

Теорија узорака

19. MART '19.

Регресионо оцењивање SRSWOR

Као и количничко оцењивање, регресионо оцењивање ('linear regression estimation') предлага се као техника оцењивања којом се побољшава прецизност оцена непознатих популацијских параметара главног обележја Y коришћењем помоћног обележја X које је у корелацији са Y . Претпоставља се, дакле, да између вредности ова два обележја постоји приближно линеарна веза, али која не пролази кроз координатни почетак. Вредност обележја X се, при томе, може одредити на произвољној јединици у популацији, а морају бити познати и: m_X , τ_X .

За сваку јединицу y (простом) случајном узорку S без понављања обима n из дате популације обима N регистроване су вредности и главног и помоћног обележја.

Регресионо оцењивање

Кључни резултати у вези са регресионим оцењивањем непознате популацијске средње вредности m_Y , када је **приступ заснован на методу** SRSWOR одабира узорка, дати су у наставку:

► тачкаста оцена:

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

где је b_1 оцена промене вредности обележја Y када се вредност обележја X повећа за 1

👉 У највећем броју примера примене регресионог оцењивања b_1 је статистика, тј. функција узорка чија се реализована вредност израчунава на основу реализованог узорка. Међутим, некада је могуће / погодно изабрати вредност b_1 унапред, као детерминистичку вредност – константу

$$b_1 = 1 - \text{'difference estimation'}$$

Регресионо оцењивање (наставак)

- ▶ Ако је b_1 унапред изабрана константа:

тачкаста оцена $\hat{m}_Y^{lr}$	$E\hat{m}_Y^{lr}$	$D\hat{m}_Y^{lr}$	тачкаста оцена $D\hat{m}_Y^{lr}$
$\bar{Y} + b_1 \cdot (m_X - \bar{X})$	m_Y	$\frac{\sigma_d^2}{n} \left(1 - \frac{n}{N}\right)$	$\frac{\bar{S}_e^2}{n} \left(1 - \frac{n}{N}\right)$

где су:

$$\sigma_d^2 := \frac{1}{N-1} \sum_{k=1}^N \left(y_k - (m_Y + b_1 \cdot (x_k - m_X)) \right)^2$$
$$\bar{S}_e^2 := \frac{1}{n-1} \sum_{k \in \mathcal{S}} \left(y_k - (\bar{Y} + b_1 \cdot (x_k - \bar{X})) \right)^2$$

- ▶ Који је „најбољи“ избор константе b_1 ?

Вредност b_1 која минимизује $D\hat{m}_Y^{lr}$, дату у формули из горње табеле, је:

$$b_1 = B_1 = \frac{\sum_{k=1}^N (x_k - m_X)(y_k - m_Y)}{\sum_{k=1}^N (x_k - m_X)^2}$$

и назива се (популацијски) **линеарни регресиони коефицијент** од Y по X

- Ако се b_1 одређује на основу узорка $\mathcal{S}$, најприродније је да b_1 представља оцену (непознатог) параметра B_1 са претходног слајда:

$$b_1 = \hat{B}_1 = \frac{\sum_{k \in \mathcal{S}} (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k \in \mathcal{S}} (x_k - \bar{X})^2}$$

ова оцена добијена је тзв. **МЕТОДОМ**
НАЈМАЊИХ КВАДРАТА

валидна само за „ДОВОЉНО“ велике узорке

тачкаста оцена $\hat{m}_Y^{lr}$	непристрасност	<u>апроксимација</u> дисперзије $D\hat{m}_Y^{lr}$	оцена дисперзије $D\hat{m}_Y^{lr}$
$\bar{Y} + b_1 \cdot (m_X - \bar{X})$ $= \bar{Y} - b_1 \cdot (\bar{X} - m_X)$	$\times$ пристрасност: $-\text{cov}(b_1, \bar{X})$	$\frac{\sigma_d^2}{n} \left(1 - \frac{n}{N}\right)$	$\frac{\bar{S}_e^2}{n} \left(1 - \frac{n}{N}\right)$

где су:

$$\sigma_d^2 := \frac{1}{N-1} \sum_{k=1}^N \left(y_k - (m_Y + B_1 \cdot (x_k - m_X)) \right)^2$$

$$\bar{S}_e^2 := \frac{1}{n-1} \sum_{k \in \mathcal{S}} \left(y_k - (\bar{Y} + b_1 \cdot (x_k - \bar{X})) \right)^2$$

👉 код „мањих“ узорка може се заменити са $n - 2$

Дакле, ако су парови вредности помоћног и главног обележја (подаци су представљени као тачке у правоуглом координатном систему) уједначено „разбацани“ око праве која не пролази нужно кроз координатни почетак:

$$y = B_0 + B_1x,$$

оцењивање применом просте линеарне регресије је корисна техника.

Тачкаста оцена коефицијента правца праве ('slope') B_1 дата је формулом:

$$b_1 = \hat{B}_1 = \frac{\sum_{k \in \mathcal{S}} (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k \in \mathcal{S}} (x_k - \bar{X})^2},$$

а тачкаста оцена одсечка ординате ('intercept') B_0 формулом:

$$b_0 = \hat{B}_0 = \bar{Y} - \hat{B}_1 \cdot \bar{X}.$$

☝ Приметити да је оцена $\hat{m}_Y^{lr}$ непознате популацијске средње вредности m_Y формирана „предвиђањем“ вредности главног обележја користећи фитовану регресиону праву, при чему је вредност помоћног обележја једнака m_X , тј:

$$\hat{m}_Y^{lr} := \hat{B}_0 + \hat{B}_1 \cdot m_X$$

Регресионо оцењивање

Када је приступ заснован на моделу, слично као код количничког оцењивања, природно се намеће **линеарни регресиони модел** као модел популације: $x_1, x_2, \dots, x_N$ су познате константе (које су строго позитивне) а случајне величине $Y_1, Y_2, \dots, Y_N$ су **независне и овде дате са**:

$$Y_k = \beta_0 + \beta_1 x_k + \varepsilon_k,$$

где је $E\varepsilon_k (= E_M \varepsilon_k) = 0$ и $D\varepsilon_k (= D_M \varepsilon_k) = \sigma^2$, а β_0 и β_1 су фиксиране, али непознате вредности.

