

Количничко оцењивање

Из популације обима N бира се (прост) случајан узорак без понављања $\mathcal{S}$ обима n . Нека су Y и X , редом, главно и помоћно обележје, а y_k и x_k одговарајуће вредности обележја Y , односно X , ентитета означеног са k . Приступ приликом закључивања заснован је на методу одабира узорка.

Као тачкаста оцена непознатог популацијског количника B предлаже се узорачки количник

$$b := \hat{B} = \frac{\sum_{k \in \mathcal{S}} y_k}{\sum_{k \in \mathcal{S}} x_k} = \frac{\bar{Y}}{\bar{X}}.$$

Иако су узорачке средње вредности $\bar{Y}$ и $\bar{X}$ непристрасне оцене одговарајућих непознатих популацијских средњих вредности m_Y и m_X , узорачки количник **није** непристрасна оцена параметра B . Може се одредити и тачан израз за **пристрасност оцене b** , на основу следећег:

$$\text{cov}(b, \bar{X}) = E(b\bar{X}) - Eb \cdot E\bar{X} = E\bar{Y} - Eb \cdot E\bar{X} = m_Y - Eb \cdot m_X,$$

па је

$$Eb - B = \frac{m_Y - \text{cov}(b, \bar{X})}{m_X} - \frac{m_Y}{m_X} = \frac{-\text{cov}(b, \bar{X})}{m_X}. \quad (1)$$

Затим се може добити следећа процена пристрасности:

$$|Eb - B| = \left| \frac{-\text{cov}(b, \bar{X})}{m_X} \right| \leq \frac{\sqrt{Db} \cdot \sqrt{D\bar{X}}}{|m_X|} = \sqrt{Db} \cdot |CV(\bar{X})|.$$

Коришћено је да је коефицијент корелације $\rho_{Z,W} := \frac{\text{cov}(Z,W)}{\sqrt{DZ}\sqrt{DW}}$ двеју случајних величина Z и W по апсолутној вредности мањи или једнак 1.

Одатле следи закључак да је пристрасност оцене b популацијског количника мала у поређењу са њеном стандардном грешком ако је коефицијент варијације сл. величине $\bar{X}$ мали. Код случајног узорка без понављања важи:

$$|CV(\bar{X})| \leq \frac{\sqrt{\sigma_X^2}}{\sqrt{n}|m_X|},$$

тако да коефицијент варијације $CV(\bar{X})$ опада (по апсолутној вредности) са повећањем обима узорка.

Према томе, b није непристрасна тачкаста оцена непознатог популацијског количника B , али се може показати да је квадрат њене пристрасности занемарљиво мали у односу на дисперзију за довољно велике узорке.

Једнакост (1) даје **тачну** формулу за рачунање пристрасности оцене b , али њу обично није једноставно примењивати на конкретне податке добијене у истраживању. Зато се приступа одређивању тзв. линеарне апроксимације (првог реда) пристрасности и дисперзије (онда и средње квадратне грешке) оцене b , која почива на дисперзијама и коваријацији сл. величина $\bar{Y}$ и $\bar{X}$.

Примењује се линеарна апроксимација (нелинеарне) функције две променљиве $g(x, y)$, која се добија развојем у Тејлоров ред ф-је g у околини тачке (x_0, y_0) :

$$g(x, y) = \sum_{n=0}^{+\infty} \left\{ \frac{1}{n!} \sum_{k=0}^n \binom{n}{k} \frac{\partial^n g}{\partial x^{n-k} \partial y^k} \Big|_{(x,y)=(x_0,y_0)} (x-x_0)^{n-k} (y-y_0)^k \right\}$$

за "мале" вредности $|x-x_0|$ и $|y-y_0|$.

