sum_{h=1}^{L} \cancel{Y_h^2} \cdot \frac{N_h^2}{\cancel{Y_h^2}} \cdot \frac{1-f_h}{n_h} \left(S^2_{x_h} + R_h^2 S^2_{y_h} - 2R_h S_{xy_h}\right) =$$

$$= \sum_{h=1}^{L} N_h^2 \frac{(1-f_h)}{n_h} \left[S^2_{x_h} + R_h^2 S^2_{y_h} - 2R_h S_{xy_h}\right].$$

y se puede estimar por:

$$\hat{V}[\hat{X}_{RS}] = \sum_{h=1}^{L} N_h^2 \cdot \frac{(1-f_h)}{n_h} \left[\hat{S}^2_{x_h} + \hat{R}_h^2 \hat{S}^2_{y_h} - 2\hat{R}_h \hat{S}_{xy_h}\right]$$

$$(CR) \rightarrow V[\hat{X}_{RS}] = \ldots \simeq \sum_{h=1}^{L} N_h^2 \cdot \frac{1}{n_h} \left[\sigma^2_{x_h} + R_h^2 \sigma^2_{y_h} - 2R_h \sigma_{xy_h}\right]$$

$$\hat{V}[\hat{X}_{RS}] = \sum_{h=1}^{L} N_h^2 \cdot \frac{1}{n_h} \left[\hat{\sigma}^2_{x_h} + \hat{R}_h^2 \hat{\sigma}^2_{y_h} - 2\hat{R}_h \hat{\sigma}_{xy_h}\right]$$

Para la media poblacional:

$$V[\hat{\bar{X}}_{RS}] = V\left[\frac{\hat{X}_{RS}}{N}\right] = \frac{1}{N^2} V[\hat{X}_{RS}]$$

$$(SR) \rightarrow \sum W_h^2 \cdot \frac{1-f_h}{n_h} \left(S^2_{x_h} + R_h^2 S^2_{y_h} - 2R S_{xy_h}\right)$$

$$(CR) \rightarrow \sum W_h^2 \cdot \frac{1}{n_h} (\qquad)$$

estimación → poner ^

MIRAR ATRÁS

b) Estimador ~~combinado~~ del total:

$$\hat{X}_{RC} = \hat{R}_c Y \quad \text{con} \quad \hat{R}_c = \frac{\bar{x}_{st}}{\bar{y}_{st}}$$

Para un tamaño muestral sufic. grande $\hat{R}_c$ es ~~insesgado~~ aproximadamente insesgado de R, con:

$$V[\hat{R}_c] = V\left(\frac{\bar{x}_{st}}{\bar{y}_{st}}\right) = V\left(\frac{\bar{x}_{st}}{\bar{y}_{st}} - R\right) \sim V\left(\frac{\bar{x}_{st} - R\bar{y}_{st}}{\bar{Y}}\right) =$$

$$= \frac{1}{\bar{Y}^2}\left[V[\bar{x}_{st}] + R^2 V[\bar{y}_{st}] - 2R\,COV[\bar{x}_{st}, \bar{y}_{st}]\right]$$

$$SR \to \frac{1}{\bar{Y}^2}\left[\sum_{h=1}^{L} W_h^2 (1-f_h)\frac{S^2_{xh}}{n_h} + R^2 \sum_{h=1}^{L} W_h^2 \frac{(1-f_h)}{n_h} S^2_{y_h} - 2R\sum_{i=1}^{L} W_h^2 \frac{(1-f_h)}{n_h} S_{xy_h}\right]$$

$$= \frac{1}{\bar{Y}^2}\sum_{h=1}^{L} W_h^2 \frac{(1-f_h)}{n_h}\left[S^2_{x_h} + R^2 \cdot S^2_{y_h} - 2R\,S_{xy_h}\right]$$

$$CR \to V[\hat{R}_c] = \frac{1}{\bar{Y}^2}\sum W_h^2 \frac{1}{n_h}\left[\sigma^2_{x_h} + R^2\sigma^2_{y_h} - 2R\,\sigma_{xy_h}\right].$$

