Zobecněný lineární model

Ekonometrie

Jiří Neubauer, Jaroslav Michálek

Katedra ekonometrie FVL UO Brno
kancelář 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

Zobecněný lineární model

Lineární regresní model patří k nejpoužívanějším metodám statistické analýzy vícerozměrných dat v ekonometrii. Nabízí možnost vyjádření vztahu mezi vysvětlovaná proměnnou (odezvou) a množinou vysvětlujících proměnných (regresorů) pomocí regresní funkce, která je lineární funkcí neznámých odhadovaných parametrů. V některých situacích ale předpoklad linearity není splněn a potom je potřeba přejít ke složitějším matematickým modelům a zabývat se modely, kde regresní funkce není lineární funkcí neznámých parametrů. V mnohých vybraných situacích se vystačí s regresní funkcí, která je sice nelineární funkcí vybraných parametrů, ale je funkcí lineární kombinace vysvětlujících proměnných, přičemž koeficienty této lineární kombinace jsou neznámé parametry. Takové modely se nazývají **zobecněné lineární modely**.

Zobecněný lineární model

Použití lineárního modelu je limitováno čtyřmi základními podmínkami (P1), (P2), (P3) a (P4)

(P1) Střední hodnota $E(\epsilon_i) = 0$, $i = 1, 2, \ldots, n$, tj. náhodné chyby jsou **nesystematické**.

(P2) Rozptyl $D(\epsilon_i) = \sigma^2$, $i = 1, 2, \ldots,$, tj. náhodné chyby jsou **homogenní** se stejným neznámým rozptylem σ^2.

(P3) Náhodné chyby ϵ_i jsou nezávislé.

V případě, kdy je třeba provádět testy hypotéz o neznámých parametrech a konstruovat intervaly spolehlivosti pro neznámé parametry modelu, zavádí se v LRM další předpoklad:

(P4) Náhodné chyby ϵ_i mají normální rozdělení.

Když v obecném lineárním modelu nahradíme tyto čtyři podmínky podmínkami obecnějšími, dospějeme k **zobecněnému lineárnímu modelu**.

Zobecněný lineární model

Pokud jde o podmínku (P1), zavedeme nejdříve funkci

$$\eta = \eta(X_1, X_2, \ldots, X_k) = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k. \tag{1}$$

Funkce η je lineární kombinací regresorů $X_1, X_2, \ldots, X_k$ a koeficienty této lineární kombinace jsou neznámé parametry $\beta_1, \beta_2, \ldots \beta_k$. Dále ji budeme ji nazývat **lineárním prediktorem**. Pro lineární regresní model lze vyjádřit střední hodnotu μ odezvy Y pomocí funkce η identickým vztahem

$$\mu = E(Y) = \eta = \eta(X_1, \ldots, X_k) = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k.$$

Tedy v lineárním regresním modelu predikujeme střední hodnotu μ náhodné veličiny Y pomocí vztahu $\mu = \eta$.

Zobecněný lineární model

Když označíme η_i hodnotu prediktoru η při hodnotách regresorů $X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_k = x_{ik}$, lze pak lineární regresní model přepsat do tvaru $\mu_i = \eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$. Tedy střední hodnota i-tého pozorování odezvy Y je podle podmínky (P1) přímo rovna hodnotě lineárního prediktoru η_i pro $X_1 = x_{i1}, \ldots, X_k = x_{ik}$.

Podmínka (P1) se ve zobecněném lineárním modelu nahrazuje novou podmínkou, která nahrazuje identický vztah mezi střední hodnotou $\mu = E(Y)$ a lineárním prediktorem η obecnějším vztahem. Předpokládá se, že μ a η jsou v obecném funkčním vztahu, který je určen tzv. **linkovací funkcí** g. Tedy podmínku (P1) z lineárního modelu lze přepsat jako novou podmínku zobecněného lineárního modelu tvaru:

(ZP1) $\eta = g(\mu)$,

přičemž o funkci g se předpokládá, že je ryze monotónní a existuje funkce h, která je inverzní funkcí k funkci g. Na základě podmínky (ZP1) lze střední hodnotu μ odezvy Y zapsat jako funkci lineárního prediktoru η ve tvaru $\mu = h(\eta)$. V zobecněném lineárním modelu uvažujeme novou modelovou rovnici

$$\mu_i = E(Y_i) = h(\eta_i) = h(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}), \ i = 1 \ldots, n. \tag{2}$$

V tomto modelu už $E(Y_i)$ obecně není lineární funkcí lineárního prediktoru η_i, ale jedná se o speciální případ nelineárního modelu.

