Técnicas de Investigación Social

Grado en Relacións Laborais e Recursos Humanos

Curso 2019-2020

Tema 5. Mostraxe de razón e regresión

Alejandro Saavedra Nieves

Estimadores de Razón

Estimadores de razón

O obxectivo deste tema pasa por traballar sobre pares de datos (X, Y) de maneira conxunta.

Para cada unidade da mostra extraída, además de obter información sobre unha certa variable Y, tamén obteremos información doutra variable X asociada.

En adiante, supoñemos que X e Y están correlacionadas.

O estimador da razón permitirá estimar o total ou a media dunha variable cunha precisión mellor que a da mostraxe aleatoria simple.

Estimadores de razón: motivación

Supoñamos que $X = \sum_{i=1}^{N} X_i$ denota o total da poboación da cidade dos N barrios de Vigo no ano 2016 e $Y = \sum_{i=1}^{N} Y_i$ o total da poboación na mesma cidade no ano 2019.

Dacordo co INE, Vigo conta cunha poboación de $X = 292826$ persoas no ano 2016 e $Y = 295323$ persoas no 2019.

Desta maneira, a ratio

$$R = \frac{Y}{X} = \frac{295323}{292826} = 1.008527$$

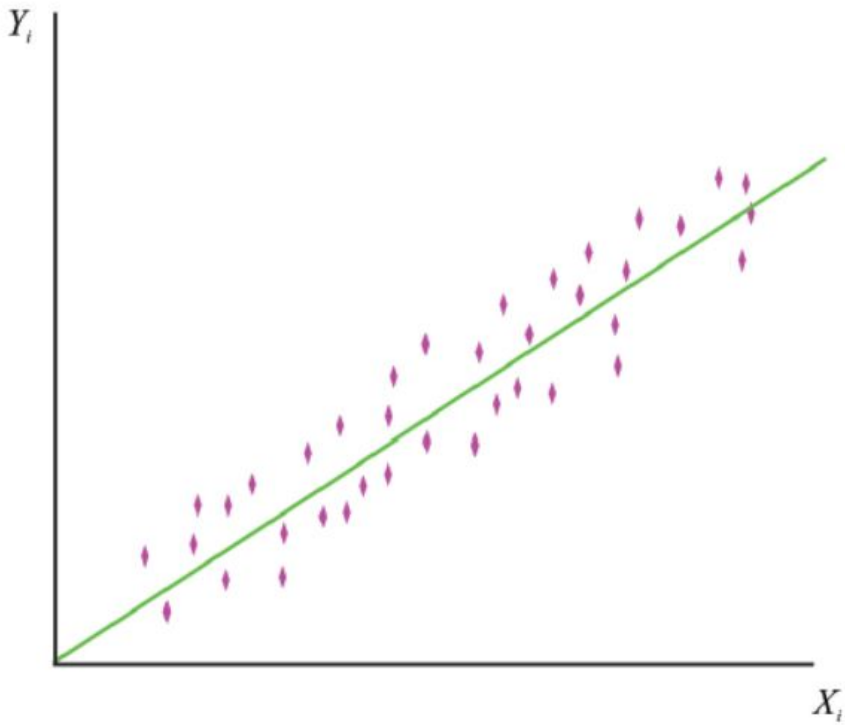
denota o crecemento da poboación de Vigo. En concreto, o volume de poboación aumentou nun 0.85 %.

Ese valor de R é que denominaremos por razón. Pódese aplicar sobre os totais, sobre medias, ou sobre calquera magnitude que desexemos comparar en dous instantes de tempo.

Caso 1. Estimadores de Razón (con medias coñecidas)

Estimadores de Razón (con medias coñecidas)

Sexan (X, Y) o par de variables baixo estudo, entre as cales é posible a existencia de relacións (de tipo lineal, por exemplo).