Функција g две променљиве, која се појављује у вези са количничким оцењивањем, дефинисана је са:

$$g(x, y) = \frac{y}{x},$$

а апроксимира се развијањем, на описани начин, око тачке (m_X, m_Y) . Користи се апроксимација првог реда ($n=1$), којом се добија:

$$b = g(\bar{X}, \bar{Y}) \approx \frac{m_Y}{m_X} - \frac{m_Y}{m_X^2} (\bar{X} - m_X) + \frac{1}{m_X} (\bar{Y} - m_Y)$$

$$= B - \frac{m_Y}{m_X^2} (\bar{X} - m_X) + \frac{1}{m_X} (\bar{Y} - m_Y) = B + \frac{1}{m_X} (\bar{Y} - B\bar{X}). \quad (2)$$

Дакле, код случајног узорковања без понављања за узорачки количник, до на апроксимацију првог реда, важи: $Eb \approx B$. Апроксимација дисперзије, а самим тим и средње квадратне грешке, оцене b дата је са:

$$\begin{aligned} Db &\approx \left(\frac{m_Y}{m_X^2} \right)^2 D\bar{X} + \left(\frac{1}{m_X} \right)^2 D\bar{Y} - 2 \cdot \frac{m_Y}{m_X^3} \text{cov}(\bar{X}, \bar{Y}) \\ &= \frac{1}{m_X^2} D(\bar{Y} - B\bar{X}). \end{aligned} \quad (3)$$

Већ је познато (час 2, Лема 1) да је:

$$D\bar{X} = \frac{\sigma_X^2}{n} \left(1 - \frac{n}{N} \right), \quad D\bar{Y} = \frac{\sigma_Y^2}{n} \left(1 - \frac{n}{N} \right),$$

а може се показати да је:

$$\text{cov}(\bar{X}, \bar{Y}) = \frac{1}{n} \left(1 - \frac{n}{N} \right) \sum_{k=1}^N \frac{(x_k - m_X)(y_k - m_Y)}{N-1}.$$

$\sum_{k=1}^N \frac{(x_k - m_X)(y_k - m_Y)}{N-1}$ представља (непознату) популацијску коваријацију главног и помоћног обележја

Према томе, апроксимирана средње квадратна грешка оцене b биће мала онда када:

- је обим узорка n велики
- је разломак узорковања $\frac{n}{N}$ велики
- су одступања $y_k - Bx_k$, по апсолутној вредности, мала
- је $\rho_{X,Y} := \frac{\sum_{k=1}^N (x_k - m_X)(y_k - m_Y)}{(N-1)\sigma_X\sigma_Y}$ – популацијски коефицијент (линеарне) корелације близак вредности +1 онда када је B позитивно, односно близак вредности -1 онда када је B негативно.

Важно је приметити да се израз из једнакости (3) може записати и на следећи начин:

$$\frac{1}{m_X^2} E(\bar{Y} - B\bar{X})^2 = \frac{1}{m_X^2} D \left(\frac{1}{n} \sum_{k \in S} (y_k - Bx_k) \right),$$

где је дисперзија са десне стране заправо дисперзија узорачке средње вредности обележја које ћемо назвати "одступање", и чија је вредност регистрована на јединици означеној са k , заправо, одступање

$d_k := y_k - Bx_k$. Важи: $\sum_{k=1}^N d_k = 0$. Како одступања зависе од непознатог популацијског количника B , уводе се нове сл. величине e_k , дате са:

$$e_k = y_k - Bx_k, \quad k = \overline{1, N},$$

које представљају резидуале (остатке) – случајне грешке које настају услед 'фитовања' праве $y = bx$ вредностима главног и помоћног обележја.

Као тачкаста оцена дисперзије оцене b предлаже се статистика:

$$\frac{1}{m_X^2} \cdot \frac{\bar{S}_e^2}{n} \left(1 - \frac{n}{N} \right),$$

где је $\bar{S}_e^2 = \frac{1}{n-1} \sum_{k \in S} e_k^2$ узорачка дисперзија резидуала.

Регресионо оцењивање

Из популације обима N бира се (прост) случајан узорак без понављања $\mathcal{S}$ обима n . Нека су Y и X , редом, главно и помоћно обележје, а y_k и x_k одговарајуће вредности обележја Y , односно X , ентитета означеног са k . Приступ приликом закључивања заснован је на методу одабира узорка.