Zobecněný lineární model

Příklad

Zavedení linkovací funkce lze dobře osvětlit na příkladu, kdy jednotlivá pozorování odezvy Y mají logarimicko-normální rozdělení. Pak transformovaná veličina $\ln Y$ má normální rozdělení a lze uvažovat model

$$\ln E(Y) = \ln \mu = \eta = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

nebo naopak $E(Y) = \mu = \exp(\eta)$. V této situaci odpovídá linkovací funkce g logaritmické funkci a její inverzní funkce h odpovídá exponenciální funkci.

Zobecněný lineární model

Dále podmínky (P2) a (P3) lze pomocí jednotkové matice $\boldsymbol{I}$ přepsat do maticového tvaru

$$\text{var}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}$$

a v zobecněném lineárním modelu pak takto maticově vyjádřené podmínky (P2) a (P3) nahrazujeme maticovou podmínkou:

(ZP2) $\text{var}(\boldsymbol{Y}) = a(\phi)\boldsymbol{W}$,

kde $\boldsymbol{W}$ je diagonální matice, její diagonální prvky mohou záviset na vektoru neznámých parametrů $\boldsymbol{\beta}$. Dále varianční matice $\text{var}(\boldsymbol{Y})$ může záviset na dalším parametru ϕ prostřednictvím funkce $a(\phi)$. Parametr ϕ v této souvislosti nazýváme **rušivým parametrem**, předpokládáme že rušivý parametr je nějakou konstantou, v testovaných hypotézách nevystupuje, ale pro popis modelu je potřebný. Srovnáním s podmínkou lineárního regresního modelu (P2) vidíme, že v lineárním modelu byl rušivým parametrem ϕ rozptyl σ^2, funkce a byla identická funkce, tedy $a(\sigma^2) = \sigma^2$ a matice $\boldsymbol{W}$ byla rovna jednotkové matici $\boldsymbol{I}$.

Zobecněný lineární model

Konečně se ve zobecněném lineárním modelu podmínka (P4) zobecňuje a předpokládá se místo ní podmínka:

(ZP3) Rozdělení odezvy Y patří do **exponenciální třídy rozdělení**,

přičemž exponenciální třída rozdělení je speciální skupina rozdělení, která zahrnuje celou řadu známých diskrétních i spojitých rozdělení. Patří do ní např. rozdělení binomické, Poissonovo, normální, exponenciální, gamma a další.

Exponenciální třída rozdělení

Předpokládejme dále, že je dán systém hustot $f(y;\lambda)$, kde y je proměnná a λ je neznámý parametr. Pro jednoduchost budeme předpokládat, že parametr λ je jednorozměrný reálný parametr. Dále budeme předpokládat, že daný systém hustot vyhovuje jistým podmínkám regularity, které zaručí korektnost dále prováděných matematických operací.

Budeme říkat, že rozdělení pravděpodobnosti má hustotou $f(y;\lambda)$ exponenciálního typu (stručněji, že rozdělení je **exponenciálního typu**), když existují funkce $r(\lambda)$ a $q(\lambda)$ parametru λ a funkce $s(y)$ a $t(y)$ reálné proměnné y tak, že jejich prostřednictvím lze hustotu $f(y;\lambda)$ vyjádřit ve tvaru

$$f(y;\lambda) = \exp\left\{t(y)q(\lambda) + r(\lambda) + s(y)\right\}. \tag{3}$$

Pozn. Je třeba upozornit na rozdíl mezi hustotou exponenciálního typu a hustotou exponenciálního rozdělení. Jde o dva zcela odlišné pojmy.