No que segue, estimamos Y (ou $\bar{Y}$), co valor de X (ou $\bar{X}$) coñecido. Además, X e Y son características dos individuos dunha poboación de N membros.

Estimadores de Razón (con medias coñecidas)

A razón

Defínese a razón entre X e Y como o valor $R > 0$ tal que:

$$R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}$$

de tal maneira que

$$Y = R \cdot X \text{ ou } \bar{Y} = R \cdot \bar{X}.$$

Un estimador da razón baseado nunha mostra aleatoria simple de n elementos é:

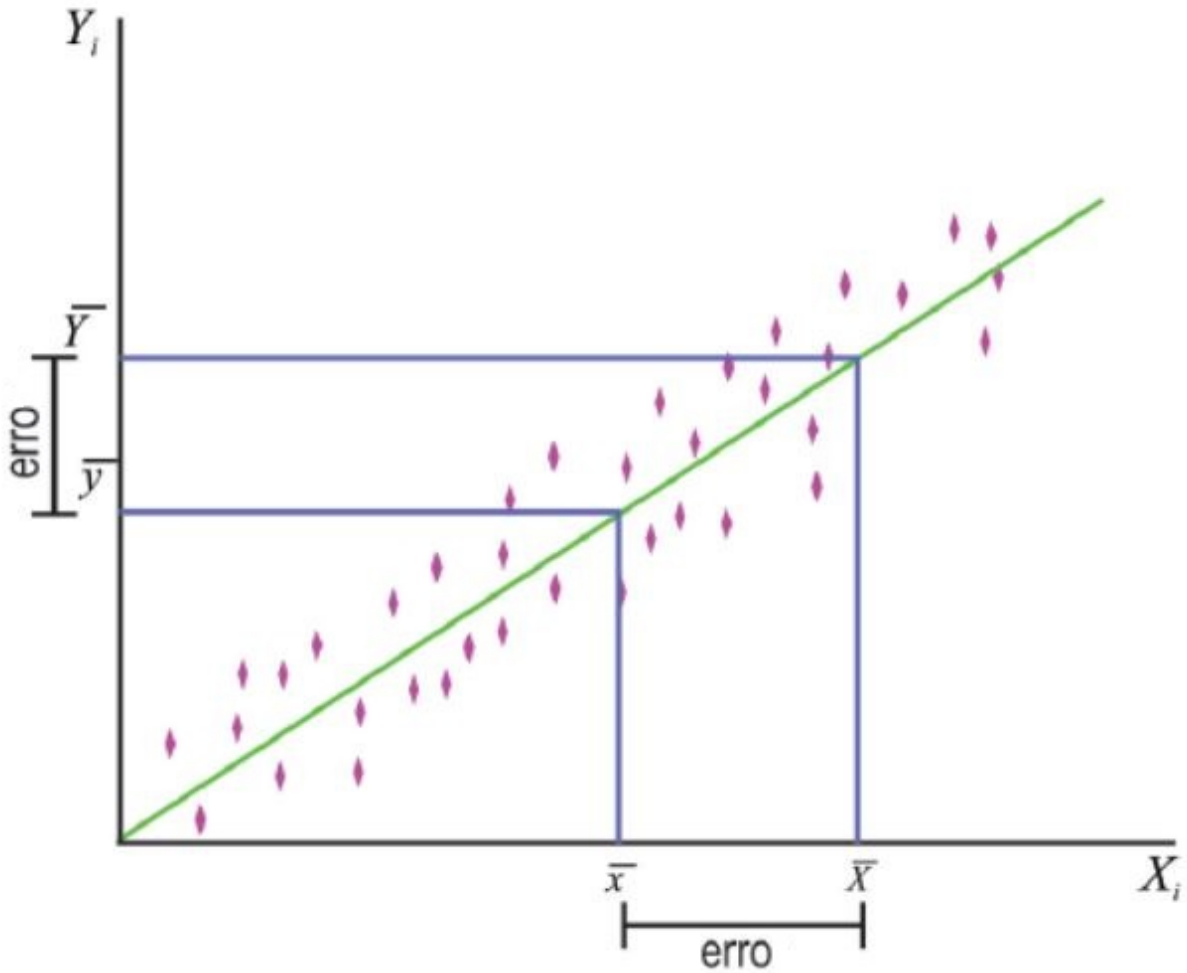
$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$$

