Linearna regresia

V texte su vynechane vektorove znacenia. Veliciny θ, y, ϵ su vektorove veliciny.

Linearna regresia je odhad vektoru parametrov θ pomocou linearneho modelu $m = \mathbf{A}\theta$, o ktorom dufame, ze pre odhadnute $\hat{\theta}$ bude $m \approx y$, kde y su data. $\mathbf{A}$ je modelova matica a pre tento problem ma tvar:

$$\mathbf{A} = \begin{pmatrix} 1 & x_0^2 \\ \vdots & \vdots \\ 1 & x_N^2 \end{pmatrix} \tag{1}$$

V linearnom pripade su odhady parametrov vypocitane ako:

$$\hat{\theta} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{W}\mathbf{A}\right)^{-1} \mathbf{A}^{\mathrm{T}}\mathbf{W}y, \tag{2}$$

kde $\mathbf{W}$ je vahova matica. V tomto probleme sa neuplatini, bude identita, kedze body maju rovnaku vahu. Vzorec je mozne odvodit z minimalizacie $g = \sum_{i=1}^{N}(y_i - m_i)w_{ij}(y_j - m_j)$.

Chyby a korelacie parametrov su dane kovariancnou maticou:

$$c_{ij} \equiv \mathrm{Cov}(x_i, x_j) = E((x_i - E(x_i))(x_j - E(x_j))) = E(x_i x_j) - E(x_i)E(x_j) \tag{3}$$

Oznacim:

$$\xi_{ij} \equiv \left(\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1} \mathbf{A}^{\mathrm{T}}\right)_{ij} \tag{4}$$

potom dosadenim odhadov $\hat{\theta}$ (uz s $W = I$):

$$\mathrm{Cov}(\hat{\theta}_i, \hat{\theta}_j) = E\left(\xi_{ki} y_i \xi_{lj} y_j\right) - E\left(\xi_{ki} y_i\right) E\left(\xi_{lj} y_j\right) = \tag{5}$$

$$\xi_{ki}\mathrm{Cov}(y_i, y_j)\xi_{jl} \tag{6}$$

Predpokladame, ze $\mathbf{C}(y, y)$ je diagonalna, ale s neznamym rozpylom (rovnaky pre vsetky body). Nevychyleny odhad rozptylu je:

$$\sigma_y^2 \approx s_y^2 = \frac{\epsilon_i \epsilon_i}{N - p}, \tag{7}$$

kde N je pocet bodov y a p je pocet fitovanych parametrov.

Intervalovy odhad

Odhadovane parametre su tiez nahodne premenne. Rozdelenie ϵ je nahodne s nulovym prvym momentom a rozpylom σ_y^2. Z toho vyplyva, ze rozdelenie $\hat{\theta} - \theta$ je tiez normalne, kedze je dane linearnym vztahom z ϵ. Rozdelenie $\frac{\epsilon_i \epsilon_i}{\sigma_y^2} \equiv R$ je χ^2 z definicie χ^2. Zaujima nas, z akeho rozdelenia bude θ. Odhadovany rozptyl parametrov $s_{\hat{\theta}}^2$ je dany diagonalou $\mathbf{C}(\hat{\theta}, \hat{\theta})$ a suvisi s odhadnutym rozpytlom s_y^2:

$$s_{\hat{\theta}i}^2 = \mathrm{Cov}(\hat{\theta}_i, \hat{\theta}_i) = \xi_{ik} \mathrm{Cov}(y_k, y_l) \xi_{li} = s_y^2 \xi_{ik} \xi_{ki} \tag{8}$$

