MUEST_T5 . ESTIMADOR de REGRESIÓN en el m.a.s.
SESGO, VARIANZA y sus ESTIMACIONES.
COMPARACIONES CON EL ~~ESTI~~ $\hat{X}_R$ y $\hat{X}$.
REFERENCIA a la ESTIM. GENERALIZADA por REGRESIÓN.

Var < bcte, bvarmin, b óptim.

---

## 1_ ESTIMADOR de REGRESIÓN en el MUEST. ALEAT. SIMPLE

Los métodos indirectos aprovechan la información conocida relativa a una var. auxiliar Y correlacionada con la var. de estudio X para obtener estimaciones más precisas para $\bar{X}$ que las obtenidas solamente con la información de $\bar{X}$.

Si entre $\bar{X}$ e $\bar{Y}$ hay una relación lineal (o aproximadamente lineal) ~~en~~ se puede utilizar un método de estimación basado en la regresión lineal de X sobre Y.
(Si la recta pasase por el origen, se utilizaría el estimador de razón).

Para la población estudiada se observa el par $(X_i, Y_i)$ de la variable en estudio y la variable auxiliar, con media poblacional conocida, $\bar{Y}$ conocido.

Si los puntos están situados sobre una recta:

$$X_i = a + bY_i \Rightarrow \sum^N X_i = Na + b\sum^N Y_i \Rightarrow \bar{X} = a + b\bar{Y}$$

Tomando una mas(n):

$$x_i = a + by_i \Rightarrow \sum^n x_i = na + b\sum^n y_i \Rightarrow \bar{x} = a + b\bar{y}$$

luego $\bar{X} - \bar{x} = b(\bar{Y} - \bar{y})$, por lo que:

$$\bar{x}_{reg} = (\hat{\bar{X}}_{reg}) = \bar{x} + b(\bar{Y} - \bar{y})$$

→ estimador de regresión de la media poblacional

$$\hat{X}_{reg} = N\bar{x}_{reg} = N\bar{x} + b(N\bar{Y} - N\bar{y}) =$$
$$= \hat{X} + b(Y - \hat{Y})$$ ← estimador de regresión del total poblacional.

En el caso de que $\bar{Y} = \bar{y}$, el estimador de regresión de la media poblacional coincide con la media muestral

$$\bar{Y} = \bar{y} \Rightarrow \bar{x}_{reg} = \bar{x}$$

si $\bar{Y} \neq \bar{y}$, también se espera que $\bar{X} \neq \bar{x}$, por lo que $b(\bar{Y} - \bar{y})$ es el sumando de ajuste para una posible estimación por defecto o por exceso.

CASOS PARTICULARES:

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

- $b = 0 \Rightarrow \bar{x}_{reg} = \bar{x}$ (estimador simple)
  ↳ ocurre cuando no hay correlación lineal entre X e Y.
- $b = \dfrac{\bar{x}}{\bar{y}} \Rightarrow \bar{x}_{reg} = \bar{x}_R$ (estimador de razón)
  $$\bar{x}_{reg} = \bar{x} + \frac{\bar{x}}{\bar{y}}(\bar{Y} - \bar{y}) = \bar{x} + \frac{\bar{x}}{\bar{y}}\bar{Y} - \bar{x} = \frac{\bar{x}}{\bar{y}}\bar{Y} = \hat{\bar{X}}_R.$$
- $b = 1 \Rightarrow \bar{x}_{reg} = \bar{x}_{dif}$ (estimador por diferencia)
  $$\bar{x}_{reg} = \bar{x} + (\bar{Y} - \bar{y}) = \bar{x} - \bar{y} + \bar{Y}$$

---

Para el total poblacional: $\hat{X}_{reg} = N\bar{x}_{reg}$

* César Pérez da la expresión para los estimadores de regresión del total de clase y de la proporción poblacional, ¿tiene sentido la regresión?

$$\hat{P}_{reg} = \hat{P}_X + b(P_Y - \hat{P}_Y) \quad \text{y} \quad \hat{A}_{reg} = N\hat{P}_{reg}.$$