por lo que

$$V[\hat{X}_{RC}] = V[\hat{R}_c Y] = Y^2 V[\hat{R}_c]$$

, y para tamaños muestrales sufic. grandes

$$V[\hat{X}_{RC}] \sim \frac{Y^2}{\bar{Y}^2}\left[V[\bar{x}_{st}] + R^2 V[\bar{y}_{st}] - 2R\,COV[\bar{x}_{st}, \bar{y}_{st}]\right] =$$

$$= N^2 [\qquad\qquad]$$

que se puede detallar para cada tipo de muestreo:

$$SR \to V[\hat{X}_{RC}] = \sum_{h=1}^{L} N_h^2 \cdot \frac{(1-f_h)}{n_h}\left[S^2_{x_h} + R^2 S^2_{y_h} - 2R S_{xy_h}\right]$$

$$\hat{V}[\hat{X}_{RC}] = \sum N_h^2 \frac{(1-f_h)}{n_h}\left[\hat{S}^2_{x_h} + \hat{R}^2 \hat{S}^2_{y_h} - 2\hat{R}\hat{S}_{xy_h}\right].$$

$$CR \rightarrow V[\hat{X}_{RC}] = \sum_{h=1}^{L} N_h^2 \cdot \frac{1}{n_h} [\sigma_{x_h}^2 + R^2 \sigma_{y_h}^2 - 2R\,\sigma_{xy_h}]$$

$$\hat{V}[\hat{X}_{RC}] = \sum_{h=1}^{L} N_h^2 \cdot \frac{1}{n_h} [\hat{S}_{x_h}^2 + \hat{R}^2 \hat{S}_{y_h}^2 - 2\hat{R}\,\hat{S}_{xy_h}]$$

Para la media poblacional:

$$\text{Ahora } \hat{\bar{X}}_{RC} = \hat{R}_c \cdot \bar{Y} = \frac{\hat{X}_{RC}}{N}$$

$$V[\hat{\bar{X}}_{RC}] = \frac{1}{N^2} V[\hat{X}_{RC}]$$

$$SR \rightarrow V[\hat{\bar{X}}_{RC}] = \sum_{h=1}^{L} W_h^2 \cdot \frac{(1-f_h)}{n_h} (S_{x_h}^2 + R^2 S_{xy_h}^2 - 2R\,S_{xy_h})$$

$$\hat{V}[\hat{\bar{X}}_{RC}] = \sum_{h=1}^{L} W_h^2 \cdot \frac{(1-f_h)}{n_h} (\hat{S}_{x_h}^2 + \hat{R}^2 \hat{S}_{xy_h}^2 - 2\hat{R}\,\hat{S}_{xy_h})$$

$$CR \rightarrow V[\hat{\bar{X}}_{RC}] = \sum_{h=1}^{L} W_h^2 \cdot \frac{1}{n_h} (\sigma_{x_h}^2 + R^2 \sigma_{y_h}^2 - 2R\,\sigma_{xy_h})$$

$$\hat{V}[\hat{\bar{X}}_{RC}] = \sum_{h=1}^{L} W_h^2 \cdot \frac{1}{n_h} (\hat{S}_{x_h}^2 + \hat{R}^2 \hat{S}_{y_h}^2 - 2\hat{R}\,\hat{S}_{xy_h}).$$

Nótese que las varianzas aproximadas de los estimadores separado y combinado solamente difieren en las razones poblacionales que aparecen:

$\hat{X}_{RS} \rightarrow R_h$ (de cada estrato)

$\hat{X}_{RC} \rightarrow R$ (total)

por lo que los estimadores serán igual de precisos cuando las razones poblac. en todos los estratos coincidan, y a su vez coincidan con la razón total.

3 – COMPARACIÓN de PRECISIONES

La diferencia de las expresiones* generales de las varianzas de los estimadores de razón es: (SR)

$$V(\hat{X}_{RC}) - V(\hat{X}_{RS}) = \sum_{h=1}^{L} N_h^2 \cdot \frac{(1-f_h)}{n_h} \left[(R^2 - R_h^2) \cdot S_{y_h}^2 - 2(R-R_h) S_{xy_h} \right]$$

La interpretación no es sencilla, pero se puede hacer ~~en~~ atendiendo a dos elementos:

- el tamaño muestral en cada estrato, n_h.
- la variabilidad de las razones ~~de~~ entre los estratos, R_h.

