Empīriskā ticamības funkcija un regresija

Māra Vēliņa

Latvijas Universitāte

10.11.2010

Saturs

EL un parametriskā regresija

Pirmoreiz EL metodi lineārai regresijai apskatīja Owen (1991).

Aplūkojam regresijas modeli formā

$Y_i = m(X_i; \beta) + \epsilon_i,\ i = 1, \ldots, n,$
kur $m(x; \beta)$ - zināma regresijas funkcija ar nezināmu parametru $\beta \in \mathbb{R}^p$ $(p < n)$ un ϵ_i - neatkarīgi g.l., kam spēkā $E(\epsilon_i|X_i) = 0$ un $D(\epsilon_i|X_i) = \sigma^2(X_i)$.

Parametriskā regresijas funkcija iekļauj gadījumus:

- Lineārā regresija $m(x; \beta) = x^T\beta$;
- Visparinātais lineārais modelis (McCullagh Nelder, 1989) ar $m(x; \beta) = G(x^T\beta)$ un $\sigma^2(x) = \sigma_0^2 D(G(x^T\beta))$ zināmai saites funkcijai G, zināmai dispersijas funkcijai $D(.)$ un nezināmai konstantei $\sigma_0^2 > 0$.

- Parametra β mazāko kvadrātu (MK) novērtējumu iegūst, minimizējot funkciju

$$S_n(\beta) := \sum_{i=1}^{n} (Y_i - m(X_i;\beta))^2.$$

- β MK novērtējums $\hat{\beta}_{LS} = \arg\inf_\beta S_n\beta$, ir sekojoša vienādojuma atrisinājums:

$$\sum_{i=1}^{n} \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta)) = 0.$$

Saskaņā ar Owen(1988) un (1991) EL funkcija parametram β ir formā

$$L_n = \max \prod_{i=1}^{n} p_i \tag{1}$$

pie ierobežojumiem

$$\sum_{i=1}^{n} p_i = 1, \tag{2}$$

$$\sum_{i=1}^{n} p_i \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta)) = 0. \tag{3}$$

- Iegūst optimizācijas problēmu

$$T(p,\lambda_0,\lambda_1)=\sum_{i=1}^{n}\log p_i+\lambda_0(\sum_{i=1}^{n}p_i-1)+\lambda\sum_{i=1}^{n}p_i\frac{\partial m(X_i;\beta)}{\partial\beta}(Y_i-m(X_i;\beta)),$$

kur $p=(p_1,\cdots,p_n)^T$.

- Var parādīt, ka $\lambda_0=-n$ un, definējot $\lambda=-n\lambda_1$, optimālos p_i var izteikt formā

$$p_i=\frac{1}{n}\frac{1}{1+\lambda^T\frac{\partial m(X_i;\beta)}{\partial\beta}(Y_i-m(X_i;\beta))},$$

kur λ spēkā

$$\sum_{i=1}^{n}\frac{\frac{\partial m(X_i;\beta)}{\partial\beta}(Y_i-m(X_i;\beta))}{1+\lambda^T\frac{\partial m(X_i;\beta)}{\partial\beta}(Y_i-m(X_i;\beta))}=0 \qquad (4)$$

Seko, ka EL pieņem formu

$$L_n(\beta) = \prod_{i=1}^{n} \frac{1}{n} \frac{1}{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta))}$$

Logaritmiskā EL

$$\log\{L_n(\beta)\} = -\sum_{i=1}^{n} \log\{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta))\} - n \log n \tag{5}$$

- Lai novērtētu EL, jārisina attiecībā pret λ nelineārs vienādojums, kas ir atkarīgs no β
- Var risināt duālo optimizācijas problēmu (Owen, 1990)

Īpaši rezultāti:

- No 5 var definēt logaritmisko EL attiecību

$$r_n(\beta) = -2\log\{L_n(\beta)/Ln(\hat{\beta})\} = 2\sum_{i=1}^{n} \log\{1+\lambda^T \frac{\partial m(X_i;\beta)}{\partial\beta}(Y_i - m(X_i;\beta))$$

- Owen (1991) parādīja, ka EL lineārai regresijai ir spēkā Vilksa teorēma (Wilks, 1938):

$$r_n(\beta_0) \xrightarrow{d} \chi^2_p, \text{ kad } n \to \infty.$$

- Izmantojot Vilksa teorēmu, var konstruēt EL ticamības apgabalus regresijas parametram β_0 (ar ticamības līmeni $(1-\alpha)$):

$$I_{1-\alpha} = \{\beta : r_n(\beta) \le \chi^n_{p,1-\alpha}\},$$

kur $\chi^n_{p,1-\alpha}$ ir $\chi^2_{p,1-\alpha}$ sadalījuma $(1-\alpha)$–kvantile.