## 2. SESGO, VARIANZA y ESTIMACIONES

SESGO:

El estimador de regresión es, en general, sesgado.

$$E[\bar{x}_{reg}] = E[\bar{x} + b(\bar{Y} - \bar{y})] = E[\bar{x}] + \bar{Y}E[b] - E[b\bar{y}] =$$
$$= E[\bar{x}] - \left(E[b\bar{y}] - E[b]E[\bar{y}]\right) = \bar{X} - COV(b, \bar{y})$$

$$B[\bar{x}_{reg}] = E[\bar{x}_{reg}] - \bar{X} = -COV(b, \bar{y})$$

El estimador de regresión será insesgado si $COV(b, \bar{y})$ es 0, que puede ocurrir cuando:

- $b = 0 \Rightarrow$ no hay correlación lineal entre $X$ e $Y$
- $b = cte \Rightarrow$ se elige el valor de b antes de observar la muestra.

VARIANZA del ESTIMADOR:

1) Para b cte ($\neq 0$), el estimador de regresión es insesgado y su varianza es:

$$V[\bar{x}_{reg}] = V[\bar{x} + b(\bar{Y} - \bar{y})] = V[\bar{x}] + \underset{\text{b cte}}{b^2} V[\underset{\text{conocido}}{\bar{Y}} - \bar{y}] - 2b\,COV[\bar{x}, \bar{y} - \bar{Y}] =$$
$$= V[\bar{x}] + b^2 V[\bar{y}] - 2b\,COV[\bar{x}, \bar{y}] \longleftarrow \text{Expresión general}$$

expresión que se puede concretar para los $\neq$ tipos de muestreo:

$$SR \longrightarrow V[\bar{x}_{reg}] = \frac{1-f}{n}\left(S_x^2 + b^2 S_y^2 - 2bS_{xy}\right)$$

que tiene por estimador insesgado

$$\hat{V}[\bar{x}_{reg}] = \frac{1-f}{n}\left(\hat{S}_x^2 + b^2\hat{S}_y^2 - 2b\hat{S}_{xy}\right)$$

Para el total poblacional:

$$V[\hat{X}_{reg}] = N^2 V[\bar{x}_{reg}] = N^2\frac{(1-f)}{n}\left(S_x^2 + b^2 S_y^2 - 2bS_{xy}\right)$$
$$\hat{V}[\hat{X}_{reg}] = N^2\hat{V}[\bar{x}_{reg}] = N^2\frac{(1-f)}{n}\left(\hat{S}_x^2 + b^2\hat{S}_y^2 - 2b\hat{S}_{xy}\right)$$

CR $\rightarrow V[\bar{x}_{reg}] = \frac{1}{n}(\sigma_x^2 + b^2\sigma_y^2 - 2b\sigma_{xy})$

que tiene por estimador insesgado:

$$\hat{V}[\bar{x}_{reg}] = \frac{1}{n}(\hat{S}_x^2 + b^2\hat{S}_y^2 - 2b\hat{S}_{xy})$$

Para el total poblacional:

$$V[\hat{X}_{reg}] = N^2 V[\bar{x}_{reg}] = \frac{N^2}{n}(\sigma_x^2 + b^2\sigma_y^2 - 2b\sigma_{xy})$$

$$\hat{V}[\hat{X}_{reg}] = N^2\hat{V}[\bar{x}_{reg}] = \frac{N^2}{n}(\hat{S}_x^2 + b^2\hat{S}_y^2 - 2b\hat{S}_{xy}).$$

2) Para b MÍNIMA varianza

Como la varianza del estimador de regresión depende de la de b, que normalmente se determina después de observar la muestra, se puede elegir el valor de la de b que haga mínima la varianza del estimador.

$$\min_b V(\bar{x}_{reg}) = \phi(b) \longrightarrow \begin{array}{l} CN: \phi'(b) = 0 \Rightarrow b_{MIN} \\ CS: \phi''(b)\big|_{b_{MIN}} > 0 \Rightarrow \text{MÍNIMO}. \end{array}$$