- En tamaños muestrales de los estratos pequeños, no se puede admitir la validez de la aproximación de la varianza, por lo que se recomienda el estimador combinado, que no acumula sesgo.

- Si las razones de los estratos coinciden, $R_h = R$, también se recomienda el estimador combinado, porque al ser las varianzas iguales se elige el más sencillo (requiere menos información).

- Si R_h varía entre los estratos, se recomienda el uso del estimador separado. Será más preferible cuanto más próxima a una recta que pase por el origen esté la relación entre X e Y en cada estrato.

- Cuando se trabaje con muestras grandes en todos los estratos tb es preferible el estimador separado.

4_ REFERENCIA A LOS ESTIMADORES DE REGRESIÓN EN EL MUESTREO ESTRATIFICADO

Al igual que ocurre con los estimadores de razón, en el muestreo estratificado existen dos métodos para construir el estimador de regresión:

a) Estimador separado de regresión:

En cada estrato, se construye un estimador de regresión para la media poblacional. El estimador separado de regresión de la media poblacional se obtiene como media ponderada de los estimadores de los estratos.

Para cada estrato h $\longrightarrow$ $\overline{x}_{regh} = \overline{x}_h + b_h(\overline{Y}_h - \overline{y}_h)$

$$\hat{\overline{X}}_{regs} = \overline{x}_{regs} = \sum_{h=1}^{L} W_h \cdot \overline{x}_{regh} \qquad / \; W_h = \frac{N_h}{N}$$

~~Es un estimador insesgado de la media poblacional~~

~~$E[\overline{x}_{regs}] = \overline{X}$~~

En general, es un estimador sesgado, salvo en el caso de que las pendientes de regresión sean constantes prefijadas

Para cada estrato, $B(\overline{x}_{regh}) = -\operatorname{cov}(b_h, \overline{y}_h) = 0$ si $b_h = cte$

$$B(\overline{x}_{regs}) = \sum_{h=1}^{L} W_h\, B(\overline{x}_{regh}) = -\sum_{h=1}^{L} W_h \operatorname{cov}(b_h, \overline{y}_h)$$

- En el supuesto $b = b_0$ cte, la varianza del estimador separado es:

$$(SR) \longrightarrow V(\overline{x}_{regs}) = \sum_{h=1}^{L} W_h^2 \cdot \frac{1-f_h}{n_h}\left[S^2_{xh} + b_0^2 S^2_{yh} - 2b_0 S_{xyh}\right]$$

$$(CR) \longrightarrow V(\overline{x}_{regc}) = \sum_{h=1}^{L} W_h^2 \cdot \frac{1}{n_h}\left[\sigma^2_{xh} + b_0^2 \sigma^2_{yh} - 2b_0 \sigma_{xyh}\right]$$

Estimaciones con ^

- La varianza será mínima cuando lo sean $V(\bar{x}_{reg\,h})$, es decir, cuando $b_{MIN} = \beta_h = \dfrac{S_{xyh}}{S^2_{yh}} = \dfrac{cov(x_h y_h)}{V(y_h)}$

$$V_{MIN}(\bar{x}_{reg\,s}) = \sum_{h=1}^{L} W_h^2 \cdot V(\bar{x}_{reg\,h}) =$$

$$(SR) \longrightarrow \quad = \sum_{h=1}^{L} W_h^2 \cdot \frac{1-f_h}{n_h} \left(S^2_{xh} + b^2_{MIN} S^2_{yh} - 2b_{MIN} S_{xyh}\right) =$$

$$= \dots = \sum_{h=1}^{L} W_h^2 \cdot \frac{1-f_h}{n_h} S^2_{xh} (1-\rho_h^2)$$

$$(CR) \longrightarrow \quad = \sum_{h=1}^{L} W_h^2 \cdot \frac{1}{n_h} \left(\sigma^2_{xh} + b^2_{MIN} \sigma^2_{yh} - 2b_{MIN} \sigma_{xyh}\right) =$$

$$= \dots = \sum_{h=1}^{L} W_h^2 \cdot \frac{1}{n_h} \sigma^2_{xh} (1-\rho^2_{xyh})$$

b) Estimador <u>combinado</u> de regresión:

A partir del muestreo estratificado, se estiman las medias poblacionales de la variable en estudio $\bar{X}$ y de la var. auxiliar $\bar{Y}$ mediante los estimadores $\bar{x}_{st}$, $\bar{y}_{st}$. A partir de estos estimadores se construye el estimador combinado de regresión para la media poblacional:

$$\bar{x}_{reg\,c} = \bar{x}_{st} + b(\bar{Y} - \bar{y}_{st})$$

El sesgo es

$B(\bar{x}_{reg\,c}) = -cov(b, \bar{y}_{st})$, que vale 0 cuando b es uno ya prefijado.

