Iterativne numerične metode v posplošenih linearnih modelih

Mitja Mandić
Mentor: izr. prof. dr. Jaka Smrekar

1. april 2021

Posplošeni linearni modeli

- Slučajni del, sistematični del, povezovalna funkcija

Posplošeni linearni modeli

- Slučajni del, sistematični del, povezovalna funkcija
- Linearna regresija:
$$\mathbb{E}(Y) = x^T \beta$$
- Problem - ni najboljša. Rešitev? Transformacija pričakovane vrednosti

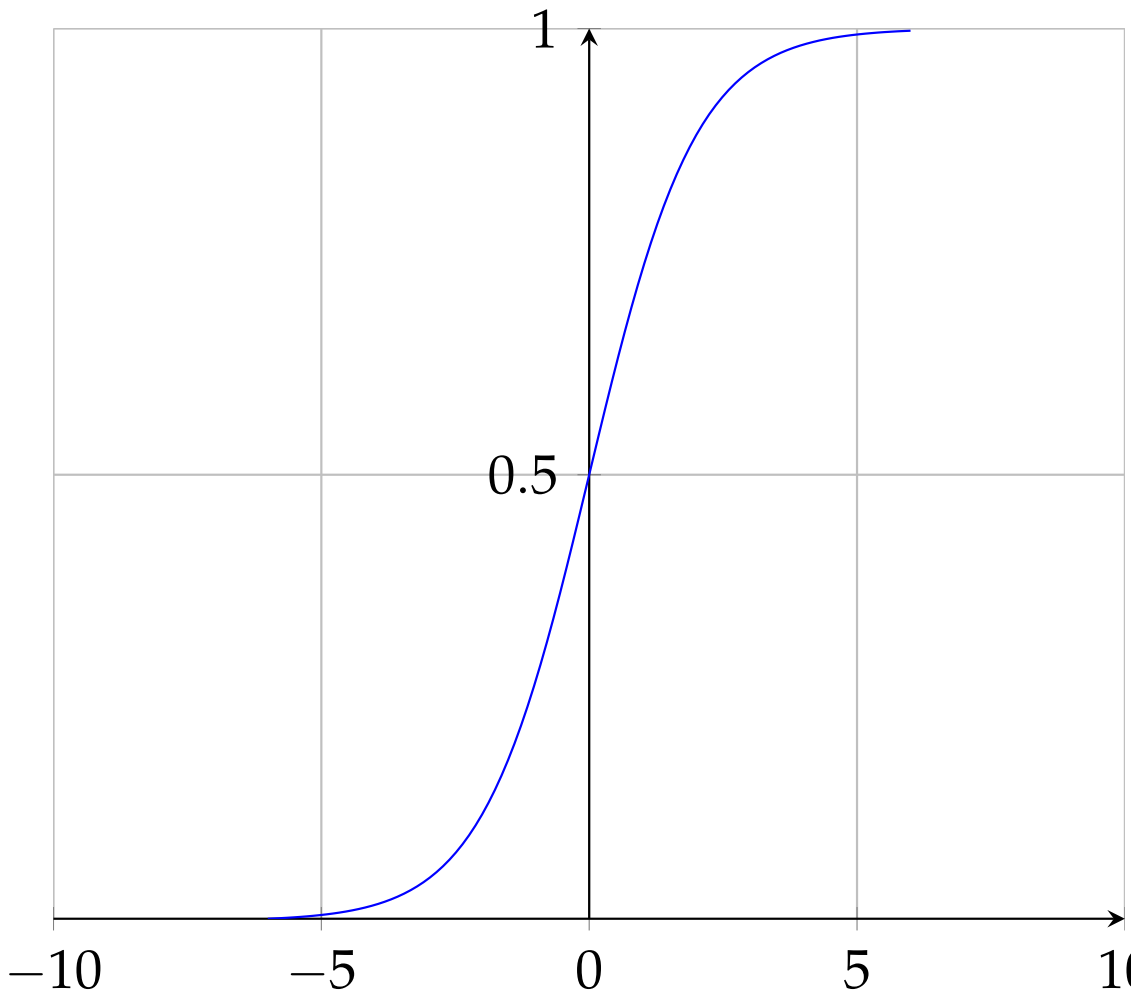
Logistični model

- Za kategorične podatke $\rightarrow$ binomska porazdelitev

Logistični model

- Za kategorične podatke $\rightarrow$ binomska porazdelitev
- $\text{logit}(p_i) = \log(\frac{p_i}{1-p_i}) = x^T\beta$