Exponenciální třída rozdělení

Příklad – *Poissonovo rozdělení Po*(λ)
Hustota Poissonova rozdělení (tj. jeho pravděpodobnostní funkce podle úmluvy uvedené výše) je tvaru

$$f(y;\lambda) = \mathrm{e}^{-\lambda}\frac{\lambda^y}{y!} \quad \text{pro} \quad y \in \{0,1,\dots\}, \quad \lambda > 0 \quad \text{je parametr.}$$

Uvedenou hustotu lze snadno převést na tvar

$$f(y;\lambda) = \exp\{y\ln\lambda - \lambda + \ln(y!)\}.$$

V uvedeném vztahu a rovněž v dalším textu funkce $\ln(x)$ značí přirozený logaritmus. Jestliže položíme $t(y) = y, q(\lambda) = \ln(\lambda), r(\lambda) = -\lambda$ a $s(y) = y!$, pak je ihned zřejmé, že hustota $f(y;\lambda)$ je tvaru (3) a je tedy exponenciálního typu.

Exponenciální třída rozdělení

Příklad – *Exponenciální rozdělení Ex*(λ)
Hustota exponenciálního rozdělení je tvaru

$$f(y;\lambda) = \lambda \mathrm{e}^{-\lambda y} \quad \text{pro} \quad y > 0, \quad \lambda > 0 \quad \text{je parametr.}$$

Snadno nahlédneme, že při volbě $t(y) = y, q(\lambda) = -\lambda, r(\lambda) = \ln(\lambda)$ a $s(y) = 0$ dostaneme

$$f(y;\lambda) = \mathrm{e}^{-\lambda y + \ln(\lambda)} = \mathrm{e}^{t(y)q(\lambda)+r(\lambda)+s(y)},$$

takže je zřejmé, že hustota exponenciálního rozdělení je exponenciálního typu.

Exponenciální třída rozdělení

V obou uvedených příkladech je $t(y) = y$. Tato skutečnost motivuje zavedení následující terminologie. Říkáme, že hustota exponenciálního typu je **v kanonickém tvaru**, když ve vztahu (3) platí, že $t(y) = y$. Dále lze v hustotě exponenciálního typu (3), která je v kanonickém tvaru, provést reparametrizaci a zavést nový parametr θ vztahem $\theta = q(\lambda)$. Tento nový parametr θ pak nazýváme **kanonickým parametrem**.

V případě Poissonova rozdělení $Po(\lambda)$ je kanonickým parametrem parametr $\theta = \ln(\lambda)$ a v případě exponenciálního rozdělení $Ex(\lambda)$ je kanonickým parametrem parametr $\theta = -\lambda$.

Exponenciální třída rozdělení

V některých situacích s hustotou exponenciálního typu tvaru (3) nevystačíme. Často se v praxi stává, že pravděpodobnostní rozdělení, s nimiž pracujeme, obsahují rušivý parametr ϕ. Ten sice není bezprostředně středem našeho zájmu, ale jak již bylo zmíněno následně po zavedení podmínky (ZP2) v definici zobecněného lineárního modelu, je třeba věnovat mu pozornost i přes to, že testované hypotézy na něm nezávisí.

Roli rušivého parametru lze demonstrovat na jednoduchém případě s normálním rozdělením $N(\mu, \sigma^2)$, kdy je třeba testovat hypotézu o jeho střední hodnotě μ při neznámém rozptylu σ^2. Pak tento rozptyl σ^2 vstupuje do rozhodovacího procesu, ale v nulové hypotéze, která se týká se pouze parametru μ, se neobjevuje. V popsané situaci je tedy rušivým parametrem ϕ parametr σ^2.

Exponenciální třída rozdělení

V dalších úvahách budeme i nadále rušivý parametr označovat písmenem ϕ a budeme uvažovat rozdělení s hustotou exponenciálního typu v kanonickém tvaru s parametrem θ, s rušivým parametrem ϕ a s hustotu $f(y;\theta,\phi)$ tvaru

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{4}$$

kde $b(\theta), a(\phi)$ a $c(y,\phi)$ jsou dané funkce svých argumentů. Snadno lze najít jejich vyjádření pomocí funkcí $t(y), q(\lambda), r(\lambda)$ a $s(y)$ použitých v definičním vztahu (3). Porovnáním (3) a (4) zjistíme, že v (3) je $a(\phi) = 1$ a dále platí $\theta = q(\lambda), b(\theta) = b(q(\lambda)) = = r(\lambda), c(y,\phi) = s(y)$. V tomto vztahu budeme parametr θ opět nazývat **kanonickým parametrem**.

