Lineární modely včetně zobecněných, logistická regrese, analýza přežívání v R

—

17VSADR – Skriptování a analýza dat v jazyce R

Lubomír Štěpánek[1, 2]

[1]Oddělení biomedicínské statistiky
Ústav biofyziky a informatiky
1. lékařská fakulta
Univerzita Karlova v Praze

[2]Katedra biomedicínské informatiky
Fakulta biomedicínského inženýrství
České vysoké učení technické v Praze

Obsah

Zavedení lineární regrese

- buď $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ vektor spojité závisle proměnné, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ je vektor lineárních koeficientů, nakonec $\boldsymbol{X}$ je designová matice (též datová matice či matice modelu) právě k nezávisle proměnných (spojitých či kategorických) tak, že

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix}$$

- pak lineární regresí (vícerozměrnou pro $k > 1$) nazveme model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ je vektor chybové složky (též méně přesně vektor reziduí)

Hledání řešení modelu lineární regrese

- iniciálně známe vektor $\boldsymbol{y}$ a matici $\boldsymbol{X}$
- hledáme takový vektor odhadů lineárních koeficientů $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^T$, aby součet druhých mocnin členů vektoru chybové složky $\boldsymbol{\varepsilon}$ byl co nejmenší
- tedy nalézt model lineární regrese mezi danými k nezávisle proměnnými a jednou závisle proměnnou znamená nalézt odhady lineárních koeficientů $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ tak, aby

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \varepsilon_i^2 \right\} = \arg\min_{\boldsymbol{\beta}} \left\{ \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} \right\} = \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\} = \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \ldots - \beta_k x_k)^2 \right\}\end{aligned}$$

Tvar řešení lineární regrese

- řešení vede na soustavu normalizovaných rovnic, jejichž analytickým (!) řešením je vektor odhadů lineárních koeficientů

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

- za předpokladu, že vektor chybové složky $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ splňuje tzv. slabou sadu předpokladů, tj.
  - (i) nulová střední hodnota každého rezidua, tj. $\mathbf{E}(\varepsilon_i) = 0$ pro $\forall i \in \{1, 2, \ldots, n\}$
  - (ii) (homoskedasticita) konečný konstantní rozptyl každého rezidua, tj. $\mathrm{var}(\varepsilon_i) = \sigma^2 < \infty$ pro $\forall i \in \{1, 2, \ldots, n\}$
  - (iii) (nekorelovanost) lineární nezávislost dvou různých reziduí, tj. $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ pro $\forall i, j \in \{1, 2, \ldots, n\}$ a $i \neq j$

  lze ukázat, že řešení $\hat{\boldsymbol{\beta}}$ má kýžené vlastnosti, např.

$$\mathbf{E}(\hat{\beta}_j - \beta_j) = 0$$

  pro $\forall j \in \{0, 1, 2, \ldots, k\}$ apod.

Tvar řešení lineární regrese

- současně lze nahlédnout, že odhady tzv. vyrovnaných hodnot $\boldsymbol{y}$ vyjádříme za slabé sady předpokladů jako

$$\hat{\boldsymbol{y}} = \mathbf{E}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{E}(\boldsymbol{X}\boldsymbol{\beta}) + \mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{E}(\boldsymbol{X}\boldsymbol{\beta}) + \mathbf{0} = \boldsymbol{X}\mathbf{E}(\boldsymbol{\beta}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

  a po rozepsání

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

  kde tvar $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ nazýváme též projekční či hat maticí a značíme $\boldsymbol{H} \equiv \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
- znalost projekční matice $\boldsymbol{H}$ je výhodná pro zkoumání vlivu závisle proměnné mezi různými pozorováními
- vidíme tedy, že jak lineární koeficienty $\hat{\boldsymbol{\beta}}$, tak vyrovnané hodnoty $\hat{\boldsymbol{y}}$ odhadneme pouze s pomocí apriorně známých hodnot, tj. datové matice $\boldsymbol{X}$ a vektoru závisle proměnné $\boldsymbol{y}$

Některé důsledky pro výpočetní statistiku a analýzu dat

- protože $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ a $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, všechny výpočty jsou maticové, tj. nevyžadují striktně numerické přístupy a obvykle existuje jednoznačné číselné řešení
- maticové výpočty mohou být náročné na paměť, méně na výpočetní čas – to je rozdíl oproti zobecněným lineárním či nelineárním modelům (time-memory trade-off)
- caveat může nastat ve fázi výpočtu $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, tedy inverze k $(\boldsymbol{X}^T\boldsymbol{X})$; má-li $(\boldsymbol{X}^T\boldsymbol{X})$ nízkou hodnost v důsledku např. multikolinearity či *numerical fuzz*, je řešení číselně „nestabilní“ nebo vůbec neexistuje

Literatura

- Karel Zvára. *Základy statistiky v prostředí R*. Praha, Česká republika: Karolinum, 2013. ISBN: 978-80-246-2245-3.
- Hadley Wickham. *Advanced R*. Boca Raton, FL: CRC Press, 2015. ISBN: 978-1466586963.

Děkuji za pozornost!

lubomir.stepanek@lf1.cuni.cz
lubomir.stepanek@fbmi.cvut.cz

GitHub

github.com/LStepanek/17VSADR_Skriptovani_a_analyza_dat_v_jazyce_R