$$p = \frac{e^{x^T \beta}}{1 + e^{x^T \beta}}$$



Točkovno ocenjevanje

- *Cenilka* je funkcija vzorca, s katero ocenjujemo določeno karakteristiko

Točkovno ocenjevanje

- *Cenilka* je funkcija vzorca, s katero ocenjujemo določeno karakteristiko
- Dve glavni metodi za določanje: metoda momentov in metoda največjega verjetja

Metoda momentov

- Enostavnejša za računanje brez računalnika

Metoda momentov

- Enostavnejša za računanje brez računalnika
- Karakteristiko izrazimo kot funkcijo momentov
  $c(X) = g(m_1(X), m_2(X), \ldots, m_r(X))$ in jo ocenimo z $g(\hat{m}_1, \ldots, \hat{m}_r)$

Metoda največjega verjetja

- Najprej privzemimo gostote oblike $f_X(x;\theta) = f(x;\theta_1,\ldots,\theta_r)$ za nek $\theta \in \Theta$
- Pri fiksni realizaciji poskusa definiramo *funkcijo verjetja*

$$\ell(X_1,\ldots,X_n;\underbrace{\theta_1,\ldots,\theta_r}_{\theta}) = f(X_1,\theta)\cdots f(X_n,\theta)$$

- Iščemo θ, kjer bo imela maksimum, kar bo natanko ničla odvoda $\log \ell$
- Sistemu

$$\frac{\partial}{\partial\theta_j}\log \ell(X,\theta) = 0$$

pravimo sistem enačb verjetja, njegova rešitev je *cenilka največjega verjetja*

- Niso nujno nepristranske, so pa dosledne, če je rešitev enolična
- Običajno niso eksplicitno rešljive

Eksponentna družina

$$f_Y(y; \theta, \phi) = \exp\left(\frac{(y\theta - b(\theta))}{a(\phi)} + c(y, \phi)\right)$$

za neke $a(\cdot), b(\cdot)$ in $c(\cdot)$. θ imenujemo tudi naravni parameter.

Eksponentna družina

Iz predavanj STAT1 se spomnimo, da velja

$$\mathbb{E}(\nabla \ell) = 0.$$

Na podoben način z uporabo $\int f_Y(y;\theta)\, dy = 1$ pa dokažemo tudi *informacijsko enakost*

$$\mathbb{E}(\frac{\partial^2}{\partial \theta^2}\ell(\theta)) = -\mathbb{E}(\frac{\partial}{\partial \theta}\ell(\theta))^{\not\in}$$

Z uporabo prve enakosti sledi

$$\frac{\partial}{\partial\theta}\ell = \frac{y - b'(\theta)}{a(\phi)} \rightarrow \mu = b'(\theta)a(\phi)$$

Iz druge pa:

$$\mathbb{E}(\frac{\partial^2}{\partial\theta^2}\ell) = \mathbb{E}(\frac{b''(\theta)}{a(\phi)}) = \frac{b''(\theta)}{a(\phi)}$$
$$\mathbb{E}((\frac{\partial}{\partial\theta}\ell)^2) = \frac{1}{a(\phi)^2}\mathbb{E}((y-\mu)^2) = \frac{Var(Y)}{a(\phi)^2}$$

od koder direktno sledi

$$Var(Y) = -b''(\theta)a(\phi)$$

Zgled z binomsko porazdelitvijo

Naj bo $Y \sim \text{Bin}(n, p)$ Verjetnost

$$\text{P}(Y = y) = \binom{n}{y} p^y (1-p)^{n-y} = \exp\left(y \log(\frac{p}{1-p}) + n \log(1-p) + \log \binom{n}{k} \right),$$

od koder direktno sledi

$$\theta = \log \frac{p}{1-p} = \log \frac{\mu}{1-\mu},\ b(\theta) = \log(1+e^\theta),\ \phi = 1, a(\phi) = \frac{\phi}{n},$$