Лема 1. Статистика $\hat{m}_Y^{lr}$ дата са:

$$\hat{m}_Y^{lr} = \bar{Y} + b_1 \cdot (m_X - \bar{X}),$$

где је b_1 унапред изабрана константа, је непристрасна оцена непознате популацијске средње вредности m_Y . Дисперзија ове оцено је:

$$\frac{\sigma_d^2}{n} \left(1 - \frac{n}{N}\right), \quad (4)$$

где је

$$\sigma_d^2 = \frac{1}{N-1} \sum_{k=1}^N (y_k - (m_Y + b_1 \cdot (x_k - m_X)))^2.$$

Доказ.

Користећи својства узорачке средње вредности главног, односно помоћног, обележја лако се добија:

$$E\hat{m}_Y^{lr} = E\bar{Y} + b_1 \cdot (m_X - E\bar{X}) = m_Y + b_1 \cdot 0 = m_Y.$$

Приметити да се сл. величина $\hat{m}_Y^{lr}$ може записати и на следећи начин:

$$\hat{m}_Y^{lr} = \frac{1}{n} \sum_{k \in \mathcal{S}} (y_k - b_1(x_k - m_X)),$$

па се најједноставнији начин за одређивање њене дисперзије своди на то да се $\hat{m}_Y^{lr}$ посматра као узорачка средина новог обележја чија је вредност регистрована на јединици означеној са k једнака $y_k - b_1(x_k - m_X)$ (популацијска средња вредност овог новог обележја је: m_Y). Одатле директно следи једнакост (4), која је еквивалентна изразу:

$$\frac{1}{n} \left(1 - \frac{n}{N}\right) \left(\sigma_Y^2 - 2b_1 \cdot \sum_{k=1}^N \frac{(y_k - m_Y)(x_k - m_X)}{N-1} + b_1^2 \cdot \sigma_X^2 \right),$$

σ_Y^2 и σ_X^2 су непознате популацијске дисперзије, редом, главног и помоћног обележја. ■

Лема 2. Вредност b_1 која минимизује $D\hat{m}_Y^{lr}$ је:

$$b_1 = B_1 = \frac{\sum_{k=1}^N (y_k - m_Y)(x_k - m_X)}{\sum_{k=1}^N (x_k - m_X)^2} = \rho_{X,Y} \cdot \frac{\sigma_Y}{\sigma_X},$$

где је $\rho_{X,Y}$ непознат популацијски коефицијент корелације главног и помоћног обележја. Минимална дисперзија која се тада добија једнака је:

$$\frac{\sigma_Y^2}{n} \left(1 - \frac{n}{N}\right) (1 - \rho_{X,Y}^2).$$

Ако парови вредности обележја Y и X указују на присуство приближно линеарне везе:

$$y = B_0 + B_1 x,$$

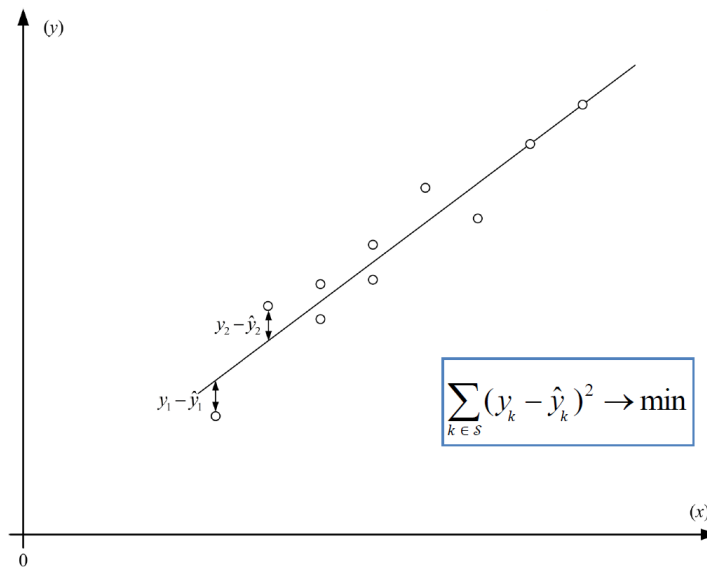
методом најмањих квадрата лако се добијају тачкасте оцено $b_0 = \hat{B}_0$ и $b_1 = \hat{B}_1$, редом, одсечка ординате и линеарног регресионог коефицијента.