~~$V(\bar{x}_{reg\,c})$~~

- Si $b = b_0$, cte, la varianza es:

$$V(\bar{x}_{reg\,c}) = V(\bar{x}_{st} + b_0(\bar{Y} - \bar{y}_{st})) = V(\bar{x}_{st}) + b_0^2 V(\bar{y}_{st}) - 2b_0 \, COV(\bar{x}_{st}, \bar{y}_{st}) =$$

(SR) $\longrightarrow \displaystyle\sum_{h=1}^{L} W_h^2 \cdot \frac{1-f_h}{n_h} \left[S^2_{xh} + b_0^2 S^2_{yh} - 2b_0 S_{xyh} \right]$

(CR) $\longrightarrow \displaystyle\sum_{h=1}^{L} W_h^2 \cdot \frac{1}{n_h} \left[\sigma^2_{xh} + b_0^2 \sigma^2_{yh} - 2b_0 \sigma_{xyh} \right]$

- El valor de la constante que hace mínima la varianza es:

$$b_{MIN} = \frac{\sum W_h^2 \frac{1-f_h}{n_h} S_{xyh}}{\sum W_h^2 \frac{1-f_h}{n_h} S^2_{yh}}$$ ← SR, media ponderada de las pendientes de regresión de los estratos

$$\frac{\sum W_h^2 \frac{1}{n_h} S_{xyh}}{\sum W_h^2 \frac{1}{n_h} S^2_{y}}$$ ← CR

y la expresión de la varianza mínima es:

(SR) $\longrightarrow V_{MIN}(\bar{x}_{reg\,c}) = V(\bar{x}_{st})(1 - \rho^2_{\bar{x}_{st}, \bar{y}_{st}})$

(CR) $\longrightarrow \boxed{V_{MIN}(\bar{x}_{reg\,c}) = V(\bar{x}_{st})(1 - \rho^2_{x_{st}, y_{st}})}$ No sé, supongo...

Si queremos comparar estos dos estimadores, al igual que en el caso del estimador de razón:

- Si las pendientes de regresión en cada estrato, b_h, difieren entre sí, se recomienda el estimador separado. Cuando $b_h = cte$, los dos estimadores son igual de precisos.

- Si los sesgos de los estimadores de regresión en cada estrato son todos positivos, pueden acumularse, y se recomienda utilizar el estimador combinado.

MUEST_T8. ESTIMADORES DE RAZÓN en el muestr. ESTRATIFICADO (SEPARADO Y COMBINADO).
SESGO, VARIANZAS y SUS ESTIMACIONES.
COMPARACIÓN de PRECISIONES.
REFERENCIA A LOS ESTIMADORES DE REGRESIÓN EN EL MUESTREO

0_MUESTREO ESTRATIFICADO y estim. razón

Concepto de muestreo estratificado:

¿por qué?

¿cómo?

¿cuándo?

1 - ESTIMADORES de RAZÓN en el MUESTREO ESTRATIFICADO: SEPARADO y COMBINADO

En el muestreo estratificado pueden considerarse también estimaciones de la razón.

Existen dos técnicas distintas de obtención de estimadores:

- estimación simple o separada:
  - ↳ obtener estimadores de razón para cada estrato y agruparlos
- estimación combinada:
  - ↳ realizan estimaciones para los parámetros poblacionales directamente mediante razones de estimadores estratificados de la variable en estudio y la variable auxiliar.