V tem primeru velja torej $\theta = \text{logit}(\mu) = X^\top \beta = \eta$ in *logit* je kanonična povezovalna funkcija za logistični model. Zakaj je to koristno?

Obvoz v numerične metode

- Za ocenjevanje parametrov β običajno rešujemo sistem enačb največjega verjetja
- v splošnem ni eksplicitno rešljiv in zato potrebujemo numerične metode

Obvoz v numerične metode

- Za ocenjevanje parametrov β običajno rešujemo sistem enačb največjega verjetja
- v splošnem ni eksplicitno rešljiv in zato potrebujemo numerične metode
- Newtonova metoda še vedno zelo aktualna zaradi kvadratična konvergence:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

Je pa to metoda za iskanje ničel in ne ekstremov!

Prilagoditev Newtonove metode

Najprej zapišimo Taylorjev polinom okoli iskane vrednosti θ

$$L(\theta) \approx L(\theta_n) + dL(\theta_n)(\theta - \theta_n) + \frac{1}{2}(\theta - \theta_n)^\top d^2 L(\theta_n)(\theta - \theta_n)$$

Prilagoditev Newtonove metode

Najprej zapišimo Taylorjev polinom okoli iskane vrednosti θ

$$L(\theta) \approx L(\theta_n) + dL(\theta_n)(\theta - \theta_n) + \frac{1}{2}(\theta - \theta_n)^\top d^2 L(\theta_n)(\theta - \theta_n)$$

Iščemo ekstrem funkcije torej ničlo odvoda in zato

$$\ell(\theta*) = 0 \approx \ell(\theta_0)(\theta * -\theta_0)$$

Izrazimo in dobimo iteracijski korak

$$\theta_t = \theta_{t-1} - \frac{\ell(\theta_{t-1})}{\ell'(\theta_{t-1})}$$

Prilagoditev Newtonove metode

Najprej zapišimo Taylorjev polinom okoli iskane vrednosti θ

$$L(\theta) \approx L(\theta_n) + dL(\theta_n)(\theta - \theta_n) + \frac{1}{2}(\theta - \theta_n)^\top d^2 L(\theta_n)(\theta - \theta_n)$$

Iščemo ekstrem funkcije torej ničlo odvoda in zato

$$\ell(\theta*) = 0 \approx \ell(\theta_0)(\theta * -\theta_0)$$

Izrazimo in dobimo iteracijski korak

$$\theta_t = \theta_{t-1} - \frac{\ell(\theta_{t-1})}{\ell'(\theta_{t-1})}$$

Kaj gre lahko narobe?

Prilagoditev Newtonove metode

- Invertiranju Hessiana se lahko izognemo z uporabo premikov:

$$\theta_t = \theta_{t-1} + h_{t-1}\ell'(\theta_{t-1})h_{t-1} = -\ell(\theta_{t-1})$$

Prilagoditev Newtonove metode

- Invertiranju Hessiana se lahko izognemo z uporabo premikov:

$$\theta_t = \theta_{t-1} + h_{t-1}\ell'(\theta_{t-1})h_{t-1} = -\ell(\theta_{t-1})$$

- Newtonova metoda ni naraščajoč algoritem $\Rightarrow$ ne vemo ali se bo premikal navzgor ali navzdol

Prilagoditev Newtonove metode

- Invertiranju Hessiana se lahko izognemo z uporabo premikov:

$$\theta_t = \theta_{t-1} + h_{t-1}\ell'(\theta_{t-1})h_{t-1} = -\ell(\theta_{t-1})$$

- Newtonova metoda ni naraščajoč algoritem $\Rightarrow$ ne vemo ali se bo premikal navzgor ali navzdol
- Če je Hessejeva matrika pozitivno definitna je algoritem konstanten!