Метод најмањих квадрата састоји се у томе да се непознати параметри B_0 и B_1 оцено тако да збир квадрата одступања вредности y_k од њене "процењене" вредности $\hat{y}_k := B_0 + B_1 x_k$ буде најмањи, тј:

$$(b_0, b_1) = \arg \min_{(B_0, B_1) \in \mathbb{R}^2} \sum_{k \in \mathcal{S}} (y_k - (B_0 + B_1 x_k))^2.$$

Коришћењем извода, на следећи начин лако се добијају оцено из узорка:

$$\begin{aligned}
 SSR &:= \sum_{k \in S} (y_k - (B_0 + B_1 x_k))^2 = \sum_{k \in S} (y_k^2 - 2y_k(B_0 + B_1 x_k) + (B_0 + B_1 x_k)^2) \\
 \frac{\partial SSR}{\partial B_0} &= -2 \sum_{k \in S} y_k + 2 \sum_{k \in S} (B_0 + B_1 x_k) & \sum_{k \in S} (-y_k + b_0 + b_1 x_k) &= 0 \\
 & & \Rightarrow b_0 &= \bar{Y} - b_1 \cdot \bar{X} \\
 \frac{\partial SSR}{\partial B_1} &= -2 \sum_{k \in S} y_k x_k + 2 \sum_{k \in S} (B_0 + B_1 x_k) x_k & \sum_{k \in S} (-y_k + b_0 + b_1 x_k) x_k &= 0 \\
 & & \Rightarrow b_1 &= \frac{\sum_{k \in S} x_k y_k - \bar{Y} \cdot \sum_{k \in S} x_k}{\sum_{k \in S} x_k^2 - \left(\sum_{k \in S} x_k \right)^2 / n}
 \end{aligned}$$



Статистика $\hat{m}_Y^{lr}$ дата са:

$$\hat{m}_Y^{lr} = b_0 + b_1 \cdot m_X,$$

где су b_1 и b_0 предложене тачкасте оцено, није непристрасна:

$$\begin{aligned}
 E\hat{m}_Y^{lr} - m_Y &= Eb_0 + Eb_1 \cdot m_X - m_Y = E\bar{Y} - E(b_1 \cdot \bar{X}) + Eb_1 \cdot m_X - m_Y \\
 &= -(E(b_1 \cdot \bar{X}) - Eb_1 \cdot E\bar{X}) = -\text{cov}(b_1, \bar{X}).
 \end{aligned}$$

Метод коришћен код количничког оцењивања за апроксимацију дисперзије може се применити, на аналоган начин, и код регресионог оцењивања. Њиме се уједно врши и апроксимација средње квадратне грешке, јер је и овде пристрасност занемарљиво мала код великих узорака.

Апроксимирана средње квадратна грешка оцено $\hat{m}_Y^{lr}$ биће мала онда када је:

- обим узорка n велики
- разломак узорковања $\frac{n}{N}$ велики
- стандардна грешка σ_Y мала
- коефицијент корелације $\rho_{X,Y}$ близак вредности $+1$ или -1 .

Као тачкаста оцена дисперзије оцено $\hat{m}_Y^{lr}$ предлаже се статистика:

$$\frac{\bar{S}_e^2}{n} \left(1 - \frac{n}{N}\right),$$

где је $\bar{S}_e^2 = \frac{1}{n-1} \sum_{k \in S} e_k^2$ узорачка дисперзија резидуала $e_k = y_k - (b_0 + b_1 x_k)$.