5 Lineární regrese

Průvodce studiem

Tato kapitola je pro pochopení možností a technik analýzy vícerozměrných dat klíčová. Proto na tuto kapitolu počítejte nejméně se třemi hodinami usilovného studia s tím, že se k probírané látce budete ještě vracet po pochopení dalších souvislostí.

Regrese je jednou z velice často aplikovaných statistických metod, uvádí se, že dokonce naprostá většina aplikací $(70 - 90\%)$ je nějakou formou regresních metod. Počátky metody nejmenších čtverců jsou dokumentovány již na začátku 19. století (Legendre, Gauss), minimalizace součtu absolutních hodnot odchylek je připisována Galileovi ještě o pár desítek let dříve.

5.1 Klasický lineární model, metoda nejmenších čtverců

Klasický model lineární regrese lze zapsat jako

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{3}$$

kde

y_i je pozorovaná hodnota náhodné veličiny Y

$x_{i1}, \cdots, x_{ik}$ jsou hodnoty vysvětlujících proměnných (regresorů, prediktorů)

$\beta_0, \beta_1, \cdots, \beta_k$ jsou parametry modelu (fixní, leč neznámé hodnoty)

ε_i je náhodná složka

$i = 1, 2, \cdots, n$ je index pozorování (objektu)

Rovnici (3) můžeme zapsat maticově

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

[pro i-tý řádek pak

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{5}$$

Vektory jsou označeny příslušnými malými tučnými písmeny, matice velkými tučnými písmeny. Matice $\mathbf{X}$ má $k+1$ sloupců, v prvním sloupci jsou jedničky, dalších sloupcích jsou hodnoty vysvětlujících veličin.

Obvyklými předpoklady v klasickém lineární modelu jsou:

1. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, tj. vektor středních hodnot náhodné složky je roven nulovému vektoru

2. $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n, \quad \sigma^2 > 0, \quad \mathbf{I}$ je jednotková matice řádu n, tj. náhodné složky jsou nekorelované a jejich rozptyl je konstantní
3. $\mathbf{X}$ je nenáhodná matice typu $n \times (k+1)$
4. hodnost matice $\mathbf{X}$, $h(\mathbf{X}) = k+1 \leq n$, tj. sloupce matice $\mathbf{X}$ nejsou lineárně závislé a počet pozorování je alespoň roven počtu parametrů

Z rovnice (**??**) dostaneme pro vektor podmíněných středních hodnot

$$E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \tag{6}$$

a pro kovarianční matici s využitím rov. (1)

$$\begin{gathered}\mathrm{cov}(\mathbf{y} \mid \mathbf{X}) = E\left\{[\mathbf{y} - E(\mathbf{y} \mid \mathbf{X})][\mathbf{y} - E(\mathbf{y} \mid \mathbf{X})]^T \mid \mathbf{X}\right\} = \\ = E\left\{[\mathbf{y} - \mathbf{X}\beta][\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]^T \mid \mathbf{X}\right\} = E\left\{(\boldsymbol{\varepsilon} \mid \mathbf{X})(\boldsymbol{\varepsilon} \mid \mathbf{X})^T \mid \mathbf{X}\right\} = \sigma^2\mathbf{I}\end{gathered} \tag{7}$$

Neznámé parametry $\boldsymbol{\beta}$ lze odhadnout metodou nejmenších čtverců, tj. nalezením takového vektoru $\mathbf{b}$, pro který je nejmenší tzv. residuální suma čtverců (*RSS*) odchylek pozorovaných hodnot od jejich odhadů z modelu

$$\mathrm{RSS} = (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) \tag{8}$$

Položíme-li derivaci výrazu (8) podle vektoru $\mathbf{b}$, tj.

$$\begin{gathered}\frac{\partial\,\mathrm{RSS}}{\partial\mathbf{b}} = \frac{\partial}{\partial\mathbf{b}}(\mathbf{y}^T\,\mathbf{y} - \mathbf{b}^T\,\mathbf{X}^T\,\mathbf{y} - \mathbf{y}^T\,\mathbf{X}\,\mathbf{b} + \mathbf{b}^T\,\mathbf{X}^T\,\mathbf{X}\,\mathbf{b}) = \\ = \frac{\partial}{\partial\mathbf{b}}(-2\,\mathbf{b}^T\,\mathbf{X}^T\,\mathbf{y} + \mathbf{b}^T\,\mathbf{X}^T\,\mathbf{X}\,\mathbf{b}) = -2\,\mathbf{X}^T\,\mathbf{y} + 2\mathbf{X}^T\,\mathbf{X}\,\mathbf{b}\end{gathered}$$

rovnu nulovému vektoru, dostaneme soustavu normálních rovnic

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\mathbf{b} \tag{9}$$

a vzhledem k platnosti předpokladu (4) můžeme řešení této soustavy lineárních rovnic (vzhledem k $\mathbf{b}$) vyjádřit explicitně jako