Exponenciální třída rozdělení

Příklad – *Normální rozdělení* $N(\mu, \sigma^2)$
Hustota normálního rozdělení $N(\mu, \sigma^2)$ má tvar

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right\} = \exp\left\{-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\}.$$

Když v tomto posledním vztahu položíme $\phi = \sigma^2, a(\phi) = \phi, c(y, \phi) = -\frac{1}{2}(\frac{y^2}{\phi} + \ln(2\pi\phi))$ a $b(\mu) = \frac{\mu^2}{2}$, vidíme, že uvedená hustota $f(y; \mu, \sigma^2)$ patří do exponenciální třídy, je v kanonickém tvaru (4), $\theta = \mu$ je kanonický parametr a $\phi = \sigma^2$ je rušivý parametr.

Výpočet charakteristik pro rozdělení exponenciálního typu

Nejdříve zavedeme funkci l parametru λ vztahem $l(\lambda; y) = \ln(f(y; \lambda))$ a nazveme ji **logaritmickou věrohodnostní funkcí**. Dále zavedeme náhodnou veličinu

$$U(\lambda) = \frac{\partial l(\lambda; Y)}{\partial \lambda}$$

a nazveme ji **skórem**. Rozptyl skóru $D(U(\lambda))$ zřejmě závisí na parametru λ a nazývá se **Fisherovou mírou informace o parametru** λ, která je obsažena v rozdělení náhodné veličiny Y. Budeme ji značit $J(\lambda)$. Protože integrál z libovolné hustoty (nebo součet všech hodnot pravděpodobnostní funkce) je roven jedné, snadno nahlédneme, že platí

$$E(U(\lambda)) = 0$$

a pomocí tohoto vztahu odvodíme, že pro druhou derivaci logaritmické věrohodnostní funkce platí

$$-E\left(\frac{\partial^2 l(\lambda; Y)}{\partial \lambda^2}\right) = -E\left(\frac{\partial l(\lambda)}{\partial \lambda}\right)^2 = E(U^2(\lambda)) = D(U(\lambda)) = J(\lambda).$$

Fisherovu míru informace o parametru λ dostáváme ve tvaru

$$J(\lambda) = -E\left(\frac{\partial^2 l(\lambda; Y)}{\partial \lambda^2}\right). \tag{5}$$

Vztah (5) se někdy užívá pro definici Fisherovy míry informace o parametru λ.

Výpočet charakteristik pro rozdělení exponenciálního typu

Je-li hustota f exponenciálního typu tvaru (3), pak logaritmickou věrohodnostní funkci lze zapsat ve tvaru

$$l(\lambda; y) = t(y)q(\lambda) + r(\lambda) + s(y)$$

a pro její derivace (derivaci značíme čárkou u příslušné funkce) dostaneme

$$U(\lambda) = \frac{\partial l(\lambda; Y)}{\partial \lambda} = t(Y)q'(\lambda) + r'(\lambda)$$

a

$$U'(\lambda) = \frac{\partial^2 l(\lambda; Y)}{\partial \lambda^2} = t(Y)q''(\lambda) + r''(\lambda).$$

Odtud, protože $E(U(\lambda)) = 0$, lze vyjádřit střední hodnotu a rozptyl statistiky $t(y)$ ve tvaru

$$E(t(Y)) = -\frac{r'(\lambda)}{q'(\lambda)}, \tag{6}$$

a

$$D(t(Y)) = \frac{1}{[q'(\lambda)]^3}\left[q''(\lambda)r'(\lambda) - q'(\lambda)r''(\lambda)\right]. \tag{7}$$

Vztahy (6) a (7) dávají návod, jak snadno nalézt střední hodnotu a rozptyl rozdělení, která mají hustotu exponenciálního typu tvaru (3).