cumprindo que

$$\hat{Y}_R = \hat{R} \cdot X \text{ ou } \bar{y}_R = \hat{R} \cdot \bar{X}.$$

(Nota: $\hat{Y}_R$ e $\bar{y}_R$ son as estimacións do total e da media usando a razón)

Como se mide o erro co estimador da razón?



Como se mide o erro co estimador da razón?

Baixo o suposto da proporcionalidade,

$$Y_i \cong RX_i$$

para cada $i = 1, \ldots, N$.

- Desta forma $Y_i = RX_i + \varepsilon_i$, para todo $i = 1, \ldots, N$ e sendo ε_i o erro cometido.

Como coñecemos $\bar{X}$, a taxa de erro é $\frac{\bar{X}}{\bar{x}}$, entón:

$$\bar{y}_R = \hat{R}\bar{X} = \frac{\bar{y}}{\bar{x}}\bar{X} = \bar{y}\left(\frac{\bar{X}}{\bar{x}}\right)$$

Exemplo. Se $\frac{\bar{X}}{\bar{x}} = 1.2$, sabemos que $\bar{X} = 1.2\bar{x}$. É dicir, subestimamos a media ($\bar{x} < \bar{X}$).

Exemplo: estimación do total

Supoñamos que se ten unha m.a.s. de 49 cidades dun total de 196 de certas rexións do país, da que se coñece o número de habitantes (en miles) no ano 2010, coa mostra $\{x_1, \ldots, x_{49}\}$ e no ano 2019 a través da mostra $\{y_1, \ldots, y_{49}\}$.

Quérese estimar o total de habitantes na rexión no 2019, coñecendo o total de habitantes en 2010 ($X = 22919$).

Exemplo: estimación do total

Supoñamos que se ten unha m.a.s. de 49 cidades dun total de 196 de certas rexións do país, da que se coñece o número de habitantes (en miles) no ano 2010, coa mostra $\{x_1, \ldots, x_{49}\}$ e no ano 2019 a través da mostra $\{y_1, \ldots, y_{49}\}$.

Quérese estimar o total de habitantes na rexión no 2019, coñecendo o total de habitantes en 2010 ($X = 22919$).

Solución. Temos que:

$$\sum_{i=1}^{49} x_i = 5054 \text{ e } \sum_{i=1}^{49} y_i = 6262.$$

O estimador do total baixo a razón é:

$$\hat{Y}_R = \hat{R}X = \frac{\sum_{i=1}^{49} y_i}{\sum_{i=1}^{49} x_i} X = \frac{6262}{5054} \cdot 22919 = 28397.$$

Baixo m.a.s., $\hat{Y}_{m.a.s.} = N\bar{y} = 196 \left(\frac{6262}{49}\right) = 25048$ e o total real no 2019 é 29351.

Exemplo: estimación do total (continuación)

Quérese estimar a media de habitantes na rexión no 2019, coñecendo o total de habitantes en 2010 ($X = 22919$).

Solución. O estimador da media baixo a razón é:

$$\bar{y}_R = \hat{R}\bar{X} = \frac{\sum_{i=1}^{49} y_i}{\sum_{i=1}^{49} x_i}\bar{X} = \frac{6262}{5054} \cdot \frac{22919}{196} = 144,883.$$

Baixo m.a.s., $\bar{y}_{m.a.s.} = \left(\frac{6262}{49}\right) = 127.7959$.

Se queremos estimar o total da característica dada por Y na poboación, adoita non coñecerse N (o total de elementos da poboación) para usalo no estimador usual $\hat{Y} = N\bar{y}$.