Fisher scoring

$$\beta_{i+1} = \beta_i + \frac{\dot{l}(\beta_i)}{E(\ddot{l}(\beta_i))}$$

Fisher scoring

$$\beta_{i+1} = \beta_i + \frac{\dot{l}(\beta_i)}{E(\ddot{l}(\beta_i))}$$

Nazaj k eksponentni družini:

$$\log f_y(y;\theta) = L(y;\theta) = \frac{y\theta - b(\theta)}{a(\phi)}$$

$$\frac{\partial L}{\partial \beta_j} = (\frac{\partial L}{\partial \theta})(\frac{\partial \theta}{\partial \mu})(\frac{\partial \mu}{\partial \eta})(\frac{\partial \eta}{\partial \beta_j})$$

- $\frac{\partial L}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}$

$$\frac{\partial L}{\partial \beta_j} = (\frac{\partial L}{\partial \theta})(\frac{\partial \theta}{\partial \mu})(\frac{\partial \mu}{\partial \eta})(\frac{\partial \eta}{\partial \beta_j})$$

- $\frac{\partial L}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}$
- Z uporabo $(b')^{-1}(\mu) = \theta$ dobimo $\frac{\partial \theta}{\partial \mu} = \frac{1}{b''(\theta)} = \frac{a(\phi)}{var(Y)}$

$$\frac{\partial L}{\partial \beta_j} = (\frac{\partial L}{\partial \theta})(\frac{\partial \theta}{\partial \mu})(\frac{\partial \mu}{\partial \eta})(\frac{\partial \eta}{\partial \beta_j})$$

- $\frac{\partial L}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}$
- Z uporabo $(b')^{-1}(\mu) = \theta$ dobimo $\frac{\partial \theta}{\partial \mu} = \frac{1}{b''(\theta)} = \frac{a(\phi)}{var(Y)}$
- $(\frac{\partial \mu}{\partial \eta})$ bo odvisen od povezovalne funkcije, s tem se bomo ukvarjali pozneje

$$\frac{\partial L}{\partial \beta_j} = (\frac{\partial L}{\partial \theta})(\frac{\partial \theta}{\partial \mu})(\frac{\partial \mu}{\partial \eta})(\frac{\partial \eta}{\partial \beta_j})$$

- $\frac{\partial L}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}$
- Z uporabo $(b')^{-1}(\mu) = \theta$ dobimo $\frac{\partial \theta}{\partial \mu} = \frac{1}{b''(\theta)} = \frac{a(\phi)}{var(Y)}$
- $(\frac{\partial \mu}{\partial \eta})$ bo odvisen od povezovalne funkcije, s tem se bomo ukvarjali pozneje
- $(\frac{\partial \eta}{\partial \beta_j}) = x_{ij}$

$$\frac{\partial L}{\partial \beta_j} = (\frac{\partial L}{\partial \theta})(\frac{\partial \theta}{\partial \mu})(\frac{\partial \mu}{\partial \eta})(\frac{\partial \eta}{\partial \beta_j})$$

- $\frac{\partial L}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}$
- Z uporabo $(b')^{-1}(\mu) = \theta$ dobimo $\frac{\partial \theta}{\partial \mu} = \frac{1}{b''(\theta)} = \frac{a(\phi)}{var(Y)}$
- $(\frac{\partial \mu}{\partial \eta})$ bo odvisen od povezovalne funkcije, s tem se bomo ukvarjali pozneje
- $(\frac{\partial \eta}{\partial \beta_j}) = x_{ij}$

Končno,

$$\frac{\partial L}{\partial \beta_j} = \frac{y - \mu}{var(Y)} \frac{\partial \mu}{\partial \eta} x_{ij}$$