Výpočet charakteristik pro rozdělení exponenciálního typu

Je-li hustota $f(y;\theta)$ v kanonickém tvaru (4), lze srovnáním s hustotou (3) získat $\lambda = \theta,\ t(y) = y,\ q(\theta) = \frac{\theta}{a(\phi)},\ r(\theta) = -\frac{b(\theta)}{a(\phi)},\ s(y) = c(y,\phi)$ a ze vzorců (6) a (7) plyne, že

$$\mu = E(Y) = b'(\theta) \tag{8}$$

a

$$D(Y) = b''(\theta)a(\phi). \tag{9}$$

Ze vzorce (9) plyne, že rozptyl $D(Y)$ je součinem dvou funkcí. První činitel $b''(\theta)$ je funkcí kanonického parametru θ, a když existuje inverzní funkce b'_{-1} k funkci b', plyne ze (8), že $\theta = b'_{-1}(\mu)$. Když položíme $V(\mu) = b''(b_{-1'}(\mu))$, lze rozptyl ve (9) zapsat ve tvaru součinu $D(Y) = V(\mu)a(\phi)$, kde první činitel $V(\mu)$ závisí pouze na μ a druhý $a(\phi)$ závisí pouze na rušivém parametru ϕ. Dostaneme tedy, že pro rozdělení s hustotou exponenciálního typu v kanonickém tvaru (4) platí

$$E(Y) = \mu = b'(\theta) \quad \text{a} \quad D(Y) = V(\mu)a(\phi). \tag{10}$$

Z uvedeného vztahu je dobře patrné, že rozptyl uvažovaného rozdělení při dané hodnotě rušivého parametru závisí pouze na střední hodnotě a tato závislost je popsána funkcí $V(\mu)$. Proto funkci $V(\mu)$ budeme dále nazývat **variační funkcí**. Variační funkce má v teorii zobecněných lineárních modelů důležité místo.

Volba linkovací funkce

Budeme se zabývat otázkou, jak vhodně zvolit linkovací funkci g zavedenou v definici zobecněného lineárního modelu v podmínce (ZP1). Je-li hustota, s níž pracujeme, v kanonickém tvaru (4), můžeme jednoduše zavést tzv. *kanonickou linkovací funkci* g. Položme

$$\eta = \theta = \theta(\mu), \tag{11}$$

a odtud užitím podmínky (ZP1) dostaneme, že linkovací funkce g je dána vztahem

$$g(\mu) = \theta(\mu).$$

Srovnáním (8) s podmínkou (ZP1) vidíme, že pro tuto linkovací funkci platí

$$g(\mu) = b'_{-1}(\mu). \tag{12}$$

Funkci g zavedenou vztahem (11) pak nazýváme **kanonickou linkovací funkcí**.

Volba linkovací funkce

Při aplikacích zobecněných lineárních modelů se často pracuje s rozdělením normálním, binomickým, Poissonovým a gamma. Všechna tato rozdělení jsou exponenciálního typu a lze ji zapsat v kanonickém tvaru (4). V následujícím přehledu jsou pro tato rozdělení uvedeny funkce b, a a c, dále střední hodnota μ, kanonická linkovací funkce g a varianční funkce $V(\mu)$.

Normální rozdělení $N(\mu, \sigma^2)$

Hustota: $f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right\}$ *Parametry:* $\mu \in (-\infty, \infty), \quad \sigma > 0$

Obor hodnot: $(-\infty, \infty)$

Kanonický parametr: $\theta = \mu$ *Rušivý parametr:* $\phi = \sigma^2$

Funkce: $b(\theta) = \frac{\mu^2}{2}, \quad a(\phi) = \phi, \quad c(y,\phi) = -\frac{1}{2}(\frac{y^2}{\phi} + \ln(2\pi\phi))$

Střední hodnota: $\mu = \mu(\theta) = E(Y;\theta)) = \theta$

Kanonická linkovací funkce: $g(\mu) = \mu$

Variační funkce: $V(\mu) = 1$

Volba linkovací funkce

Rozdělení relativní četnosti tj. binomické rozdělení $Bi(m,\pi)/m$
Hustota: $f(y;m,\pi) = \binom{m}{my}\pi^{my}(1-\pi)^{m-my}$, *Parametry:* $m \in \{1,2,\dots\}$, $\pi \in (0,1)$
Obor hodnot: $\{0, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}, 1\}$
Kanonický parametr: $\theta = \ln\frac{\pi}{1-\pi}$ *Rušivý parametr:* $\phi = \frac{1}{m}$
Funkce: $b(\theta) = \ln(1+\mathrm{e}^{\theta})$, $a(\phi) = \phi$, $c(y,\phi) = \ln\binom{m}{my}$
Střední hodnota: $\mu = \mu(\theta) = E(Y;\theta)) = \frac{\mathrm{e}^{\theta}}{1+\mathrm{e}^{\theta}}$
Kanonická linkovací funkce: logitová: $g(\mu) = \ln(\frac{\mu}{1-\mu})$
Variační funkce: $V(\mu) = \mu(1-\mu)$

Poissonovo rozdělení $Po(\lambda)$
Hustota: $f(y;\lambda) = \mathrm{e}^{-\lambda}\frac{\lambda^y}{y!}$ *Parametr:* $\lambda > 0$
Obor hodnot: $\{0,1,2,\dots\}$
Kanonický parametr: $\theta = \ln(\lambda)$ *Rušivý parametr:* $\phi = 1$
Funkce: $b(\theta) = \mathrm{e}^{\theta}$, $a(\phi) = 1$ $c(y,\phi) = -\ln(y!)$
Střední hodnota: $\mu = \mu(\theta) = E(Y;\theta)) = \mathrm{e}^{\theta}$
Kanonická linkovací funkce: $g(\mu) = \ln(\mu)$
Variační funkce: $V(\mu) = \mu$

Volba linkovací funkce

Rozdělení gamma $G(\alpha, \nu)$

Hustota: $f(y; \alpha, \nu) = \frac{1}{\Gamma(\nu)} (\frac{\nu}{\alpha})^{\nu} \mathrm{e}^{-\frac{\nu}{\alpha} y} y^{\nu-1}$ *Parametry:* $\alpha > 0, \nu > 0$

Obor hodnot: $(0, \infty)$

Kanonický parametr: $\theta = -\frac{1}{\alpha}$ *Rušivý parametr:* $\phi = \nu^{-1}$

Funkce: $b(\theta) = -\ln(-\theta), \quad a(\phi) = \phi, \quad c(y, \phi) = \nu \ln(\nu y) - \ln(y) - \ln(\Gamma(\nu))$

Střední hodnota: $\mu = \mu(\theta) = E(y; \theta)) = -\frac{1}{\theta}$

Kanonická linkovací funkce: $g(\mu) = -\frac{1}{\mu}$

Variační funkce: $V(\mu) = \mu^2$

Volba linkovací funkce

V řadě experimentálních situací, zejména při výběrech malého rozsahu, se upřednostňuje kvalitní proložení modelové funkce daty před optimálními statistickými vlastnostmi modelu. V této situaci se potom využívají nejen kanonické linkovací funkce, ale i linkovací funkce jiného typu, které vedou k dobrým proložením.

a) **Probitová linkovací funkce**

$$\eta = \Phi_{-1}(\mu),$$

kde Φ_{-1} je inverzní funkce k distribuční funkci standardizovaného normálního rozdělení. Tato linkovací funkce se používá v probitové analýze.

b) **Komplementární log-log linkovací funkce**

$$\eta = \ln(-\ln(1-\mu)).$$

Komplementární log-log funkci lze získat jako inverzní funkci k distribuční funkci rozdělení extrémního typu, Gumbelova rozdělení.

c) **Mocninná linkovací funkce**

$$\eta = \mu^{\kappa} \quad \text{pro} \quad \kappa > 0 \quad \text{a} \quad \eta = \ln \mu \quad \text{pro} \quad \kappa \to 0$$

nebo

$$\eta = \frac{\mu^{\kappa} - 1}{\kappa} \quad \text{pro} \quad \kappa > 0 \quad \text{a} \quad \eta = \ln \mu \quad \text{pro} \quad \kappa \to 0.$$

V obou těchto transformacích je potřeba nejdříve provést odhad parametru κ.