Si se desea estimar el total poblacional:

$$\theta = X = \sum_{h=1}^{L} \sum_{i=1}^{N_h} X_{hi}$$

a) Estimador separado de razón:

Para cada estrato se consideran estimaciones para el total:

$$\hat{X}_{Rh} = \hat{R}_h \cdot Y_h = \frac{\bar{x}_h}{\bar{y}_h} \cdot Y_h \qquad \text{para } h = 1, \dots L$$

El estimador del total se obtiene como suma de los estimadores de los totales en cada estrato:

$$\underset{\substack{\uparrow \\ \text{separado}}}{\hat{X}_{RS}} = \sum_{h=1}^{L} \hat{X}_{Rh} = \sum_{h=1}^{L} \hat{R}_h \cdot Y_h = \hat{R}_S \cdot Y$$

$$\hat{R}_S = \sum \frac{Y_h}{Y} \hat{R}_h$$

N_h

Ventaja: Tiene en cuenta que la ratón en cada estrato, $R_h = \frac{X_h}{Y_h}$, puede variar.
ofrece información al subnivel de estratos.
Inconveniente: Requiere conocer el total poblacional de la variable auxiliar en cada estrato, Y_h.
Se acumulan los sesgos de los estratos para el sesgo total.

b) Estimador combinado de ratón

A partir del muestreo estratificado, se obtienen los estimadores de los totales X e Y con $\hat{X}_{st}$, $\hat{Y}_{st}$.

El estimador de ratón se obtiene multiplicando la ratón de estos estimadores por el total de la variable auxiliar:

$$\underset{\text{combinado}}{\hat{X}_{RC}} = \frac{\hat{X}_{st}}{\hat{Y}_{st}} \cdot Y = \frac{\sum_{h=1}^{L} N_h \cdot \bar{x}_h}{\sum_{h=1}^{L} N_h \cdot \bar{y}_h} Y = \hat{R}_c \cdot Y$$

donde: $\hat{R}_c$ = ratón combinado, estimador de R

$\boxed{\text{MIRAR ATRÁS}}$

Ventaja: No requiere conocer el total de la variable auxiliar en cada estrato Y_h, basta conocer el total poblacional Y. No acumula los sesgos de los estratos.

Inconveniente: De forma implícita supone que la ratón es constante en todos los estratos. No permite disponer de información al subnivel de estrato.

Para la media poblacional:

$$\hat{\bar{X}}_{RS} = \sum_{h=1}^{L} W_h \hat{\bar{X}}_{Rh} = \sum_{h=1}^{L} W_h \hat{R}_h \cdot \bar{Y}_h = \sum_{h=1}^{L} \frac{\cancel{N_h}}{N} \cdot \hat{R}_h \cdot \frac{Y_h}{\cancel{N_h}} = \frac{1}{N} \sum_{h=1}^{L} \hat{R}_h \cdot Y_h =$$

$$= \sum_{h=1}^{L} \frac{\hat{X}_{Rh}}{N}$$

$$\hat{\bar{X}}_{RC} = \hat{R}_c \cdot \bar{Y}$$

2_ SESGOS, VARIANZAS y SUS ESTIMACIONES

a)

SESGOS:

a) Estimador separado para el total:

El estimador de razón en general es sesgado:

$$E[\hat{R}_h] = E\left[\frac{\bar{x}_h}{\bar{y}_h}\right] \neq \frac{E[\bar{x}_h]}{E[\bar{y}_h]} = R_h \quad , \; h=1\dots L$$

$$cov[\hat{R}_h, \bar{y}_h] = E[\hat{R}_h \cdot \bar{y}_h] - E[\hat{R}_h]E[\bar{y}_h] =$$

$$= E\left[\frac{\bar{x}_h}{\cancel{\bar{y}_h}} \cdot \cancel{\bar{y}_h}\right] - E[\hat{R}_h]\cdot \bar{Y}_h =$$

$$= \bar{X}_h - E[\hat{R}_h]\cdot \bar{Y}_h = \bar{Y}_h\left(\frac{\bar{X}_h}{\bar{Y}_h} - E[\hat{R}_h]\right) =$$

$$= \bar{Y}_h\left(R_h - E[\hat{R}_h]\right)$$

por lo que el sesgo:

$$B(\hat{R}_h) = E[\hat{R}_h] - R_h = -\frac{cov[\hat{R}_h, \bar{y}_h]}{\bar{Y}_h}$$

