Uporabna statistika

Gregor Dolinar

Fakulteta za elektrotehniko
Univerza v Ljubljani

18. december 2013

smo naredili $n > k$ opazovanj (zakaj mora biti $n > k$?). Podatke opazovanj označimo

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), \quad i = 1, 2, \ldots, n, \quad n > k.$$

Iščemo take koeficiente $\beta_1, \ldots, \beta_k$, da podatki opazovanj najbolje ustrezajo modelu

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i,$$

$i = 1, 2, \ldots, n$.

Zapišemo funkcijo

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 .$$

Iščemo minimum funkcije glede na spremenljivke $\beta_0, \beta_1, \ldots, \beta_k$. Funkcijo parcialno odvajamo po spremenljivkah $\beta_0, \beta_1, \ldots, \beta_k$ in dobljene funkcije izenačimo z 0. Dobimo $k + 1$ linearnih enačb za $k + 1$ neznank.

Matrični zapis

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

kjer je

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Naj bo $\hat{\beta}$ rešitev enačb. Potem velja

$$\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{y}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Obrnljivost! Moore-Penroseov inverz.

Regresijski model se potem zapiše v obliki

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

Residuale označimo

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

Definicija

Kovariančna matrika je matrika

$$\mathrm{cov}\hat{\beta} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Po diagonali so variance β_i, izven diagonale so kovariance

Ocena za varianco ϵ v modelu multiple linearne regresije s p parametri je

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}.$$

Dobro načrtovan poskus lahko:

- izboljša kakovost
- zmanjša variabilnost
- skrajša čas potreben za razvoj izdelka
- zmanjša stroške

Načrtovanje poskusa:

- predpostavke (vpliv parametrov, ...)
- poskus (slučajnost)
- analiza
- zaključek

Primer

Izboljšanje kakovosti papirnatih vrečk. Preverjamo vpliv koncentracije trdega lesa v papirnati masi.
Analiziramo vpliv pri koncentracijah: 5 %, 10 %, 15 % in 20 %.
Grafična predstavitev (škatlasti diagram)
Statistična analiza?

Zanima nas, kako različne vrednosti parametra vplivajo na rezultat poskusa.
Zanima nas, kako različne vrednosti neke neodvisne spremenljivke vplivajo na vrednost slučajne spremenljivke.

Statistični model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

razredi $i = 1, 2, \ldots, a$,
število meritev v vsakem razredu $j = 1, 2, \ldots, n$,
μ pričakovana vrednost,
τ_i vpliv posameznih razredov
(privzamemo

$$\sum_{i=1}^{a} \tau_i = 0),$$

ϵ_{ij} odstopanja ij-tega poskusa od i-tega razreda.

Oziroma

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

$\mu_i = \mu + \tau_i$, $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, n$, kjer je μ_i povprečje posameznega razreda.
Privzamemo ϵ_{ij} normalno porazdeljene slučajne spremenljivke, matematično upanje 0, varianca σ^2.

Pozor!

- Določeni razredi - naključno izbrani razredi.
- Popolnoma slučajen poskus.

Ker so τ_i definirani kot odstopanja od μ, je

$$\sum_{i=1}^{a} \tau_i = 0.$$

razred					vsota	povprečje
1	y_{11}	y_{12}	$\dots$	y_{1n}	$y_{1.}$	$\overline{y}_{1.}$
2	y_{21}	y_{22}	$\dots$	y_{2n}	$y_{2.}$	$\overline{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\dots$	y_{an}	$y_{a.}$	$\overline{y}_{a.}$
					$y_{..}$	$\overline{y}_{..}$

Najprej preverimo, ali sprememba vrednosti neodvisne spremenljivke vpliva na rezultat. Če ne vpliva, so vse vrednosti τ_i enake.
Postavimo domnevo.

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0$$

$$H_1 : \tau_i \neq 0 \text{ za vsaj eno vrednost } i$$

Označimo (SS summed squares):

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{..})^2$$

celotna vsota kvadratov odstopanj

$$SS_A = n\sum_{i=1}^{a}(\overline{y}_{i.} - \overline{y}_{..})^2$$

vsota kvadratov med razredi

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{i.})^2$$

nepojasnjena odstopanja

$$SS_T = SS_A + SS_E$$

$$E(SS_A) = (a-1)\sigma^2 + n\sum_{i=1}^{a}\tau_i^2$$

$$E(SS_E) = a(n-1)\sigma^2$$

$$F_0 = \frac{SS_A/(a-1)}{SS_E/(a(n-1))} = \frac{MS_A}{MS_E}$$

F porazdelitev z $a-1$ in $a(n-1)$ prostostnimi stopnjami.
Hipotezo H_0 zavrnemo, če je $f_0 > f_{\alpha,a-1,a(n-1)}$

Fisherjev test LSD (least significant difference)

Vemo, da eden izmed faktorjev vpliva. Kateri?
Za vsak par preverimo hipotezo

$$H_0 : \mu_i = \mu_j$$

s pomočjo t-testa

$$t_0 = \frac{\overline{y}_{i.} - \overline{y}_{j.}}{\sqrt{\frac{2MS_E}{n}}}.$$

Par bomo definirali kot bistveno različen, če je

$$|\overline{y}_{i.} - \overline{y}_{j.}| > \text{LSD},$$

kjer je

$$\text{LSD} = t_{\alpha/2,a(n-1)}\sqrt{\frac{2MS_E}{n}}.$$

Preverjanje modela (normalnost ϵ_{ij})

Bločna ANOVA

Statistični model:

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij},$$

razredi $i = 1, 2, \ldots, a$,
bloki $j = 1, 2, \ldots, b$,

razred					vsota	povprečje
1	y_{11}	y_{12}	$\cdots$	y_{1b}	$y_{1.}$	$\overline{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2b}	$y_{2.}$	$\overline{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{ab}	$y_{a.}$	$\overline{y}_{a.}$
					$y_{..}$	$\overline{y}_{..}$

Imamo b blokov.

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{..})^2$$

$$SS_A = b\sum_{i=1}^{a}(\overline{y}_{i.} - \overline{y}_{..})^2$$

$$SS_B = a\sum_{j=1}^{b}(\overline{y}_{.j} - \overline{y}_{..})^2$$

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{.j} - \overline{y}_{i.} + \overline{y}_{..})^2$$

$$SS_T = SS_A + SS_B + SS_E$$