Ujemanje F-S in N-R

Če pa uprabimo kanonično povezovalno funkcijo je $\theta = \eta$ in zato $\frac{\partial \mu}{\partial \theta} = b''(\theta)$ in funkcija zbira postane

$$\frac{\partial L}{\partial \beta_j} = \frac{y - \mu}{var(Y)} b''(\theta) x_{ij} = \frac{y - \mu}{a(\phi)} x_{ij}$$

Ujemanje F-S in N-R

Če pa uprabimo kanonično povezovalno funkcijo je $\theta = \eta$ in zato $\frac{\partial \mu}{\partial \theta} = b''(\theta)$ in funkcija zbira postane

$$\frac{\partial L}{\partial \beta_j} = \frac{y - \mu}{var(Y)} b''(\theta) x_{ij} = \frac{y - \mu}{a(\phi)} x_{ij}$$

Uporabimo $\mathrm{E}(\frac{\partial^2 L}{\partial \theta^2}) = -\mathrm{E}((\frac{\partial L}{\partial \theta})^2)$:

$$\begin{aligned} -\mathbf{E}(\frac{\partial^2 L}{\partial \beta_j \partial \beta_k}) &= \mathbf{E}((\frac{\partial L}{\partial \beta_j})(\frac{\partial L}{\partial \beta_k})) \\ &= \mathbf{E}(\frac{y - \mu}{var(Y)^2})(\frac{\partial \eta}{\partial \mu})^2 x_{ij} x_{ik} \\ &= \frac{1}{var(Y)}(\frac{\partial \eta}{\partial \mu})^2 x_{ij} x_{ik} \\ &= \frac{b''(\theta)}{a(\phi)} x_{ij} x_{ik} \end{aligned}$$

Po drugi strani pa je odvod funkcije zbira

$$\begin{aligned}\frac{\partial^2 L}{\partial\beta_j\partial\beta_k} &= \frac{\partial}{\beta_k}\{\left(\frac{\partial L}{\partial\theta}\right)\left(\frac{\partial\theta}{\partial\beta_j}\right)\}\\ &= \frac{\partial}{\partial\theta}\left(\frac{\partial^2\theta}{\partial\beta_j\partial\beta_k}\right)+\left(\frac{\partial\theta}{\partial\beta_j}\right)\left(\frac{\partial^2 L}{\partial\theta^2}\frac{\partial\theta}{\partial\beta_k}\right)\\ &= 0+\frac{\partial^2 L}{\partial\theta^2}x_{ij}x_{ik},\end{aligned}$$

videli pa smo že da je

$$\frac{\partial^2 L}{\partial\theta^2} = -\frac{b''(\theta)}{a(\phi)}.$$

Sledi torej, da za kanonično povezovalno funkcijo Fisher-scoring in Newton Raphson sovpadata!

Še več, iz predavanj se spomnimo da je

$$FI(\theta) = var(\frac{\partial}{\partial \theta} L)$$

==> Hessejeva matrika je za kanonično povezovalno funkcijo pozitivno definitna in Fisher scoring metoda je naraščajoča!

Enačbe verjetja v logističnem modelu

Imejmo slučajni vektor $Y = (Y_1, \ldots, Y_n)$ z NEP komponentami porazdeljenimi binomsko

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{m_i - y_i}$$

Funkcija verjetja se glasi:

$$\begin{aligned}
\ell(p_i) &= \log\{\prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{m_i - y_i}\} \\
&= \sum_{i=1}^{n} \{y_i \log p_i + (m_i - y_i) \log(1 - p_i)\} \\
&= \sum_{i=1}^{n} \{m_i \log(1 - p_i) + y_i \log\left(\frac{p_i}{1 - p_i}\right)\}
\end{aligned}$$

Enačbe verjetja v logističnem modelu

Upoštevamo še $\log \frac{p_i}{1-p_i} = x_i^\top \beta$ in dobimo

$$\ell(\beta) = \sum_{i=1}^{n} \left(y_i (x_i^\top \beta) - m_i \log(1 + \exp x_i^\top \beta) \right)$$