Si el estimador de razón es sesgado, también lo será el estimador separado para el total:

$$\boxed{\hat{R}_S = \sum_{h=1}^{L} \frac{Y_h}{Y}\hat{R}_h} \leftarrow \text{suma ponderada de los estim. razón en cada estrato}$$

(N_h)

$$E[\hat{R}_S] = E\left[\sum_{h=1}^{L} \frac{Y_h}{Y}\hat{R}_h\right] = \sum_{h=1}^{L}\frac{Y_h}{Y}E[\hat{R}_h]$$

$$B[\hat{R}_S] = E[\hat{R}_S] - R = \sum_{h=1}^{L}\frac{Y_h}{Y}E[\hat{R}_h] - R = \sum_{h=1}^{L}\frac{Y_h}{Y}\left[R_h - \frac{cov[\hat{R}_h, \bar{y}_h]}{\bar{Y}_h}\right] - R$$

$$= \sum_{h=1}^{L}\frac{\cancel{Y_h}}{Y}\cdot\frac{X_h}{\cancel{Y_h}} - \sum_{h=1}^{L}\frac{\cancel{Y_h}}{Y}\cdot\frac{cov[\hat{R}_h, \bar{y}_h]}{\cancel{Y_h}/N_h} - R =$$

$$= \cancel{R} - \frac{1}{Y}\sum_{h=1}^{L} N_h \cdot cov[\hat{R}_h, \bar{y}_h] - \cancel{R} = -\frac{1}{Y}\sum_{h=1}^{L} N_h\cdot\frac{cov[\hat{R}_h, \bar{y}_h]}{\bar{Y}_h}\cdot\bar{Y}_h$$

$$\boxed{= \sum_{h=1}^{L} B(\hat{R}_h)\cdot Y_h/Y}$$

Girar atrás

$$\hat{X}_{RS} = \sum_{h=1}^{L} \hat{R}_h Y_h \qquad X = \sum_{h=1}^{L} X_h = \sum_{h=1}^{L} R_h \cdot Y_h \qquad \hat{X}_{RS} = \hat{R}_S \cdot Y$$

$$B(\hat{X}_{RS}) = E[\hat{X}_{RS}] - X = E\left[\sum_{h=1}^{L} \hat{R}_h Y_h\right] - X =$$
$$= E\left[\sum_{h=1}^{L} \hat{R}_h Y_h\right] - \sum_{h=1}^{L} R_h Y_h = \sum Y_h E[\hat{R}_h] - \sum R_h Y_h =$$
$$= \sum_{h=1}^{L} Y_h \left(E[\hat{R}_h] - R_h\right) = \sum_{h=1}^{L} Y_h \cdot B(\hat{R}_h) \leftarrow \text{suma ponderado de los sesgos}$$

y sustituyendo la expresión del sesgo de $\hat{R}_h$:

$$B(\hat{X}_{RS}) = -\sum_{h=1}^{L} Y_h \cdot \frac{cov(\hat{R}_h, \bar{y}_h)}{\bar{Y}_h} = -\sum_{h=1}^{L} N_h \cdot cov(\hat{R}_h, \bar{y}_h).$$

Por lo que los estimadores $\hat{R}_S$ y $\hat{X}_{RS}$ serán insesgados si $cov(\hat{R}_h, \bar{y}_h) = 0$, $h = 1 \ldots L$, es decir, si la variable en estudio y la variable auxiliar están incorrelacionadas.

Suponiendo que todos los sesgos tienen el mismo signo (para que no se compensen al sumarse), se puede acotar el valor del sesgo del estimador de razón separado:

$$cov(\hat{R}_h, \bar{y}_h) = \sigma_{\hat{R}_h} \cdot \sigma_{\bar{y}_h} \cdot \rho_{\hat{R}_h, \bar{y}_h} = \bar{Y}_h \underbrace{(R_h - E[\hat{R}_h])}_{-B(\hat{R}_h)} \Longrightarrow$$

$$B(\hat{R}_h) = -\rho_{\hat{R}_h, \bar{y}_h} \cdot \sigma_{\hat{R}_h} \cdot \frac{\sigma_{\bar{y}_h}}{\bar{Y}_h} \Longrightarrow \frac{|B(\hat{R}_h)|}{\sigma_{\hat{R}_h}} = |\rho_{\hat{R}_h, \bar{y}_h}| \cdot \underbrace{\frac{\sigma_{\bar{y}_h}}{\bar{Y}_h}}_{C_{\bar{y}_h}}$$

