LSM

Bojana Milošević

11/25/2019

Klasifikacija

- Predvideti zavisnu promenljivu
- Šta utiče na raspodelu zavisne promenljive?

$$Y_i|X_i : \begin{pmatrix} 0 & 1 \\ 1-\pi(X_i) & \pi(X_i) \end{pmatrix}$$

Regresiona funkcija je

$$E(Y_i|X_i) = \pi(X_i)$$

Zašto linearni model nije adekvatan?

1. Greške modela ne mogu se modelirati normalnom raspodelom, ili nekom drugom absolutno neprekidnom i smetričnom oko nule.
2. Disperzija grešaka modela nije konstantna. Važi: $D(Y_i|X_i) = \pi(X_i)(1 - \pi(X_i)) = D(\varepsilon_i)$.
3. Regresiona funkcija verovatnoća pa treba da bude zadovoljeno da je $\pi(X_i) \in [0, 1]$.

Transformacija modela

$$F^{-1}(\pi(X_i)) = X\beta$$

1. $F(x) = \Phi(x)$ PROBIT regresija
2. $F(x) = \frac{1}{1+e^{-x}}$ LOGISTIČKA regresija
3. $F(x) = 1 - e^{-e^x}$ LOG-VEJBULOVA regresija

Probit regresija

$$\Phi^{-1}(\pi(X_i)) = \beta_0^* + \beta_1^* X_i$$

Pretpostavimo da ispitujemo zavisnost temperature Y od vlažnosti vazduha X i da se proglašava vanredno stanje ukoliko temperatura pređe neki kritični nivo C. Neka je Y_c indikator vanrednog stanja. Pod pretpostavkom da je $Y_i = aX_i + b + \varepsilon_i$, gde $\{\varepsilon_i\}$ je Gausov beli shum, modeliranje $E(Y_c|X)$ se svodi na probit regresiju.

Parametri se mogu oceniti metodom maksimalne verodostojnosti.

Log – Vejbulova regresija

$$F_G^{-1}(\pi(X_i)) = \log(-\log(1-\pi(X_i))) = \beta_0 + \beta_1 X_1$$

zbog svoje asimetričnosti, se najčešće koristi za modelovanje malih i velikih verovatnoća uspeha

Parametri se mogu oceniti metodom maksimalne verodostojnosti.

Logistička regresija

$$F_L^{-1}(\pi(X_i)) = \log\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right) = \beta_0 + \beta_1 X_i$$

$\lambda(p) = \log\left(\frac{p}{1-p}\right)$ logit transformacija

$\lambda(X_i) = \log\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right)$

Količnik $\frac{\pi(X_i)}{1-\pi(X_i)}$ se naziva *kvota.*

Interpretacija kvote

Ocena parametara

Metod maksimalne verodostojnosti

$$l(\beta) = \sum_{i=1}^{n} \left(Y_i \log \left(\frac{\pi(X_i)}{1-\pi(X_i)} \right) + \log(1-\pi(X_i)) \right)$$
$$= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log\left(1 + e^{\beta_0+\beta_1 X_i}\right)$$

Numerički se rešava sistem $\frac{\partial l(\beta)}{\partial \beta} = 0$

Ocenjena logit funkcija je

$$\hat{\lambda}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

Ocena parametara

U opštem slučaju

$$\hat{\lambda}(X) = X\hat{\beta}$$

$\hat{\beta}$ ima normalnu $\mathcal{N}(\beta, I^{-1}(\beta))$ raspodelu kao ocena maksimalne verodostojnosti pa se mogu napraviti intervali poverenja za $\lambda(X)$ a zatim i za $\pi(X)$.