Od tu vidimo, da je res odvisna le od parametra β

Enačbe verjetja v logističnem modelu

Sedaj potrebujemo še odvode.

$$\begin{aligned}\frac{\partial}{\partial \beta_j}(x_i^\top \beta) &= \frac{\partial}{\partial \beta_j}\left(\beta_0 + x_{i1}\beta_1 + \ldots x_{ir}\beta_r\right) \\ &= x_{ij}\end{aligned}$$

Enačbe verjetja v logističnem modelu

$$\begin{aligned}\frac{\partial}{\partial \beta_j} \log(1 + \exp(x_i^\top \beta)) &= \frac{\frac{\partial}{\partial \beta_j} \exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)} \\ &= \frac{\exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)} \frac{\partial}{\partial \beta_j} (x_i^\top \beta) \\ &= p_i(\beta) x_{ij},\end{aligned}$$

kjer smo upoštevali $p_i = \frac{\exp x_i^\top \beta}{1 + \exp x_i^\top \beta}$.

Iščemo torej ničlo

$$\frac{\partial}{\partial \beta_j} \ell(\beta) = \sum_{i=1}^{n} \left(x_{ij}(y_i - m_i p_i(\beta)) \right)$$

za $j = 0, \ldots, r$

Enačbe verjetja v logističnem modelu

Za uporabo Newtonove metode bomo potrebovali še drugi odvod, za kar moramo izračunati še

$$\begin{aligned}\frac{\partial p_i(\beta)}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k} \frac{\exp x_i^\top \beta}{1 + \exp x_i^\top \beta} \\ &= x_{ik} p_i(\beta)(1 - p_i(\beta))\end{aligned}$$

in sestaviti to v

$$\frac{\partial^2}{\partial \beta_j \partial \beta_k} \ell(\beta) = -\sum_i^n \left(x_{ij} x_{ik} m_i p_i(\beta)(1 - p_i(\beta)) \right)$$

za $j, k = 0, 1, \ldots, r$, kar pa lahko poenostavimo v

$$\ddot{\ell}(\beta) = -\sum_{i=1}^n \left(x_{ij} x_{ik} v_i(\beta) \right)$$

Enačbe verjetja v logističnem modelu

Preglednejši in priročnejši je zapis v matrični obliki:

$$\ell(\beta) = y^\top \mathbf{X}\beta - n^\top \log(1 + \exp \mathbf{X}\beta)$$
$$\dot{\ell}(\beta) = \mathbf{X}^\top (y - m \circ p(\beta)) = \mathbf{X}^\top (y - m \circ p(\beta)) = X^\top (y - \mu(\beta))$$

Za drugi odvod definirajmo diagonalno matriko
$v(\beta) = \text{diag}\{m_1 p_1(1 - p_1), \ldots, m_n p_n(1 - p_n)\}$ in povzemimo

$$\ddot{\ell}(\beta) = -\mathbf{X}^\top v(\beta)\mathbf{X}$$

Fisher scoring za logistični model

$$\hat{\beta}_{i+1} = \hat{\beta}_i + (X^T v(\hat{\beta}_i) X)^{-1} X^T (y - \mu(\hat{\beta}_i))$$
$$= \hat{\beta}_i + (\text{inverz info})(\text{score})$$

Fisher scoring za logistični model

$$\hat{\beta}_{i+1} = \hat{\beta}_i + (X^T v(\hat{\beta}_i) X)^{-1} X^T (y - \mu(\hat{\beta}_i))$$
$$= \hat{\beta}_i + (\text{inverz info})(\text{score})$$

Računanje inverza je lahko problematično. To rešimo takole:

$$h = \hat{\beta}_{i+1} - \hat{\beta}_i$$
$$X^T v(\hat{\beta}_i) X = h * X^T (y - \mu(\hat{\beta}_i))$$

References I