$$\frac{|B(\hat{R}_h)|}{\sigma_{\hat{R}_h}} = \underbrace{|\rho_{\hat{R}_h, \bar{y}_h}|}_{\leq 1} \cdot C_{\bar{y}_h} \leq C_{\bar{y}_h}$$

luego
$$\left.\begin{array}{l} B(\hat{R}_S) \simeq L \cdot B(\hat{R}_h) \\ \sigma_{\hat{R}_S} \simeq \sqrt{L} \cdot \sigma_{\hat{R}_h} \end{array}\right\} \Longrightarrow \frac{|B(\hat{R}_S)|}{\sigma_{\hat{R}_S}} \leq \sqrt{L}\, C_{\bar{y}_h}$$

Por tanto, si el nº de estratos es muy grande, el sesgo también puede serlo y su contribución al ECM $\rightarrow ECM = V(\hat{\theta}) + B^2(\hat{\theta})$ puede ser importante. En la práctica, el sesgo suele ser mucho menor que la cota superior.

Sin embargo, el sesgo decrece al aumentar ~~con~~ el tamaño de la muestra, y es prácticamente despreciable para muestras de los estratos sufic. grandes. Ver T4

→ Expresión aproximada del sesgo: $b(\hat{R}_h) \simeq \frac{R_h V(\bar{y}_h) - Cov(\bar{x}_h \bar{y}_h)}{\bar{Y}_h^2} = R_h (C_{\bar{y}}^2 - C_{\bar{x}\bar{y}})$

Expresión del sesgo para ≠ muestreos:

$$SR \rightarrow B(\hat{X}_{RS}) = \sum_{h=1}^{L} Y_h \cdot \frac{(1-f_h)}{n_h \bar{Y}_h^2} (R_h S_{Y_h}^2 - S_{X_h Y_h}) =$$

$$= \sum_{h=1}^{L} \frac{N_h^2}{Y_h} \cdot \frac{(1-f_h)}{n_h} (R_h S_{Y_h}^2 - S_{X_h Y_h})$$

$$CR \rightarrow B(\hat{X}_{RS}) = \sum_{h=1}^{L} Y_h \cdot \frac{1}{n_h \bar{Y}_h^2} (R_h \sigma_{Y_h}^2 - \sigma_{XY_h}) =$$

$$= \sum_{h=1}^{L} \frac{N_h^2}{Y_h} \cdot \frac{1}{n_h} (R_h \sigma_{Y_h}^2 - \sigma_{XY_h}).$$

que pueden estimarse por:

$$SR \rightarrow \hat{B}(\hat{X}_{RS}) = \sum_{h=1}^{L} \frac{N_h^2}{Y_h} \cdot \frac{(1-f_h)}{n_h} (\hat{R}_h \hat{S}_{Y_h}^2 - \hat{S}_{XY_h})$$

$$CR \rightarrow \hat{B}(\hat{X}_{RS}) = \sum_{h=1}^{L} \frac{N_h^2}{Y_h} \cdot \frac{1}{n_h} \cdot (\hat{R}_h \hat{\sigma}_{Y_h}^2 - \hat{\sigma}_{XY_h}).$$ estim. insesg. de σ es $\hat{S}$

Para la media poblacional:

$$B(\hat{\bar{X}}_{RS}) = E[\hat{\bar{X}}_{RS}] - \bar{X} = E\left[\frac{\hat{X}_{RS}}{N}\right] - \frac{X}{N} = \frac{1}{N} B(\hat{X}_{RS}) = \sum B(\hat{R}_h) \cdot \frac{Y_h}{N}$$

$$SR \rightarrow B(\hat{\bar{X}}_{RS}) = \sum_{h=1}^{L} \frac{W_h (1-f_h)}{n_h \bar{Y}_h} (R_h S_{Y_h}^2 - S_{XY_h}).$$

$$CR \rightarrow B(\hat{\bar{X}}_{RS}) = \sum_{h=1}^{L} \frac{W_h}{n_h \bar{Y}_h} (R_h \sigma_{Y_h}^2 - \sigma_{XY_h})$$

$\hat{B}$ → poner ^