Sen embargo, temos que

$$\frac{X}{\bar{X}} = \frac{X}{X/N} = N.$$

Desta maneira, como $\bar{x}$ é a estimación de $\bar{X}$, estimamos N por

$$\hat{N} = \frac{X}{\bar{x}},$$

quedando

$$\hat{Y}_R = \hat{R}X = \frac{\bar{y}}{\bar{x}}X = \bar{y}\left(\frac{X}{\bar{x}}\right).$$

Caso 2. Estimadores de Razón (con medias descoñecidas)

Estimadores de Razón (con medias descoñecidas)

Un estimador da razón baseado nunha mostra aleatoria simple de n elementos é:

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}.$$

Exemplos. Nunha enquisa de familias, mídese o ingreso total familiar (y_i) e o número de membros da familia (x_i). Entón, $\hat{R}$ denota a estimación do ingreso per cápita.

Exemplos deste tipo surxen cando a unidade de mostraxe (no exemplo, a familia), comprende un conxunto de elementos (membros da familia) e o noso interese é estimar a media por elemento.

Tamén son aplicables na estimación da proporción de certa característica en relación ó total de tódalas características.

Exemplo. Quérese estudar a intención de voto nunha enquisa. Entón:

$$\text{votos al partido } q = \frac{\text{total de votos ó partido } q}{\text{total de votos}}$$

onde

$$\text{total de votos} = \text{votos ó partido 1} + \text{votos ó partido 2} + \cdots$$

Estimadores de Razón

O estimador da razón poboacional R para unha mostra de tamaño n é:

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}.$$

- É un estimador sesgado.
- A varianza estimada é

$$\hat{V}(\hat{R}) = \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{x}^2} \sum_{i=1}^{n} \frac{(y_i - \hat{R}x_i)^2}{n-1}$$

O intervalo de confianza para R, cunha confianza do $(1-\alpha)$, é:

$$IC_R = \left(\hat{R} - z_{1-\alpha/2}\sqrt{\hat{V}(\hat{R})}, \hat{R} + z_{1-\alpha/2}\sqrt{\hat{V}(\hat{R})}\right)$$

O estimador da total poboacional Y para unha mostra de tamaño n é:

$$\hat{Y}_R = \hat{R}X.$$

- A varianza estimada é

$$\hat{V}(\hat{Y}_R) = N^2 \left[\left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{R}x_i)^2}{n-1} \right]$$

O intervalo de confianza para Y, cunha confianza do $(1-\alpha)$, é:

$$IC_Y = \left(\hat{Y}_R - z_{1-\alpha/2}\sqrt{\hat{V}(\hat{Y}_R)}, \hat{Y}_R + z_{1-\alpha/2}\sqrt{\hat{V}(\hat{Y}_R)} \right)$$

O estimador da media poboacional Y para unha mostra de tamaño n é:

$$\bar{y}_R = \hat{R}\bar{X}.$$

- A varianza estimada é

$$\hat{V}(\bar{y}_R) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{R}x_i)^2}{n-1}$$

O intervalo de confianza para $\bar{Y}$, cunha confianza do $(1-\alpha)$, é:

$$IC_{\bar{Y}} = \left(\bar{y}_R - z_{1-\alpha/2}\sqrt{\hat{V}(\bar{y}_R)}, \bar{y}_R + z_{1-\alpha/2}\sqrt{\hat{V}(\bar{y}_R)}\right)$$

Determinación do tamaño de mostra

Analizamos o caso da estimación da media $\bar{Y}$ mediante estimadores de razón.

