ԳԾԱՅԻՆ ՌԵԳՐԵՍԻԱ

07/04/2016

1 ՆԵՐԱԾՈՒԹՅՈՒՆ

Վիճակագրության հիմնական նպատակն է վերլուծել տարբեր աղբյուրներից ստացված տվյալները: Վերլուծության արդյունքում պետք է տվյալների միջից տարանջատել (օգտակար) կախվածությունները՝ օրինաչափությունները, և (անօգուտ) աղմուկը:

Տվյալները հիմնականում ներկայացված են հետևյալ տեսքով: Պատահական ձևով ընտրված n անհատի վրա կատարված են p չափումներ: Օրինակ՝ աշխարհի տարբեր երկրների համար չափված են համախառն ներքին արդյունքը, կրթվածության մակարդակը, մեկ բնակչին բաժին հասնող երեխաների քանակը, կանանց միջին կյանքի տևողությունը, և այլն: Ստացված է մի աղյուսակ

$$
\begin{array}{l}
x_1^1, \dots, x_1^p \\
x_2^1, \dots, x_2^p \\
\vdots \\
x_n^1, \dots, x_n^p
\end{array}
$$

Այս աղյուսակի ամեն տող ներկայացնում է մեկ անհատ (մեր օրինակում՝ երկիր), իսկ ամեն սյունակ՝ մեկ հատկանիշ (մեր օրինակում՝ ՀՆԱ, կրթ. մակարդակ, և այլն):

Ենթադրենք նույն անհատների վրա կատարվել է ևս մեկ չափում, որը նշանակենք $y_1, \dots, y_n$: Մեր օրինակում դա կարող է լինել միջին աշխատավարձը:

Ռեգրեսիոն անալիզի նպատակն է՝ բացահայտել

2) կապը այս նոր հատկանիշի՝ y-ի և $x^1,\dots,x^p$ հատկանիշների միջև: Սովորաբար կիրառվում են հետևյալ եզրույթները՝

$x^1,\dots,x^p$ – բացատրող փոփոխականներ

y – բացատրվողի կամ բացատրության ենթակա փոփոխական:

Խնդիրն է՝ կառուցել մի ֆունկցիա

$$f : (x^1,\dots,x^p) \longmapsto z \in \mathbb{R}$$

այնպիսին որ $f(x^1,\dots,x^p)$-ն մոտ է y-ին բոլոր երկրների համար:

2) ՄԱԹԵՄԱՏԻԿԱԿԱՆ ՄՈԴԵԼ

Մաթ. մոդելը սահմանելու համար ներմուծվում են վեկտորական նշանակումներ՝

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} x_1^1 & \dots & x_1^p \\ \vdots & \ddots & \vdots \\ x_n^1 & \dots & x_n^p \end{bmatrix}$$

Գծային ռեգրեսիայի դեպքում ցանկանում ենք, որ ամեն y_i մոտարկվի որպես ինչ-որ գծային ֆունկցիա $x_i^1,\dots,x_i^p$-ից, այսինքն

$$y_i = \sum_{j=1}^{p} \beta_j^* x_i^j + \xi_i \qquad i = 1,\dots,n$$

Այստեղ $\beta^* = \begin{bmatrix} \beta_1^* \\ \vdots \\ \beta_p^* \end{bmatrix}$ անհայտ վեկտոր է $\mathbb{R}^p$-ից:

$(\xi_1,\dots,\xi_n)$-ը մոտարկման սխալներն են, որը սովորաբար ենթադրվում են զրոյական միջինով և համաչափ պատահական:

Այս հանգեցնում է հետևյալ մոդելին`

$$(1) \qquad y = X\beta^* + \xi \qquad \xi \overset{iid}{\sim} P \ \text{ s.t. } \mathbb{E}[\xi_i] = 0, \ \mathrm{Var}[\xi_i] = \sigma^2$$

Հետազոտողի նպատակն է գնահատել β^*-ը` հենվելով (y, X) դիտարկումներից:

(1)-ը կոչվում է գծային ռեգրեսիայի մոդել:

3. ՆՎԱԶԱԳՈՒՅՆ ՔԱՌԱԿՈՒՍԻՆԵՐԻ ՄԵԹՈԴ

Սահմանում

Հետևյալ օպտիմիզացիայի խնդրի կամայական լուծում կոչվում է նվազագույն քառակուսիների գնահատական`

$$(2) \qquad \hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \|Y - X\beta\|_2^2$$

Նկատենք, որ Ն.Ք.Գ.-ն որոշվում է ոչ միարժեքորեն: Սակայն (2) խնդրի կամայական երկու լուծումների համար` $\hat{\beta}$ և $\hat{\beta}'$, ճիշտ է հետևյալ հավասարությունը`

$$X\hat{\beta} = X\hat{\beta}';$$

Հետևաբար, եթե X մատրիցի սյուները գծորեն անկախ են, ապա (2)-ի լուծումը միակն է: Ավելին, այն կարող է գրվել բացահայտ տեսքով:

Լեմմա եթե $X^T \cdot X$ մատրիցը հակադարձելի է, ապա (2)-ի լուծումը միակն է և այն գրվում է հետևյալ բանաձևով` $\hat{\beta} = (X^T \cdot X)^{-1} X^T \cdot Y$.

Ապացույց Նշանակենք $g(\beta) = \|Y - X\beta\|_2^2$.

$g : \mathbb{R}^p \to \mathbb{R}$ ուռուցիկ ֆունկցիա է: Դրա մինիմումի կետը պետք է բավարարի $\nabla g(\hat{\beta}) = 0$ պայմանին: Նկատենք, որ էվկլիդյան նորմը բավարարում է $\|v\|_2^2 = v^T \cdot v$ հավասարությանը: Հետևաբար

$$\nabla g(\beta) = \nabla\left((Y - X\beta)^T (Y - X\beta)\right)$$
$$= \nabla\left(Y^T Y - 2Y^T X\beta + \beta^T (X^T X)\beta\right)$$
$$= -2(Y^T X)^T + 2(X^T X)\beta$$

Որտեղից բխում է, որ

$$\nabla g(\hat{\beta}) = 0 \iff X^T Y = (X^T X)\hat{\beta}$$
$$\iff (X^T X)^{-1}(X^T Y) = \hat{\beta}$$

Օգտակար է նաև Ն.Փ.Գ.-ի երկրաչափական մեկնաբանությունը: Դիտարկենք

$$E = \mathrm{Span}\{X^1, \ldots, X^p\} \subset \mathbb{R}^n$$

գծային թաղանթը՝ ձգված X մատրիցի սյուները հանդիսացող վեկտորներով: Ընդհանրապես, $X\beta$-ն պատկանում է E-ին, կամայական $\beta \in \mathbb{R}^p$-ի համար: Հետևաբար, լուծելով (2) խնդիրը մենք գտնում ենք E ենթատարածության մեջ Y-ին ամենամոտ կետը: Եթե նշանակենք $\hat{v} = X\hat{\beta}$, ապա

$$\hat{v} \in \underset{v \in E}{\arg\min} \|Y - v\|_2^2$$

վերջին խնդրի լուծումը գտնելու համար բավական է հաշվել Y-ի պրոյեկցիան E ենթատարածության վրա՝

$$\hat{v} = \mathrm{Proj}_E(Y)$$

<u>Ուստի</u> Ն.Փ.Գ.-ն բավարարում է $X\hat{\beta} = \mathrm{Proj}_E(Y)$ առնչությանը: