Уопштени линеарни модели

У неким ситуацијама су зависне променљиве дискретног типа. Најједноставнији пример би била прменљива Y која за сваку вредност предиктора узима само две вредности: 0 или 1 (категоричка променљива са две вредности: ДА или НЕ, ИСТИНА или НЕИСТИНА, ЗА или ПРОТИВ). Дакле, условна расподела предиктора је

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

Регресиона функција је

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

Јасно је, да у овој ситуацији, линеарни модел не би био адекватан. Неки од разлога су следећи.

1. Грешке модела не могу се моделирати нормалном расподелом, или неком другом абсолутно непрекидном и сметричном око нуле.

2. Дисперзија грешака модела није константна. Важи: $D(Y_i|X_i) = \pi(X_i)(1-\pi(X_i)) = D(\varepsilon_i)$.

3. С обзиром да је регресиона функција вероватноћа треба да буде задовољено да је $\pi(X_i) \in [0,1]$. За линеарну функцију то очигледно не важи.

Пробит регерсија

Ради једноставности, за сада, претпоставимо да имамо само један предиктор. $\pi(X_i) = \Phi(\beta_0^* + \beta_1^* X_i)$

Одавде је $\Phi^{-1}(\pi(X_i)) = \beta_0^* + \beta_1^* X_i$ линеарни модел. Ова трансформација је позната под називом *пробит трансформација*.

Логистичка регресија

$\pi(X_i) = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$

Одавде је $F_L^{-1}(\pi(X_i)) = \log\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right) = \beta_0+\beta_1 X_i$. Ова трансформација је позната под називом *логит трансформација*.

Количник $\frac{\pi(X_i)}{1-\pi(X_i)}$ се назива *квота* (odds.)

Параметре модела оцењујемо методом максималне веродостојности. Логаритам функције веродостојности је

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

Максимум ове функције одређује се нумерички. Нека су добијене оцене $\hat{\beta}_0$ и $\hat{\beta}_1$ за непознате коефицијенте β_0 и β_1. Одавде је оцењена регресиона функција

$$\hat{\pi}(X_i) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_i}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X_i}}$$

Оцењена логит функција је

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

Тестирање значајности коефицијената се може тестирати тестом количника веродостојнисти. Како

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

под нултом хипотезом има приближно χ_1^2 расподелу, можемо извршити тестирање значајности коефицијента уз предиктора на уобичајан начин.

Још један од начина да проверимо утицај сваке независне променљиве на посматрану зависну, као значај сваког коефицијента, односно да се процени да ли ће се избацивањем неког коефицијента изгубити на квалитету модела , је Валдов тест. Користи се Валдова тест статистика

$$Z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}, \quad i = 0, 1.$$

Ова статистика, при важењу нулте хипотезе ($\beta_i = 0$) има нормалну расподелу па се могу направити одговарајуће критичне области за тестирање и израчунати p-вредности тестова. Стандардно одступање оцене се може добити на следећи начин.

Означимо са $\hat{\pi}_i = \hat{\pi}(X_i)$, као $\omega_i = \hat{\pi}_i(1-\hat{\pi}_i)$. Тада је тежинска средина

$$\bar{X}_\omega = \frac{\sum_{i=1}^{n} \omega_i X_i}{\sum_{i=1}^{n} \omega_i}.$$

тежинска сума квадратних одступања је

$$SS_\omega = \sum_{i=1}^{n} \omega_i (X_i - \bar{X}_\omega)^2.$$

Може се показати да су стандардна одступања оцена параметара (*Fleiss et al.*2003)

$$SE(\hat{\beta}_0) = \sqrt{\frac{1}{\sum_{i=1}^{n} \omega_i} + \frac{\bar{X}_\omega^2}{SS_\omega}}$$

$$SE(\hat{\beta}_1) = \frac{1}{\sqrt{SS_\omega}}.$$

Коваријанса оцена параметара је

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{\bar{X}_\omega}{SS_\omega}.$$

Тада је

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

као

$$SE(\lambda(\hat{\pi}(X))) = \sqrt{SE(\hat{\beta})^2 + 2XCov(\hat{\beta}_0, \hat{\beta}_1) + X^2 SE(\hat{\beta}_1)^2}$$

и

$$SE(\hat{\pi}(X)) = \hat{\pi}(X)(1 - \hat{\pi}(X))SE(\lambda(\hat{\pi}(X))).$$

Сада можемо направити и интервал поверења за $\pi(X)$.

До сад нисмо говорили о типу предиктора. Уколико је предиктор дискретна случајна величина онда можемо χ^2 тестом проверити значајност коефицијената.

Чланове узорка груписаћемо на основу вредности независне променљиве. Дакле, за свако X_i из узорка формирамо подскуп који чине они елементи узорка чија је независна компонента једнака одабраном X_i.
Нека је m_j број елемената у j-тој подгрупи посматраног узорка, $j = 1, 2...J$. У оквиру сваке подгрупе се може оценити условна вероватноћа $\pi(X_j) = P\{Y = 1|X_j\}$. Нека је n_j прој елемената у подгрупи за које је вредност зависне променљиве једнака 1. Оцена поменуте вероватноће, на основу логистичког модела је $\hat{P}\{Y = 1|X_j\} = \frac{1}{1+e^{\hat{\beta}_0 + \hat{\beta}_1 X_j}}$. Тада је

очекиван просечан број елемената из узорка чија је вредност зависне променљиве 1, једнака:

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

У зависности од расподељености зависне променљиве у оквиру сваке групе, као и међусобном односу група на основу те карактеристике, користе се различите статистике за проверу квалитета добијеног логистичког модела.

Пирсонови резидуали

Пирсонов j-ти резидуал је дефинисан са

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

Пирсонова статистика је дефинисана са

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

C има приближно χ^2_{J-2}.

Квалитет Пирсонових резидуала испољава се чињеницом да је њихова очекивана вредност 0, као и да је за сваки резидуал дисперзија иста.

Резидуали девијације

Резидуал девијације, за $n_j - \hat{n}_j \geq 0$, је дефинисан са:

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

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

За $n_j - \hat{n}_j < 0$ за j-ти резидуал се узима $-d_j$,
за $n_j = 0$

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

док је за $n_j = m_j$

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

Тест статистика је

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

D има приближно χ^2_{J-2} расподелу. Показује се да ови резидуали брже теже нормално распоређеној случајној променљивој, него Пирсонови резидуали.

Лог-Вејбулова регресија

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

Ова траснформација, због своје асиметричности, се најчешће користи за моделовање малих и великих вероватноћа успеха. Позната је под називом *трансформација итерираног логаритма* (complementary log-log regression).

Уопштени линеарни модели

Ово модели се састоје од следећих компоненти:

- линеарна комбинација коефицијената модела

$$\eta_j = X_j^T \beta \ \text{ односно } \ \eta_j = \beta_0 + \sum_{i=1}^{p} X_{j_i} \beta_i$$

- "линк" функције која представља трансфорамцију коју треба применити на функцију средње вредности зависне променљиве, да би се та трансформисана променљива могла описати линеарним моделом, односно за $\mu_j = EY_j$ и линк функцију g важи

$$g(\mu_j) = \eta_j$$

- дисперизија зависне променљиве се може представити у облику

$$D(Y_j) = CV(\mu_j).$$

Уколико је $g(x) = x$, $V(x) = 1$ и $C = \sigma^2$ добијамо класичан линеарни модел.
Уколико $Y_j \sim \mathcal{B}(1, \mu_j)$ расподелу и $g(x) = F_L^{-1}(x) = \log\left(\frac{x}{1-x}\right)$ и $V(x) = x(1-x)$ добијамо логистичку регресију.
Веома често се у пракси јавља случај када $Y_j \sim \mathcal{P}(\lambda_j)$. Тада је $\mu_j = \lambda_j$ и $D(Y_j) = \lambda_j = \mu_j$ па је $V(x) = x$. Треба још да одредимо линк функцију. Приметимо да за њу треба да важи да слика $(0, \infty)$ на $(-\infty, \infty)$. Зато је природан избор линк функција $g(x) = \log(x)$.

Нормална, биномна и Пуасонова расподела припадају експоненцијалној фамилији расподела.Експоненцијалној фамилији са распоршењем (расејањем) припадају све расподеле за које се функција густине (закон расподеле) може приказати у облику:

$$f(x, \theta) = e^{\frac{c(\theta)^T T(y) - d(\theta) + S(y)}{\phi(\tau)}}$$

Параметар τ се назива параметром *распршења*. Када је $\phi(\tau)$ познато ради се о класичној експоненцијалној фамилији расподела.

Примери: Бернулијева расподела, нормална расподела, Пуасонова расподела...(на часу).

Уколико је $T(y) = y$ и $c(\theta) = \theta$ кажемо да се ради о расподели у *канонском облику*. Тада је

$$EY = -d'(\theta) = \mu$$
$$DY = d''(\theta)\phi(\tau) = V(\mu)\phi(\tau).$$

Непознати параметри модела се одређују, као и у случају логистичке регресије, методом максималне веродостојности.

Вратимо се одабиру линк функције. Уколико је линк функција одабрана тако да је за канонски параметар $\theta = \eta$ онда такву функцију називамо *канонском* линк функцијом. Јасно је да је у случају логистичке регресије канонска функција баш *logit* функција.

Предност одабира канонске линк функције је што је тада $X^T Y$ довољна статистика за β јер је

$$L(y, \theta) = e^{\frac{\sum_{i=1}^n y_i x_i^T \beta - d(x_i^T \beta) - S(y_i)}{\phi(\tau)}}$$

0.0.1 Асимптотска својства МЛ оцене

- асимптотска нормалност (на часу)
- Валдова статистика
- Девијација

0.1 Пуасонова регресија

Канонска линк функција је $g(t) = \log(t)$. Тада је

$$l(\beta) = \sum_{i=1}^{n} y_i \log(e^{x_i^T \beta}) - e^{x_i^T \beta} + Const$$

У случају засићеног модела је

$$l(\mu) = -\sum_i y_i + \sum_i y_i \log(\mu_i) + Const$$

Одавде је оцена за μ_i баш y_i Одговарајућа девијација је

$$D(y, \hat{\mu}) = 2(l(\mu) - l(\hat{\mu})) = 2\sum_{i=1}^{n} y_i \log(\frac{y_i}{\hat{\mu}_i}) - 2(y_i - \hat{\mu}_i) = 2\sum_i y_i \log(\frac{y_i}{\hat{\mu}_i})$$

Статистика D има приближно $\chi^2_{n-\text{број оцењених параметара}}$ расподелу.