O intervalo de confianza para $\bar{Y}$ permite controlar o erro na súa estimación:

$$P(|\bar{y}_R - \bar{Y}| < \varepsilon) \geq 1 - \alpha.$$

Pódese comprobar que o tamaño de mostra axeitado, se N é grande, é

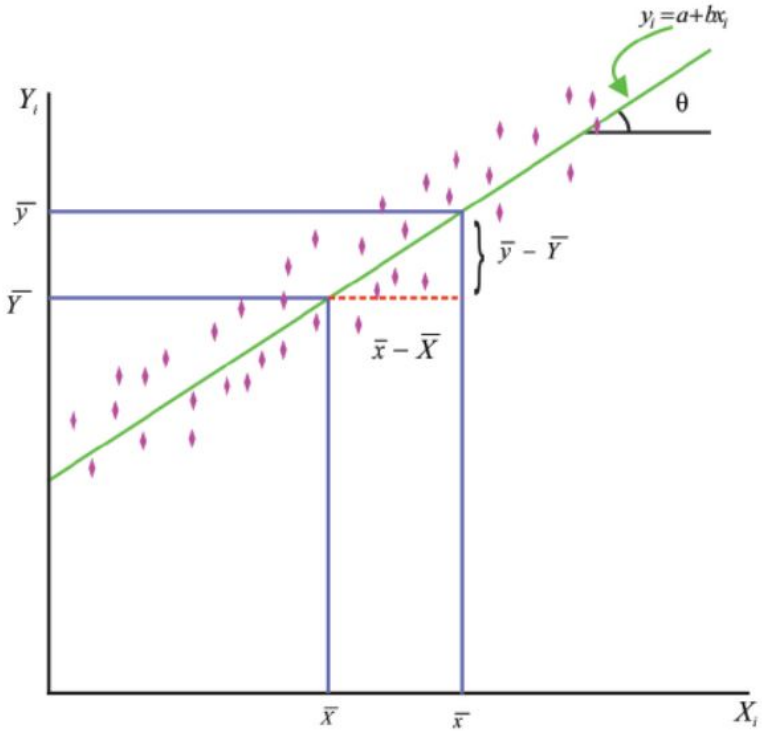
$$n = \frac{z^2_{1-\alpha/2} S^2}{\varepsilon^2}$$

onde S^2 debe ser aproximado por $S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2$.

Estimadores de Regresión (baixo m.a.s.)

Estimadores de Regresión

Sexan (X, Y) o par de variables baixo estudo. Tal e como se mencionou, é posible a existencia de relacións.



Asúmese que Y é a variable de interese e X é a variable auxiliar.

Na análise das rectas de regresión, o parámetro b é o fundamental. Desde unha perspectiva xeométrica:

$$b = tan\ \theta = \frac{\text{cateto oposto}}{\text{cateto adxacente}} = \frac{\bar{y} - \bar{Y}}{\bar{x} - \bar{X}}$$

Estimadores de Regresión

O estimador de regresión da media poboacional é:

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

O estimador de regresión do total poboacional é:

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

$$\hat{Y}_{reg} = \hat{Y}_{mas} + \hat{b}(X - \hat{X}_{mas}).$$

onde $\hat{b} = \frac{S_{XY}}{S_X^2}$, $S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ e $S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Para a análise estatística, a correlación entre X e Y xoga un papel fundamental. Defínese por ρ e, por simplicidade, usaremos a súa estimación na mostra:

$$\rho = \frac{S_{XY}}{S_X S_Y}.$$

O estimador da media poboacional $\bar{Y}$, con $\bar{X}$ coñecido e para unha mostra de tamaño n, é:

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

- A varianza estimada é

$$\hat{V}(\bar{y}_{reg}) = \frac{S_Y^2}{n}(1 - \rho^2)\left(1 - \frac{n}{N}\right)$$

O intervalo de confianza do $(1 - \alpha) \times 100\,\%$ para $\bar{Y}$ é:

$$IC_{\bar{Y}} = \left(\bar{y}_{reg} - z_{1-\alpha/2}\sqrt{\hat{V}(\bar{y}_{reg})}, \bar{y}_{reg} + z_{1-\alpha/2}\sqrt{\hat{V}(\bar{y}_{reg})}\right)$$