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{10}$$

Odhady určené podle (10) se nazývají OLS – odhady (Ordinary Least Squares). Tyto odhady jsou nestranné, neboť

$$E(\mathbf{b}) = E(\mathbf{b}|\mathbf{X}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{y} \mid \mathbf{X}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \tag{11}$$

Lze ukázat, že tyto odhady mají další dobré vlastnosti, jsou to BLU-odhady (Best Linear Unbiased). Kovarianční matice těchto odhadů je

$$\mathrm{cov}(\mathbf{b}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right]^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{12}$$

Na diagonále této matice jsou rozptyly odhadu parametrů $\text{var}(b_i)$.

Nestranný odhad parametru σ^2 je (důkaz viz např. Anděl 1978)

$$s^2 = \frac{\text{RSS}}{n-k-1} \tag{13}$$

a nestranný odhad kovarianční matice odhadů parametru $\mathbf{b}$ je

$$\mathbf{S}_{bb} = s^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{14}$$

Na diagonále matice $\mathbf{S}_{bb}$ jsou tedy nestranné odhady rozptylů odhadu parametrů, jejich odmocninu (směrodatnou odchylku odhadu) označme $s(b_i)$.

Přidáme-li k předpokladům (1) až (4) ještě předpoklad o tvaru rozdělení náhodné složky modelu (3), resp. (**??**), a to

$$(5) \qquad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \cdots, n,$$

S využitím předpokladu (2) $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ pro následující statistiku platí

$$\frac{b_i - \beta_i}{\sqrt{\text{var}(b_i)}} \sim N(0,1) \qquad i = 0, 1, \cdots, k$$

a pro

$$\frac{b_i - \beta_i}{s(b_i)} \sim t_{n-k-1}. \tag{15}$$

Tuto statistiku (15) pak můžeme užít ke stanovení intervalu spolehlivosti pro parametr β_i a testování hypotéz o tomto parametru.

5.2 Odhad parametrů metodou maximální věrohodnosti

Za předpokladu (5) můžeme odvodit i maximálně věrohodné (Maximum Likelihood) ML odhady pro klasický model lineární regrese. ML-odhady odhadují hodnoty parametrů tak, aby tyto odhady maximalizovaly tzv. věrohodnostní funkci, tj. odhady jsou určeny jako nejpravděpodobnější hodnoty parametrů pro pozorovaná data. ML-odhady obecně mají řadu dobrých vlastností:

- jsou asymptoticky nestranné (s rostoucím n jejich střední hodnota konverguje k odhadovanému parametru)
- skoro vždy jsou konsistentní (s rostoucím n rozptyl odhadu konverguje k nule)

Věrohodnostní funkce (součin hustot jednotlivých pozorování) má pro klasický model lineární regrese tvar

$$L_{ML} = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}}{2\sigma^2}\right)$$

a její logaritmus je

$$\ln(L_{ML}) = L = (-\frac{n}{2}) \ln\left(2\pi\sigma^2\right) - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \frac{\mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \tag{16}$$

Maximálně věrohodnými odhady jsou takové odhady $\boldsymbol{\beta}_{ML}$, pro které věrohodnostní funkce (16) nabývá maxima. Při hledání maxima funkce (16) položíme derivace podle hledaných proměnných rovny nule, tedy

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{0}, \qquad \frac{\partial L}{\partial \sigma^2} = 0 \tag{17}$$

a dostaneme

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_{ML} \tag{18}$$

Vidíme, že ML-odhady regresních koeficientů jsou v klasickém lineárním modelu stejné jako odhady získané metodou nejmenších čtverců (OLS-odhady), $\boldsymbol{\beta}_{ML} = \mathbf{b}$, srovnej rov.(18) s rov.(10).

ML-odhad parametru σ^2 z rov.(17) je

$$\sigma_{ML} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{\mathrm{RSS}}{n}$$

tedy se liší od OLS-odhadu v rov.(13), je pouze asymptoticky nestranný.

Shrnutí

- *lineární regresní model*
- *předpoklady v klasickém lineárním regresním modelu*
- *metoda nejmenších čtverců*
- *metoda maximální věrohodnosti*
- *kovarianční matice odhadů parametrů*

Kontrolní otázky

1. *Jaký je rozdíl mezi parametry a jejich odhady?*
2. *Jsou odhady parametrů náhodné veličiny?*
3. *Jaký je tvar kovarianční matice odhadů parametrů v klasickém modelu?*
4. *Jsou odhady získané metodou nejmešších čtverců nestranné? Dokažte to.*

Korespondeční úloha

Korespondenční úlohy budou zadávány ke každému kursu samostatně.