Chen (1993, 1994) ieviesa Bartleta korekciju lineārajai regresijai.

- Tika parādīts, ka parametriskai regresijai gan empīriskās gan parametriskās ticamības attiecības apgabaliem $\mathrm{I}_{1-\alpha}$ pārklājuma kļūdas kārta ir n^{-1}.
- Bartleta korekcija samazina pārklājuma kļūdu par vienu kārtu:

$$\mathrm{P}\{\mathrm{r}_{\mathrm{n}}^{*}(\beta_0) \le \chi^2_{\mathrm{p},1-\alpha}\} = 1 - \alpha + \mathrm{O}(\mathrm{n}^{-2})$$

.

EL un neparametriskā regresija

EL neparamteriskajai regresijai apskatīja Chen un Hall (2000, lokālais lineārais kodolu novērtējums), un Chen un Qin (2003, Nadaraya-Watson kodolu novērtējums)

- Aplūko neparametriskās regresijas modeli

$$Y_i = m(X_i) + \epsilon_i,$$

kur $m(Y_i|X_i = x)$ ir neparametriska regresijas funkcija,
X_i ir $d-$dimensionāls un
$\sigma^2(x) = D(Y_i|X_i = x)$.

Plaši pazīstamais $\mathrm{m}(\mathrm{x})$ kodolu regresijas novērtējums:

$$\hat{\mathrm{m}}(\mathrm{x}) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{K}_{\mathrm{h}}(\mathrm{x}-\mathrm{X}_{\mathrm{i}})\mathrm{Y}_{\mathrm{i}}}{\sum_{\mathrm{n}=1}^{\mathrm{n}} \mathrm{K}_{\mathrm{h}}(\mathrm{x}-\mathrm{X}_{\mathrm{i}})},$$

kur $\mathrm{K}_{\mathrm{h}}(\mathrm{t}) = \mathrm{K}(\mathrm{t}/\mathrm{h})/\mathrm{h}^{\mathrm{d}}$, un K ir $\mathrm{d}-$dimensionāla kodolu funkcija. $\hat{\mathrm{m}}(\mathrm{x})$ var iegūt kā atrisinājumu vienādojumu sistēmai

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{K}_{\mathrm{h}}(\mathrm{x}-\mathrm{X}_{\mathrm{i}})\{\mathrm{Y}_{\mathrm{i}}-\mathrm{m}(\mathrm{x})\} = 0. \tag{6}$$

EL regresijas funkcijai $m(x)$ pie fiksēta x:

(6) motivē EL definēt sekojošā veidā:

$$L_n\{\theta(x)\} = \max \prod_{i=1}^{n} p_i$$

ar ierobežojumiem

$$\sum_{i=1}^{n} p_i = 1$$

un

$$\sum_{i=1}^{n} p_i K_h(x - X_i)\{Y_i - \theta(x)\} = 0$$

.

- EL novērtējums punktā $\theta(x)$ ir

$$L_n\{\theta(x)\} = \prod_{i=1}^{n} \frac{1}{n} \frac{1}{1 + \lambda(x)K_h(x - X_i)\{Y_i - \theta(x)\}}.$$

- Logaritmiskā EL ir formā

$$\log\{L_n\{\theta(x)\}\} = -\sum_{n=1}^{n} \log[1 + \lambda(x)K_h(x - X_i)\{Y_i - \theta(x)\}].$$

- EL tiek maksimizēta punktā $p_i = n^{-1}$, no kā seko, ka $\theta(x)$ ir Nadaraya-Watson novērtējums $\hat{m}(x)$, un var definēt

$$r_n\{\theta(x)\} = -2\log[L_n\{\theta(x)\}/n^{-n}] =$$

$$= 2\sum_{n=1}^{n} \log[1 + \lambda(x)K_h(x - X_i)\{Y_i - \theta(x)\}].$$

Piezīme

Aplūkotā EL attiecas nevis uz īsto funkciju $\mathrm{m}(\mathrm{x})$, bet gan $\mathrm{E}\{\hat{\mathrm{m}}(\mathrm{x})\} = \mathrm{m}(\mathrm{x}) + \mathrm{bias}$. Lai pārietu uz $\mathrm{m}(\mathrm{x})$ EL, var veikt korekciju, atsevišķi novērtējot biasu (Hall, 1991), vai samazināt biasu, veicot nepietiekamu nogludināšanu (Neumann, 1995).

