Uogolnione modele liniowe

Jerzy Mycielski

Uniwersytet Warszawski

grudzien 2013

Uogólniony model liniowy (generalized linear model - glm)

- Zakładamy, że
$$g\left[\mathrm{E}\left(y_i | \mathbf{x}_i\right)\right] = \mathbf{x}_i \boldsymbol{\beta}$$
- Inaczej mówiąc:
$$\mathrm{E}\left(y_i | \mathbf{x}_i\right) = g^{-1}\left(\mathbf{x}_i \boldsymbol{\beta}\right)$$
- Funkcję $g\left(\cdot\right)$ nazywamy funkcją związku (link function)
- O zmiennej y_i zakładamy, że należy do rodziny dystrybuant wykładniczych (exponential family)

Rodzina dystrybuant wykładniczych

- Dystrybuanta należy do rodziny dystrybuant wykładniczych jeśli jej funkcję gęstości (prawdopodobieństwa) można przedstawić w następujący sposób:

$$f_{\boldsymbol{\theta}}(y) = h(y) \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta}) \mathbf{T}(y) - A(\boldsymbol{\theta})\right)$$

- Jeśli dystrybuanta jest sparametryzowana w ten sposób, że $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ to mówimy, że dystrybuanta ma formę kanoniczną a parametr $\boldsymbol{\eta}$ nazywamy naturalnym (kanonicznym) parametrem rokładu
- $\mathbf{T}(y)$ jest statystyką dostateczną, $A(\boldsymbol{\theta})$ jest parametrem logpodziału (log partition parameter)
- Funkcja związku jest kanoniczna jeśli $\boldsymbol{\eta}(\boldsymbol{\mu}) = g(\boldsymbol{\mu})$ a więc jest tożsama z funkcją, która przekształca parametr związany z wartością oczekiwaną na parametr kanoniczny rozkładu

- Statystyka dostateczna jest to funkcja próby $\mathbf{y} = [y_1, \ldots, y_n]$, że jeśli $\mathbf{T}(\mathbf{y}_A) = \mathbf{T}(\mathbf{y}_B)$

$$\frac{f_{\boldsymbol{\theta}}(\mathbf{y}_A)}{f_{\boldsymbol{\theta}}(\mathbf{y}_B)} = 1 \text{ dla każdego } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

- Łatwo zauważyć, że dla dystrybuant z rodziny rozkładów wykładniczych $\mathbf{T}(\mathbf{y}) = \sum_{i=1}^{n} \mathbf{T}(y_i)$
- Statystyka dostateczna zawiera w sobie całą istotną informację zawartą w próbie na temat nieznanego parametru $\boldsymbol{\theta}$
  - w związku z tym estymator efektywny jest zawsze funkcją statystyki dostatecznej

Rodzina dystrybuant wykładniczych

Dystrybuanty

- Elementami są między innymi dystrybunaty:
  - dwumianowy
  - Poissona
  - ujemny dwumianowy
  - normalny
  - odwrotny gausowski
  - gamma.

- rozkład Bernoulliego: $T(y) = y$, $h(y) = 1$, $\eta(p) = \ln \frac{p}{1-p}$, $A(\eta) = \ln(1 + \exp\eta)$

$$\Pr(y) = p^y (1-p)^{1-y} = \exp\left[\eta y - \ln(1 + \exp\eta)\right]$$

- Rzeczywiście:

$$\exp\left[\eta y + \ln(1 + \exp\eta)\right] = \exp(\eta y) \exp\left[-\ln(1 + \exp\eta)\right]$$
$$= \left(\frac{p}{1-p}\right)^y \left(\frac{1}{1 + \frac{p}{1-p}}\right) = \exp(\eta y) \left(\frac{1}{1 + \exp\eta}\right)$$
$$= \left(\frac{p}{1-p}\right)^y (1-p) = p^y (1-p)^{1-y}$$

Dystrybuanta wykładniczych z nadmierną dyspersją

- Uogólnieniem rodziny dystrybuant wykładniczych jest rodzina dystrybuant wykładniczych z nadmierną dyspersją (overdispersed exponential family)

$$f_{\boldsymbol{\theta},\boldsymbol{\tau}}(y) = h(y) \exp\left(\frac{\boldsymbol{\eta}(\boldsymbol{\theta}) \mathbf{T}(y) - A(\boldsymbol{\theta})}{d(\boldsymbol{\tau})} \right)$$

- Funkcje związku determinują relację między wartością oczekiwaną zmiennej zależnej y_i a wielkościami zmiennych niezależnych
- W zależności od typu funkcji związku różnić się będzie interpretacja parametrów
- Popularne postaci funkcji związku
  - identity: $g(\mu) = \mu$
  - log: $g(\mu) = \ln(\mu)$
  - logit: $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$
  - probit: $g(\mu) = \Phi^{-1}(\mu)$
  - cloglog: $g(\mu) = \ln\left[-\ln(1-\mu)\right]$
  - opower: $g(\mu) = \left[\left(\frac{\mu}{1-\mu}\right)^n - 1\right] \Big/ n$
  - power: $g(\mu) = \mu^n$
  - nbinominal: $g(\mu) = \ln\left(\frac{\mu}{\mu+k}\right)$
  - loglog: $g(\mu) = -\ln(-\ln \mu)$
  - logc: $g(\mu) = \ln(1-\mu)$