Testiranje značajnosti koeficijenta

Valdova statistika

$$Z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

- test količnika verodostojnosti

$$2\log\left(\frac{L(\hat{\beta}_0, \hat{\beta}_1)}{L(\hat{\beta}_0)}\right) \sim \chi^2_1$$

$$2\log\left(\frac{L(\hat{\beta})}{L(\hat{\beta}_0)}\right) \sim \chi^2_q$$

broj koeficijenata za koje se pretpostavlja da su 0

Devijacija

- meri razliku između pretpostavljenog modela i saturiranog modela
- Definiše se sa $D_0 = 2(l(y, \hat{\theta}_s) - l(y, \hat{\beta}_0))$ gde je $\hat{\theta}_s$ ocena u saturiranom modelu
- $D = 2(l(y, \hat{\theta}_s) - l(y, \hat{\beta}))$
- $D_0 - D \sim \chi^2_q$

Slučaj grupisanih podataka

najčešće kad je neki prediktor kategorička promenljiva

Za svako X_i iz uzorka formiramo podskup koji čine oni elementi uzorka čija je nezavisna komponenta jednaka odabranom X_i.

Neka je m_j broj elemenata u j-toj podgrupi posmatranog uzorka, $j = 1, 2...J$.

U okviru svake podgrupe ocenimo $\pi(X_j) = P\{Y = 1|X_j\}$. Neka je n_j proj elemenata u podgrupi za koje je vrednost zavisne promenljive jednaka 1. Tada je $\hat{\pi}\{Y = 1|X_j\} = \frac{1}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_j}}$. Očekivan broj elemenata iz svake od grupa:

$$\hat{n}_j = m_j\hat{\pi}_j = m_j\frac{1}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X_j}}.$$

Pirsonovi reziduali

$$r_j = \frac{n_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}} = \frac{n_j - \hat{n}_j}{\sqrt{\hat{n}_j(1-\frac{\hat{n}_j}{m_j})}}.$$

$$C = \sum_{j=1}^{J} r_j^2.$$

C ima približno χ^2_{J-2}. Kada se u modelu javlja $p+1$ ocenjen parametar onda C ima približno χ^2_{J-p-1} raspodelu

Reziduali devijacije

Rezidual devijacije, za $n_j - \hat{n}_j > 0$, je definisan sa:

$$d_j = \sqrt{2\left(n_j \ln \frac{n_j}{m_j \hat{\pi}_j} + (m_j - n_j) \ln \frac{m_j - n_j}{m_j(1 - \hat{\pi}_j)}\right)}$$

$$= \sqrt{2\left(n_j \ln \frac{n_j}{\hat{n}_j} + (m_j - n_j) \ln \frac{m_j - n_j}{m_j - \hat{n}_j}\right)}$$

Za $n_j - \hat{n}_j < 0$ za j-ti rezidual se uzima $-d_j$, u suprotnom nula.

Specijalni slučajevi:

za $n_j = 0$

$$d_j = -\sqrt{2m_j \left| \ln \frac{m_j}{m_j - \hat{n}_j} \right|},$$

dok je za $n_j = m_j$

$$d_j = \sqrt{2m_j \left| \ln \frac{m_j}{\hat{n}_j} \right|}.$$

Reziduali devijacije

Test statistika je

$$D = \sum_{j=1}^{J} d_j^2.$$

D ima približno χ^2_{J-2} raspodelu. Kada se u modelu javlja $p+1$ ocenjen parametar onda D ima približno χ^2_{J-p-1} raspodelu. Primetimo da je D zapravo devijacija modela.

Hosmer-Lemešov test

- Podaci se grupišu u g kategorija na osnovu sličnosti ocenjenih verovatnoća. Granice koje određuju grupu se dobijaju kao odgovarajući kvantili. Npr. prva grupa sadrži sve elemente za koje je ocenjena verovatnoća između 0 i 0.1, druga, od 0.1 do 0.2 itd. Statstika se pravi analogno Pirsonovoj u slučaju grupisanih podataka

$$C = \sum_{k=1}^{g} \frac{(o_k - M_k \bar{\pi}_k)^2}{M_k \bar{\pi}_k (1 - \bar{\pi}_k)},$$

gde je M_k broj elemenata u k-toj grupi, c_k je broj različitih elemenata u $k - toj$ grupi i $o_k = \sum_{j=1}^{c_k} Y_j$ i

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_k}{M_k}$$

Ukoliko je dobar model C ima χ^2_{g-2} raspodelu