Vilksa teorēma

Lietojot nepietiekamu nogludināšanu tādu, ka $\mathrm{n}^{2/(4+\mathrm{d})\mathrm{h}^2} \to 0$, aplūkotajai neparametriskajai regresijai ir spēkā Vilksa teorēma:

$$\mathrm{r_n}\{\mathrm{m}(\mathrm{x})\} \xrightarrow{\mathrm{d}} \chi^2_1, \text{ kad } \mathrm{n} \to \infty.$$

Tātad, neparametriskās regresijas ticamības intervāli ar pārklājuma precizitāti $(1-\alpha)$ ir formā

$$\mathrm{I}_{1-\alpha,\mathrm{el}} = \{\theta(\mathrm{x}) : \mathrm{r_n}\{\theta(\mathrm{x})\} \leq \chi^2_{1,1-\alpha}\}.$$

- Vienmērīgās ticamības joslas
  $\mathrm{I}_{1-\alpha,\mathrm{el}}$ ir punktveida ticamības intervāli. Zhu, Lin un Chen (2010) neparametriskajai regresijai konstruē globālās EL ticamības joslas, kas ir modeļu un datu adaptīvas.
- Bartleta korekcija
  Chen un Qin (2003) parādīja $\mathrm{I}_{1-\alpha,\mathrm{el}}$ pārklājuma varbūtības Edgeworth izvirzījumus, un ka EL arī neparametriskās regresijas kontekstā ir spēkā Bartleta korekcija.

Single-index regresijas modeļi

- Aplūko modeli
$$Y_i = g(\beta^T X_i) + \epsilon_i,$$
kur Y_i - 1-dimensijas atbildes mainīgais, X_i - p-dimensionālu skaidrojošo mainīgo vektors;
g-nezināma, gluda funkcija, un $E(\epsilon_i|X_i) = 0$, $D(\epsilon_i|X_i) = \sigma^2$.
β_0 - īsto parametru vektors; pieņem, ka $||\beta|| = 1$.
- Katram $\beta = (\beta_1, \dots, \beta_p)^T$: $||\beta|| = 1$ un katram $1 \leq r \leq p$ definē $\beta^r = (\beta_1, \beta_2, \dots, \beta_{r-1}, \beta_{r+1}, \dots, \beta_p)^T$. Tad
- Jakobiāna matrica
$$J_{\beta^{(r)}} = \frac{\partial \beta}{\partial \beta^{(r)}} = (\gamma_1, \dots, \gamma_p)^T$$
, kur $\gamma_s (s \neq r)$ - vienības vektors un
$\gamma_r = -(1 - ||\beta^{(r)}||^2)^{-1/2} \beta^{(r)}$.

- Seko, ka $E[\xi_i(\beta_0^{(r)})] = 0, (i = 1, \ldots, n)$, kur

$$\xi(\beta^{(r)}) = [Y_i - g(\beta^T X_i)]g'(\beta^T X_i)J^T_{\beta^{(r)}} X_i$$

- Aizstājot nezināmās funkcijas g un g' ar to lokālajiem lineārajiem novērtējumiem $\hat{g}$ un $\hat{g}'$, EL ir formā

$$R_n(\beta^{(r)}) = \max \prod_{i=1}^{n} (np_i),$$

pie ierobežojumiem $p_i \geq 0 (i = 1, \ldots, n)$, $\sum_{i=1}^n p_i = 1$, un $\sum_{i=1}^n p_i \hat{\xi}_i(\beta^{(r)}) = 0$.

- Xue, Zhu (2006) parādīja, ka pie noteiktiem regularitātes nosacījumiem

$$-2 \log R_n(\beta_0^{(r)}) \xrightarrow{d} \omega_1 \chi^2_{1,1} + \cdots + \omega_{p-1} \chi^2_{1,p-1},$$

noteiktiem svariem $\omega_1, \ldots, \omega_{p-1}$, un $\chi^2_{1,1}, \ldots, \chi^2_{1,p-1}$ neatkarīgi χ^2_1 mainīgie.

EL un regresija - citi lietojumi

- EL metodi var vispārināt arī regresijai trūkstošu datu gadījumā (gan atbildes, gan skaidrojošiem mainīgiem).
- EL metodi var vispārināt regresijai uz cenzētiem datiem, t.i., kur netiek novērota atbilde Y_i, bet gan $T_i = \min(Y_i, C_i)$, kur C_i ir cenzētais mainīgais.
- EL var izmantot, lai konstruētu goodness-of-fit testus, piemēram, parametriskai laikrindu analīzes regresijai (Chen, Hardle, Li 2003), mainīgu koeficientu regresijas modelim (Fan, Zhang, Zhang 2001).