Es un problema de optimización clásico:

$$\phi(b) = V(\bar{x}_{reg}) = V[\bar{x}] + b^2 V[\bar{y}] - 2b\,COV[\bar{x},\bar{y}]$$

$$\phi'(b) = 2bV[\bar{y}] - 2COV[\bar{x},\bar{y}] = 0 \Leftrightarrow b_{MIN} = \frac{COV[\bar{x},\bar{y}]}{V[\bar{y}]}$$

$$\phi''(b)\big|_{b_{MIN}} = 2V[\bar{y}]\big|_{b_{MIN}} = 2V[\bar{y}] > 0 \text{ siempre}$$

$$b_{MIN} = \frac{COV[\bar{x},\bar{y}]}{V[\bar{y}]}$$, sustituyendo esta expresión en la varianza del estimador:

$$V_{MIN}[\bar{x}_{reg}] = V[\bar{x}] + \left(\frac{COV[\bar{x},\bar{y}]}{V[\bar{y}]}\right)^2 V[\bar{y}] - 2\frac{COV[\bar{x},\bar{y}]}{V[\bar{y}]}\cdot COV[\bar{x},\bar{y}] =$$

$$= V[\bar{x}] - \frac{COV^2[\bar{x},\bar{y}]}{V[\bar{y}]} = V[\bar{x}](1 - \rho_{xy}^2)$$

$$\left.\begin{array}{l} b_{MIN} = \dfrac{cov[\bar{x},\bar{y}]}{V[\bar{y}]} \\ V_{MIN}(\bar{x}_{reg}) = V[\bar{x}](1-\rho^2_{xy}) \end{array}\right\} \longleftarrow \text{Expresión general}$$

Se puede particularizar para los ≠ tipos de muestreo:

$$SR \longrightarrow \begin{cases} b_{MIN} = \dfrac{S_{xy}}{S^2_y} = \beta \\ V_{MIN}(\bar{x}_{reg}) = \dfrac{1-f}{n} S^2_x (1-\rho^2_{xy}) \\ \hat{V}_{MIN}(\bar{x}_{reg}) = \dfrac{1-f}{n} \hat{S}^2_x (1-\hat{\rho}^2_{xy}) \end{cases}$$

Para el total: $V_{MIN}(\hat{X}_{reg}) = N^2 V_{min}(\bar{x}_{reg})$ $\uparrow$ SR

$\hat{V}_{min}(\hat{X}_{reg}) = N^2 \hat{V}_{min}(\bar{x}_{reg})$ $\downarrow$ CR

$$CR \longrightarrow \begin{cases} b_{MIN} = \dfrac{\sigma_{xy}}{\sigma^2_y} = \dfrac{S_{xy}}{S^2_y} = \beta \\ V_{MIN}(\bar{x}_{reg}) = \dfrac{1}{n} \sigma^2_x (1-\rho^2) \\ \hat{V}_{MIN}(\bar{x}_{reg}) = \dfrac{1}{n} \hat{S}^2_x (1-\hat{\rho}^2) \end{cases}$$

3) En principio, el valor $b_{MIN} = \dfrac{S_{xy}}{S^2_y}$ no depende de la muestra obtenida, por lo que puede fijarse de antemano. Pero los parámetros poblacionales $S_{xy}$ y $S^2_x$ no siempre son conocidos $\longrightarrow \hat{b}_{MIN} = \dfrac{\hat{S}_{xy}}{\hat{S}^2_y} = \dfrac{\sum^n (X_i-\bar{x})(Y_i-\bar{y})}{\sum^n (Y_i-\bar{y})^2}$

que es un estimador insesgado de $b_{MIN}$ para muestras grandes

## 3_ COMPARACIONES del ESTIMADOR de ~~RAZÓN~~ REGRESIÓN con EL ESTIMADOR de EXPANSIÓN y CON EL DE RAZÓN

A la hora de comparar dos estimadores, será más preciso aquél con menor varianza.

Necesitamos un tamaño muestral sufic. grande para que las aproximaciones de las varianzas de los estim. de razón y de regresión sean válidas.

Utilizamos el estimador de la media poblacional y su varianza en los diferentes tipos de muestreo.