Odhad parametrů zobecněného lineárního modelu metodou maximální věrohodnosti

Předpokládejme, že je dáno n nezávislých náhodných veličin $Y_1, \ldots, Y_n$, které se řídí zobecněným lineárním modelem s linkovací funkcí g a s hustotou exponenciálního typu v kanonickém tvaru (4). Hustota veličiny Y_i závisí na parametru θ_i $i = 1, \ldots, n$. Předpokládejme, že rušivý parametr ϕ je známý pro všechna pozorování $Y_1, \ldots, Y_n$. Pro střední hodnotu Y_i dostaneme

$$\mu_i = E(Y_i) = b'(\theta_i), \ i = 1, \ldots, n. \tag{13}$$

Pomocí linkovací funkce g lze lineární prediktor

$$\eta_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \tag{14}$$

vyjádřit jako funkci střední hodnoty μ_i ve tvaru

$$\eta_i = g(\mu_i), \ i = 1, \ldots, n. \tag{15}$$

Uvedené vztahy využijeme při odvozování věrohodnostních rovnic pro výpočet odhadů neznámých parametrů $\beta_1, \ldots, \beta_k$.

Odhad parametrů zobecněného lineárního modelu metodou maximální věrohodnosti

Označíme-li $l_i = \ln f(y_i; \theta_i, \phi)$ **logaritmickou věrohodnostní funkci** náhodné veličiny Y_i, dostaneme logaritmickou věrohodnostní funkci náhodného vektoru $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ ve tvaru

$$L(\boldsymbol{\beta}) = \ln \prod_{i=1}^{n} f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} \ln f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} l_i(\theta_i, \phi; y_i).$$

Označení $L(\boldsymbol{\beta})$ je použito proto, aby bylo zdůrazněno, že parametry θ_i závisí na parametrech $\beta_1, \ldots, \beta_k$, jak je patrno ze vztahů (13), (14) a (15). Maximálně věrohodné odhady neznámých parametrů $\beta_1, \ldots, \beta_k$ nalezneme maximalizací logaritmické věrohodnostní funkce $L(\boldsymbol{\beta})$. Vyjdeme z věrohodnostních rovnic

$$\frac{\partial L}{\partial \beta_j} = 0; \; j = 1, \ldots, k.$$

Nejdříve zavedeme **skórový vektor** $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{\beta})$ vzhledem k vektorovému parametru $\boldsymbol{\beta}$ vztahem

$$\boldsymbol{U}(\boldsymbol{\beta}) = \left(\frac{\partial L}{\partial \beta_1}, \ldots, \frac{\partial L}{\partial \beta_k} \right)'.$$

a věrohodnostní rovnice přepíšeme do maticového tvaru

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}. \tag{16}$$

Odhad parametrů zobecněného lineárního modelu metodou maximální věrohodnosti

Pro j-tou rovnici potom dostaneme

$$U_j(\beta) = \frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = 0,\ j = 1, \ldots, k. \tag{17}$$

Po výpočtu derivací uvedených v (17) a po jejich dosazení do (16) lze přepsat věrohodnostní rovnice (16) do tvaru

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{D(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} = 0,\ j = 1, \ldots, k. \tag{18}$$

Dále pomocí inverzní funkce $h = g_{-1}$ k linkovací funkci g lze získat vyjádření věrohodnostních rovnic (18) ve tvaru

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{D(Y_i)} h'(\eta_i) x_{ij} = \sum_{i=1}^{n} \frac{y_i - b'(\eta_i)}{D(Y_i)} h'(\eta_i) x_{ij} = 0,\ j = 1, \ldots, k. \tag{19}$$

Odhad parametrů zobecněného lineárního modelu metodou maximální věrohodnosti

Vzhledem k tomu, že linkovací funkce $\eta_i = g(\mu_i)$ je obecně nelineární funkcí a střední hodnoty μ_i i rozptyl $D(Y_i)$ závisí na parametrech $\beta_1, \ldots, \beta_k$ obecně nelineárně, jsou rovnice maximální věrohodnosti (19) obecně nelineární rovnice pro parametry $\beta_1, \ldots, \beta_k$. Snadno lze nahlédnout, že rovnici (19) lze zapsat v maticovém tvaru. Když označíme $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$, zavedeme diagonální matici $\boldsymbol{V} = \mathrm{diag}\{D(Y_1), \ldots, D(Y_n)\}$ a položíme

$$\boldsymbol{F} = \left(\frac{\partial \mu_i}{\partial \beta_j}\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,k}} = \left(\frac{\partial \mu_i}{\partial \eta_i} x_{ij}\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,k}} = \left(x_{ij} h'(\eta_i)\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,k}} = \boldsymbol{D}_h \boldsymbol{X},$$

kde

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk}, \end{pmatrix}$$

a $\boldsymbol{D}_h = \mathrm{diag}(h'(\eta_1), \ldots, h'(\eta_n))$.