- Modele z zerojedynkową zmienną zależną
  - rozkład zmiennej zależnej y: Bernoulliego

$$y_i = \begin{cases} 1 & \text{z p-stwem} \quad p \\ 0 & \text{z p-stwem} \quad 1-p \end{cases}$$

- Wartość oczekiwana w z rozkładzie dwumianowym $\mathrm{E}\left(y_i\right) = p$

Uogólniony model liniowy
Przykłady c.d.

- Model logitowy
  - funkcja związku: logit

$$E\left(y_i | \mathbf{x}_i\right) = \mu_i = \frac{\exp\left(\mathbf{x}_i \boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{x}_i \boldsymbol{\beta}\right)}$$

$$g\left[E\left(y_i | \mathbf{x}_i\right)\right] = \ln\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{x}_i \boldsymbol{\beta}$$

-
  - zauważmy, że w tym przypadku funkcja związku jest kanoniczna
- Model probitowy
  - funkcja związku: probit

$$E\left(y_i | \mathbf{x}_i\right) = \mu_i = \Phi\left(\mathbf{x}_i \boldsymbol{\beta}\right)$$

$$g\left[E\left(y_i | \mathbf{x}_i\right)\right] = \Phi^{-1}\left(\mu_i\right) = \mathbf{x}_i \boldsymbol{\beta}$$

	gauss.	igauss.	binominal	Poisson	nbinomina	gamma
id	x	x	x	x	x	x
log	x	x	x	x	x	x
logit			x			
probit			x			
clog			x			
pow			x			
opower			x			
nbinominal					x	
loglog			x			
logc			x			

- Najbardziej popularna jest MNW
- Inną możliwością jest zastosowanie ważonej nieliniowej metody najmniejszych kwadratów (WNMNK)
- Estymujemy nieliniwe równanie regresji

$$y_i = g^{-1}\left(\mathbf{x}_i \boldsymbol{\beta}\right) + \varepsilon_i$$

- W przypadku zmiennych losowych o dystrybuancie z rodziny rozkładów wykładniczych, wariancja jest z reguły funkcją wartości oczekiwanej $\mathrm{Var}\left(y_i \middle| \mathbf{x}_i\right) = w_i = w\left(\boldsymbol{\mu}_i\right)$
  - np. w przypadku rozkładu Bernoulliego $w\left(\boldsymbol{\mu}_i\right) = \mathrm{Var}\left(y_i \middle| \mathbf{x}_i\right) = \mathrm{Var}\left(\varepsilon_i \middle| \mathbf{x}_i\right) = \boldsymbol{\mu}_i\left(1 - \boldsymbol{\mu}_i\right)$
- Estymatorem efektywnym w takim przypadku jest WNMNK

- Funkcja kryterium:

$$\min_{\beta} \sum_{i=1}^{n} w_i^{-1} \left[y_i - g^{-1} \left(\mathbf{x}_i \boldsymbol{\beta} \right) \right]^2$$

  gdzie w_i jest proporcjonalne do wielkości wariancji.
- Ponieważ wariancja $w_i = w\left(\boldsymbol{\mu}_i\right) = w\left(g^{-1}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)$ jest funkcją nieznanych parametrów więc stosujemy wersję iterowaną WNMNK.

1. Uzyskujemy wstępne oszacowania $\widehat{\boldsymbol{\beta}}_{(0)}$ dla wag $w_i = 1$
2. Znajdujemy oszacowania wariancji $\widehat{w}_{i,(k)} = w\left(g^{-1}\left(\mathbf{x}_i \widehat{\boldsymbol{\beta}}_{(k-1)}\right)\right)$
3. Szacujemy model przy użyciu wag $\widehat{w}_{i,(k)}$, uzyskujemy oszacowania $\widehat{\boldsymbol{\beta}}_{(k)}$
4. Jeśli kryterium zbieżności nie jest spełnione przechodzimy do kroku 2.