* En el caso general, las expresiones de las varianzas difieren bastante unas de otras, por lo que resulta más agradable hacer las comparaciones para los diferentes tipos de muestreo.

Parámetro poblacional: $\theta = \overline{X}$ → media poblacional

Estimadores: $\hat{\overline{X}} = \begin{cases} \overline{x} \rightarrow \text{media muestral (estim. expansión)} \\ \hat{\overline{X}}_R = \hat{R} \cdot \overline{Y} \rightarrow \text{estim. razón} \\ \hat{\overline{X}}_{reg} = \overline{x} + b(\overline{Y} - \overline{y}) \rightarrow \text{estim. regresión} \end{cases}$

cuyas varianzas cambian de expresión al cambiar el tipo de muestreo:

1) SR → $V(\overline{x}) = \frac{(1-f)}{n} S_x^2$

$$V(\hat{\overline{X}}_R) = \frac{1-f}{n}\left(S_x^2 + R^2 S_y^2 - 2R\, S_x S_y \cdot \rho_{xy}\right)$$

$$V(\hat{\overline{X}}_{reg}) = V_{min}(\overline{x}_{reg}) = \frac{1-f}{n} S_x^2 \left(1 - \rho_{xy}^2\right)$$

Es evidente que $\underset{min}{V}(\bar{x}_{reg}) \leq V(\bar{x})$ porque:

$$1-\rho^2_{xy} \leq 1 \longrightarrow 0 \leq \rho^2_{xy} \leq 1 \Rightarrow 1-\rho^2_{xy} \leq 1.$$

$$1-\rho^2_{xy} = 1 \Longleftrightarrow \rho^2_{xy} = 0 \Longleftrightarrow X \text{ e } Y \text{ incorreladas}$$

$\Rightarrow$ La estimación por regresión es mejor que la estimación por muestreo aleatorio simple, excepto cuando las var. están incorreladas, ~~en~~ caso en que son iguales.

Por otra parte,

$$V(\bar{x}_{reg}) < V(\bar{x}_R) \Longleftrightarrow V(\bar{x}_R) - V(\bar{x}_{reg}) > 0$$

$$\frac{1-f}{n}\left[\cancel{S^2_x} + R^2 S^2_y - 2RS_xS_y\rho_{xy} - \cancel{S^2_x} + S^2_x\rho^2_{xy}\right] \geq 0$$

$$\Longleftrightarrow \underbrace{R^2S^2_y - 2RS_xS_y\rho_{xy} + S^2_x\rho^2_{xy}}_{\text{cuadrado de una dif}} \geq 0 \Longleftrightarrow$$

$$(RS_y - S_x\rho_{xy})^2 \geq 0 \longrightarrow \text{SIEMPRE}$$

↳ Se produce la igualdad si $R = \frac{S_x}{S_y}\rho_{xy} = \beta$.

$\Rightarrow$ recta regresión pasa por origen

$\Rightarrow$ La estimación por regresión es más precisa que la estimación por razón en todos los casos, salvo cuando la recta de regresión pasa por el origen, en que los 2 est. son igual de precisos.

2) CR $\longrightarrow$ $V(\bar{x}) = \frac{1}{n}\sigma_x^2$

$$V(\bar{x}_R) = \frac{1}{n}(\sigma_x^2 + R^2\sigma_y^2 - 2R\sigma_x\sigma_y\rho_{xy})$$

$$V_{min}(\bar{x}_{reg}) = \frac{1}{n}\sigma_x^2(1-\rho_{xy}^2)$$

- $V_{min}(\bar{x}_{reg}) \leq V(\bar{x})$ porque $1-\rho_{xy}^2 \leq \cancel{0}\,1$