O estimador do total poboacional Y, con X coñecido e para unha mostra de tamaño n, é:

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

- A varianza estimada é

$$\hat{V}(\hat{Y}_{reg}) = N^2 \frac{S_Y^2}{n}(1 - \rho^2)\left(1 - \frac{n}{N}\right)$$

Se n é o suficientemente grande, o intervalo de confianza do $(1 - \alpha) \times 100\,\%$ para Y é:

$$IC_Y = \left(\hat{Y}_{reg} - z_{1-\alpha/2}\sqrt{\hat{V}(\hat{Y}_{reg})}, \hat{Y}_{reg} + z_{1-\alpha/2}\sqrt{\hat{V}(\hat{Y}_{reg})}\right)$$

Exemplo: estimación do total

Examinouse a 486 candidatos ó ingresar nunha facultade e quérese avaliar a súa nota en Estatística. Destos tomouse unha m.a.s. de 10 estudantes ós que lles mediu a súa cualificación (sobre 100) en Estatística ó final do primeiro cuadimestre.
Sábese que $\bar{X} = 52$ para os 486 estudantes ó inicio de curso e deséxase estimar $\bar{Y}$, o promedio da calificación ó final do cuadrimestre.

Estudante	1	2	3	4	5	6	7	8	9	10
Inicio (X)	39	43	21	64	57	47	28	75	34	52
Final (Y)	65	78	52	82	92	89	73	98	56	75

Solución. Sabemos que:

- $N = 486$, $\bar{X} = 52$ e $\bar{x} = 46$.
- $n = 10$, $\bar{y} = 76$ e $S_Y^2 = 228.45$.
- $\hat{b} = 0.766$ e $\rho = 0.84$.

Exercicio. Comproba que, dacordo coas fórmulas proporcionadas, os resultados anteriores son correctos.

Exemplo: estimación do total (continuación)

Desta forma,

$$\bar{y}_{reg} = \bar{y} + \hat{b}(\bar{X} - \bar{x}) = 76 + 0.766(52 - 46) = 80.596$$

e

$$\hat{V}(\bar{y}_{reg}) = \frac{S_Y^2}{n}(1 - \rho^2)\left(1 - \frac{n}{N}\right)$$
$$= \frac{228.44}{10}(1 - 0.84^2)\left(1 - \frac{10}{486}\right) = 6.586$$

Facendo a aproximación á normalidad, o intervalo de confianza do 95 % para $\bar{Y}$ é

$$IC_{\bar{Y}} = \left(\bar{y}_{reg} - z_{1-\alpha/2}\sqrt{\hat{V}(\bar{y}_{reg})}, \bar{y}_{reg} + z_{1-\alpha/2}\sqrt{\hat{V}(\bar{y}_{reg})}\right)$$

Substituíndo na fórmula anterior

$$IC_{\bar{Y}} = (80.596 - 1.96\sqrt{6.586}, 80.596 + 1.96\sqrt{6.586}) = (75.567, 85.625)$$

Determinación do tamaño de mostra

Analizamos o caso da estimación da media $\bar{Y}$ usando regresión.

O intervalo de confianza para $\bar{Y}$ permite controlar o erro na súa estimación:

$$P(|\bar{y}_{reg} - \bar{Y}| < \varepsilon) \geq 1 - \alpha.$$

Pódese comprobar que o tamaño de mostra axeitado, se N é grande, é

$$n = \frac{z^2_{1-\alpha/2} S^2_Y (1 - \rho^2)}{\varepsilon^2}.$$

Se ρ (a correlación) é o suficientemente grande, n é pequeno.

É xeralmente máis pequeno que o asociado a mostraxe aleatoria simple, xa que

$$n_{mas} = \frac{z^2_{1-\alpha/2} S^2_Y}{\varepsilon^2} \quad \text{e} \quad 0 < 1 - \rho^2 < 1.$$