- Zakładamy, że

$$\mathrm{E}\left(y_{it} \mid \mathbf{x}_{it}, c_i\right) = g^{-1}\left(\mathbf{x}_{it}\boldsymbol{\beta} + c_i\right)$$

oraz, że $\mathbf{x}_i$ i c_i są niezależne

- W przypadku estymowania modelu GLM z efektami losowymi na próbie panelowej można zastosować MNW
  - w tym przypadku do obliczania funkcji wiarygodności musimy zastosować kwadratury, ponieważ funkcja wiarygodności ma postać

$$f\left(\mathbf{y}_i \mid \mathbf{x}_i\right) = \int_{-\infty}^{\infty} \prod_{t=1}^{T} f\left(\mathbf{x}_{it}\boldsymbol{\beta} + c_i \mid c_i\right) f\left(c_i\right) dc_i$$

$$\ell\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{N} \ln f\left(\mathbf{y}_i \mid \mathbf{x}_i\right)$$

  - metoda ta może okazać się w praktyce bardzo czasochłonna
- Alternatywą jest zastosowanie Ogólnych Równań Estymacyjnych (GEE)

Uogólniony model liniowy

GEE (general estimating equations)

- Specyfikujemy model dla warunkowej średniej oczekiwanej uśrednionej po wielkościach efektów indywidualnych

$$\mathrm{E}\left(y_{it} \middle| \mathbf{x}_{it}\right) = \mathrm{E}_{c_i}\left[\mathrm{E}\left(y_{it} \middle| \mathbf{x}_{it}, c_i\right)\right] = g_*^{-1}\left(\mathbf{x}_{it}\boldsymbol{\beta}\right) = \boldsymbol{\mu}_{it}$$

- Model taki określamy jako model uśredniony po populacji (PPA - population averaged model)
- Minimalizowana funkcja kryterium

$$\min_{\boldsymbol{\beta},\boldsymbol{\alpha}} \sum_{i=1}^{N} \left[\mathbf{y}_i - \boldsymbol{\mu}_i\left(\boldsymbol{\beta}\right)\right]' \left[\mathbf{W}\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right)\right]^{-1} \left[\mathbf{y}_i - \boldsymbol{\mu}_i\left(\boldsymbol{\beta}\right)\right]$$

- Nawet w przypadku błędnego wyspecyfikowania modelu dla wariancji $\mathbf{W}\left(\boldsymbol{\alpha}\right)$ estymator ten jest estymatorem zgodnym
- Estymator macierzy wariancji oszacowań $\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right)$ można zbudować tak, by był zgodny w przypadku błędnej specyfikacji $\mathbf{W}\left(\boldsymbol{\alpha}\right)$ (wersja estymatora odpornego Hubera/White'a)

- Poza szczególnym przypadkiem identycznościowej funkcji związku fukcja związku stosowane w GEE nie jest tożsama z funkcją związku w MNW (tw. Jensena):

$$g_*^{-1}\left(\mathbf{x}_{it}\boldsymbol{\beta}\right)=\mathrm{E}_{c_i}\left[g^{-1}\left(\mathbf{x}_{it}\,\boldsymbol{\beta}+c_i\right)\right]\neq g^{-1}\left[\mathbf{x}_{it}\,\boldsymbol{\beta}+\mathrm{E}_{c_i}\left(c_i\right)\right]=g^{-1}\left(\mathbf{x}_{it}\boldsymbol{\beta}\right)$$

- W szczególności za pomocą modelu wyestymowany GEE można uzyskać jedynie efekt cząstkowy uśredniony po cechach populacji (Averaged Partial Effect - APE)

$$APE=\mathrm{E}_c\left[\frac{\partial\,\mathrm{E}\left(y|\,\mathbf{x},c\right)}{\partial x}\right]=\frac{\partial g_*^{-1}\left(\mathbf{x}\boldsymbol{\beta}\right)}{\partial\mathbf{x}}$$

  - szacując APE zastępujemy $\boldsymbol{\beta}$ oszacowaniem $\widehat{\boldsymbol{\beta}}_{GEE}$

- W przypadku modelu estymowane MNW możliwe jest oszacowanie efektu cząstkowego dla danej jednostki i
$$\frac{\partial \mathrm{E}\left(y\middle|\mathbf{x},c\right)}{\partial \mathbf{x}}=\frac{g^{-1}\left(\mathbf{x}\boldsymbol{\beta}+c_i\right)}{\partial \mathbf{x}}$$
- Szacując efekt cząstkowy dla jednostki i, zastępujemy parametr $\boldsymbol{\beta}$ oszacowaniem $\widehat{\boldsymbol{\beta}}_{ML}$ a c_i oszacowaniem $\widehat{c}_i$.
- Oszacowanie c_i bazuje na rozkładzie a posteriori c_i:
$$f\left(c_i\middle|\mathbf{x}_i,\mathbf{y}_i\right)=\frac{f\left(\mathbf{y}_i\middle|\mathbf{x}_i,c_i\right)f\left(c_i\right)}{f\left(\mathbf{y}_i\middle|\mathbf{x}_i\right)}$$
- Wartość oczekiwana c_i z rozkładu a posteriori
$$\mathrm{E}\left(c_i\middle|\mathbf{x},\mathbf{y}_i\right)=\frac{1}{f\left(\mathbf{y}_i\middle|\mathbf{x}_i\right)}\int c_i f\left(\mathbf{y}_i\middle|\mathbf{x}_i,c\right)f\left(c_i\right)dc_i$$
- Oszacowanie c_i znajdujemy zastępując nieznane parametry rozkładów ich oszacowaniami i apoksymując całki kwadraturami.