  $1-\rho_{xy}^2 = \cancel{0}\,1$ si X e Y incorreladas

- $V_{min}(\bar{x}_{reg}) \leq V(\bar{x}_R)$ siempre

  $V(\bar{x}_{reg}) = V(\bar{x}_R)$ si $R = \beta \to$ recta pasa por origen

$V(\bar{x}_R) - V_{min}(\bar{x}_{reg}) \geq 0 \iff$

$$\frac{1}{n}\left[\cancel{\sigma_x^2} + R^2\sigma_y^2 - 2R\sigma_x\sigma_y\rho_{xy} - \cancel{\sigma_x^2} + \sigma_x^2\rho_{xy}^2\right] \geq 0$$

$$\frac{1}{n}\left[(R\sigma_y - \sigma_x\rho_{xy})^2\right] \geq 0 \text{ siempre}$$

$$= 0 \text{ si } R = \frac{\sigma_x}{\sigma_y}\rho_{xy} = \beta.$$

## 4_ REFERENCIA A LA ESTIMACIÓN GENERALIZADA por REGRESIÓN

la estimación por regresión se puede generalizar al caso en el que se dispone de K variables auxiliares que tengan correlación con la var. en estudio X.

Sea $\vec{Y}_i$ el vector de variables auxiliares observadas sobre el elemento poblacional $u_i$.

Para $u_i, i=1 \dots N \longrightarrow \vec{Y}_i = (Y_{1i} \dots Y_{Ki})$

Suponemos que cada valor de la var. a estudiar se puede relacionar de manera lineal con $\vec{Y}_i$:

$$X_i = \sum_{j=1}^{K} \beta_j Y_{ji} + \varepsilon_i \quad , i=1 \dots N \longrightarrow X = Y'\beta + \varepsilon$$

donde

$$E[X_i] = \sum_{j=1}^{K} \beta_j Y_{ji} \quad , i=1 \dots N \quad (\text{sp. } Y \text{ determinist})$$

$$V[X_i] = \sigma_i^2 \quad , i=1 \dots N$$

$\varepsilon_i \sim N(0, \sigma_i) \quad i=1 \dots N \rightarrow$ perturbación o término alet.

Al estar las var. auxiliares correlacionadas, el modelo incumple la hipótesis de ~~hetero~~ homoscedasticidad ($\sigma_i^2 \neq \sigma^2$ cte), por lo que para estimar el vector de coeficientes de regresión $\vec{\beta} = (\beta_1 \dots \beta_K)^t$ hay que utilizar Mínimos Cuadrados Generalizados (ponderados).

$$\hat{\vec{\beta}} = \left[ \sum_{i=1}^{N} \frac{\vec{Y}_i \vec{Y}_i^t}{\sigma_i^2} \right]^{-1} \sum \frac{\vec{Y}_i X_i}{\sigma_i^2}$$

$$\hat{\beta} = (Y^{*\prime} Y^*)^{-1} Y^{*\prime} X^* \qquad \text{donde } (\;)^* = \frac{(\;)}{\sigma_i}$$

Al depender de los valores poblacionales $X_1 \ldots X_N$, el estimador $\vec{\beta}$ no se puede obtener, por lo que hay que estimarlo a partir de las observaciones muestrales.

En el caso de m.a. ~~s~~. s.r.p. desig.:

$$\hat{\vec{\beta}} = \left( \sum_{i=1}^{n} \frac{\vec{y}_i \vec{y}_i^{\,t}}{\sigma_i^2 \pi_i} \right)^{-1}_{K\times K} \cdot ~~\text{matriz } K\times K~~ \cdot \left( \sum^{n} \frac{\vec{y}_i X_i}{\sigma_i^2 \pi_i} \right)_{K\times 1} = (\ )_{K\times 1}$$

Por lo que el estimador generalizado de regresión para el total poblacional resulta:

$$\hat{X}_{G.reg} = \hat{X} + \sum^{K} \hat{\beta}_j ( Y_j - \hat{Y}_j )$$

donde:

$Y_j$ → total poblacional de la var. auxiliar $y_j$.

$\hat{X}$ → estimador insesgado del total de $X$, $\hat{X} = \sum^{n} \frac{x_i}{\pi_i}$

$\hat{Y}_j$ → estimador insesgado del total de $Y$

$\hat{\beta}$ → estimador generalizado de regresión para $X$.

La varianza de $\hat{\beta}$ no se puede calcular, pero se pueden aplicar técnicas de ~~regresión~~ linealización para obtener una aproximación del estimador y de su varianza.