Potom vyraz:

$$t_i = \frac{\theta_i - \hat{\theta}_i}{s_{\theta i}} = \frac{\theta_i - \hat{\theta}_i}{\xi_{ik}\xi_{ki} s_y} = \frac{\theta_i - \hat{\theta}_i}{\xi_{ik}\xi_{ki}} \sqrt{\frac{(N-p)\sigma_y^2}{R}} \tag{9}$$

je zo Studentovho t-rozdelenia a je mozne ho pouzit pre konstrukciu intervalu spolahlivosti. Konkretne:

$$\Pr(-\tau < t < \tau) = 1 - \alpha, \tag{10}$$

kde τ je hodnota (kvantil), pre ktoru $T(\tau) = 1 - \frac{\alpha}{2}$, kde T je (kumulativna) Studentova distribucna funkcia. Vyraz je mozne pretvorit na tvrdenie o θ:

$$\Pr(\hat{\theta} - s_\theta \tau < \theta < \hat{\theta} + s_\theta \tau) = 1 - \alpha, \tag{11}$$

co by malo byt ekvivalentne tvdeniu, ze θ lezia s pravdepodobnostou $1 - \alpha$ v intervale:

$$\theta \in [\hat{\theta} - s_\theta \tau, \hat{\theta} + s_\theta \tau] \tag{12}$$

Pas spolahlivosti

Pre pas spolahlivosti okolo celej krivky existuje podobny postup ako vyssie (podrobne v (Casella 2002) a (Michael H Kutner 2005), alebo tiez).

$$\mathrm{Var}(a_{ij}\theta_j) = a_{ij} a_{ik} \mathrm{Cov}(\theta_j, \theta_k) \tag{13}$$

alebo v maticovom zapise:

$$s_m^2 = \mathrm{diag}\left(\mathbf{A}\mathbf{C}(\theta, \theta)\mathbf{A}^{\mathrm{T}}\right) \tag{14}$$

Znova potrebujeme najst rozdelenie vyrazu:

$$f_i = \frac{A\theta - A\hat{\theta}}{s_m} \tag{15}$$

pre vycislenie tvrdenia:

$$\Pr(\mathbf{A}\hat{\theta} - s_m\phi < \mathbf{A}\theta < \mathbf{A}\hat{\theta} + s_m\phi \text{ for all } x) = 1 - \alpha \tag{16}$$

Ukazuje sa, ze ϕ^2 je z F-rozdelenia:

$$\phi = \sqrt{2F_{\alpha;p,N-p}} \tag{17}$$

Intervalovy odhad chyby

Ako bolo pouzite vyssie rozptyl data je odhadnuty z:

$$\sigma_y^2 \approx s_y^2 = \frac{\epsilon_i\epsilon_i}{N-p} \tag{18}$$

a vyraz:

$$c = \frac{\epsilon_i\epsilon_i}{\sigma_y^2} = \frac{s_y^2(N-p)}{\sigma_y^2} \tag{19}$$

je z χ^2-rozdelenia ($N-p$ stupnov volnosti). To znamena, ze:

$$\Pr\left(\sigma_l \le \sigma_y \le \sigma_u\right) = 1 - \alpha \tag{20}$$

$$\Pr\left(\sigma_l^2 \le \sigma_y^2 \le \sigma_u^2\right) = 1 - \alpha \tag{21}$$

$$\Pr\left(\frac{(N-p)s_y^2}{\sigma_l^2} \ge \frac{(N-p)s_y^2}{\sigma_y^2} \ge \frac{(N-p)s_y^2}{\sigma_u^2}\right) = 1 - \alpha \tag{22}$$

Krajne hodnoty teda budu:

$$\sigma_u = \sqrt{\frac{(N-p)s_y^2}{c_{\alpha/2}}} \tag{23}$$

$$\sigma_l = \sqrt{\frac{(N-p)s_y^2}{c_{1-\alpha/2}}}, \tag{24}$$

kde c_β je definovana ako:

$$\beta = \int_0^{c_\beta} \chi^2(x)\mathrm{d}x \tag{25}$$

Casella, George. 2002. *Statistical Inference*. Australia Pacific Grove, CA: Thomson Learning.

Michael H Kutner, John Neter, Christopher J. Nachtsheim. 2005. *Applied Linear Statistical Models*. 5th ed. The Mcgraw-Hill/Irwin Series Operations and Decision Sciences. McGraw-Hill Irwin.