Нека је са $\mathcal{S}$ означен **реализован** (не)случајан узорак без понављања обима n . Ту имамо:

- оцене методом најмањих квадрата за β_0 и β_1 су:

$$\hat{\beta}_1 = \frac{\sum_{k \in \mathcal{S}} x_k Y_k - \sum_{k \in \mathcal{S}} x_k \cdot \sum_{k \in \mathcal{S}} Y_k / n}{\sum_{k \in \mathcal{S}} x_k^2 - (\sum_{k \in \mathcal{S}} x_k)^2 / n} \quad \text{и} \quad \hat{\beta}_0 = \frac{\sum_{k \in \mathcal{S}} Y_k}{n} - \hat{\beta}_1 \cdot \frac{\sum_{k \in \mathcal{S}} x_k}{n}$$

- непристрасан (у односу на модел) линеарни предиктор случајне величине

$$\frac{1}{N} \sum_{k=1}^N Y_k$$

дат је формулом:

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot m_X$$

Регресионо оцењивање – узорак са неједнаким вероватноћама

Уопштена регресиона оцена непознатог популацијског тотала τ_Y , када је приступ заснован на методу случајног узорка са понављањем, односно без понављања, и неједнаким вероватноћама укључења, дата је са:

$$\hat{t}_G = \hat{t}_Y + b_1 \cdot (\tau_X - \hat{t}_X),$$

где су $\hat{t}_Y$ и $\hat{t}_X$ Horvitz-Thompson-ове оцене, редом, τ_Y и τ_X , тј.

$$\hat{t}_Y = \sum_{k \in S'} \frac{y_k}{\pi_k} \quad \text{и} \quad \hat{t}_X = \sum_{k \in S'} \frac{x_k}{\pi_k},$$

а b_1 је тежинска оцена линеарног регресионог коефицијента:

$$b_1 = \frac{\sum_{k \in S'} \frac{1}{\pi_k} \left(x_k - \frac{\hat{t}_X}{\hat{N}_\pi} \right) \left(y_k - \frac{\hat{t}_Y}{\hat{N}_\pi} \right)}{\sum_{k \in S'} \frac{1}{\pi_k} \left(x_k - \frac{\hat{t}_X}{\hat{N}_\pi} \right)^2}$$

$\hat{N}_\pi = \sum_{k \in S'} \frac{1}{\pi_k}$ је Horvitz-Thompson-ова оцена обима популације N , S' је редуковани узорак, који садржи само различите јединице из оригиналног узорка S , а π_k су (познате) вероватноће укључења првог реда.

оцена средине:

$$\hat{m}_G = \hat{m}_Y + b_1 \cdot (m_X - \hat{m}_X),$$

где су:

$$\hat{m}_Y = \frac{\hat{t}_Y}{\hat{N}_\pi} \quad \text{и} \quad \hat{m}_X = \frac{\hat{t}_X}{\hat{N}_\pi}$$

► Пример – регресионо оцењивање – SRSWOR

Дати су подаци о укупном броју грађана са правом гласа за сваку од 299 изборних јединица, у 16 немачких федералних држава, и резултатима избора за немачки Бундестаг 2002. и 2005. г.

Резултати гласања дати су у облику процената „других“ гласова (сваки гласач на изборима за Бундестаг има два гласа – „првим“ гласом гласачи бирају своје локалне представнике у Бундестагу (директни мандати), а „други“ глас је, заправо, за страначку листу) за сваку од странака, који одређују и број посланичких места у парламенту. Проценти су израчунати дељењем броја гласова за странку са бројем важећих гласова.

оквир – база података *

$n = 45$, узорак се бира методом SRSWOR коришћењем програмског генератора  јединице узорковања су изборне јединице

главно обележје Y – проценат у корист Социјалдемократске партије Немачке 2005. г.

помоћно обележје X – проценат у корист Социјалдемократске партије Немачке 2002. г.

* подаци се налазе у бази election у оквиру  пакета samplingbook

► Пример – регресионо оцењивање – SRSWOR – наставак

На основу узорка оцењује се непозната средња вредност процената гласова, које је освојила Социјалдемократска партија Немачке (Sozialdemokratische Partei Deutschlands) на изборима 2005. г. Сматра се да је познат цензус за 2002. г.

Узорачка средина главног обележја је: $\bar{Y} = 34.69988\%$

Оцењена стандардна грешка статистике $\bar{Y}$ је: 1.167303%

Узорачка средина помоћног обележја је: $\bar{X} = 38.83385\%$

Популацијска средина помоћног обележја је: $m_X = 38.61344\%$

Оцењена вредност

- линеарног регресионог коефицијента је: $b_1 = \hat{B}_1 = 0.8646685$
- одсечка ординате је: $b_0 = \hat{B}_0 = 1.121475\%$
- популацијске средине је: $\hat{m}_Y^{lr} = 34.5093\%$

Оцењена вредност стандардне грешке оцено $\hat{m}_Y^{lr}$ је: 0.3698428%

Тачна вредност популацијске средине је: $m_Y = 34.26949\%$