Odhad parametrů zobecněného lineárního modelu metodou maximální věrohodnosti

Pak věrohodnostní rovnice (19) mají tvar

$$\boldsymbol{F}'\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) = \boldsymbol{X}'\boldsymbol{D}_h\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) = \boldsymbol{0}. \tag{20}$$

Při kanonické volbě linkovací funkce platí, že $\boldsymbol{D}_h = \frac{1}{a(\phi)}\boldsymbol{V}$ a věrohodnostní rovnice (19) se redukují na jednoduchý tvar

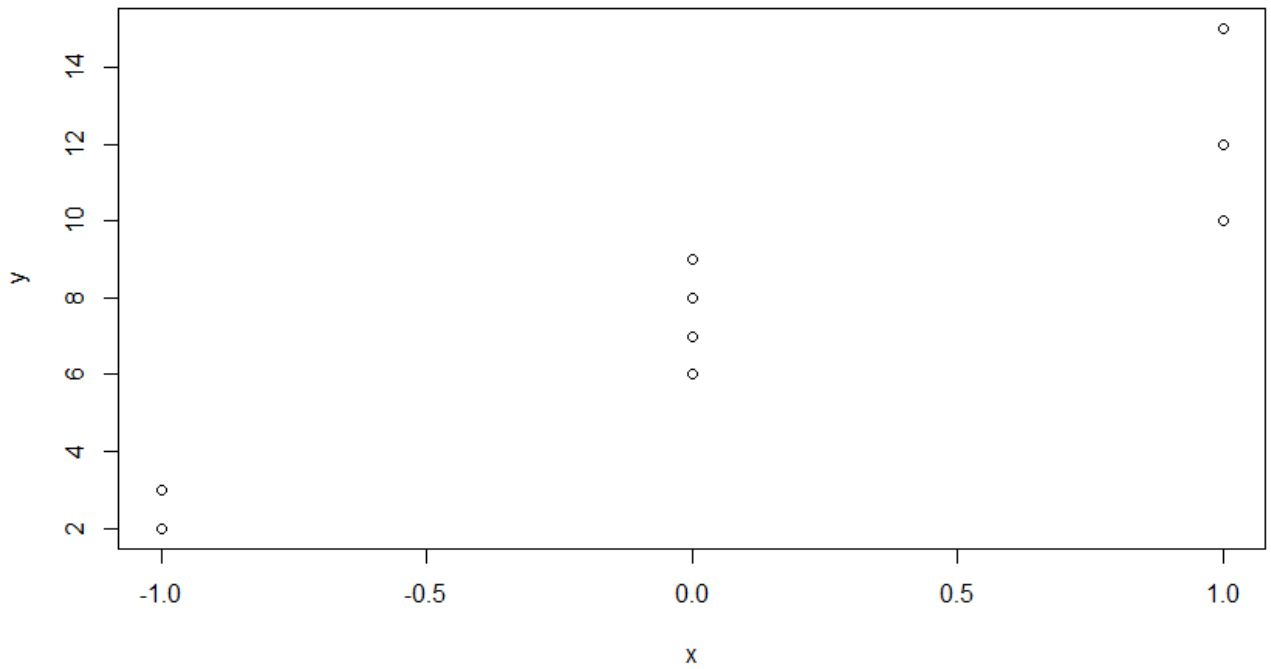
$$\boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{\mu}) = 0.$$

Jejich řešení se provádí iteračními technikami.

Odhad parametrů zobecněného lineárního modelu metodou maximální věrohodnosti – příklad

Uvažujme data z tabulky s odezvou Y a jedním regresorem x.

x	−1	−1	0	0	0	0	1	1	1
y	2	3	6	7	8	9	10	12	15



Obrázek: Poissonovská regrese

Odhad parametrů zobecněného lineárního modelu metodou maximální věrohodnosti – příklad

S rostoucí střední hodnotou odezvy roste její variabilita (viz obrázek), a proto budeme data modelovat pomocí poissonovské regrese. Připomeňme, že rozptyl a střední hodnota Poissonova rozdělení jsou stejné a rovny jeho parametru λ . Vzhledem k tomu, že obrázku je patrná lineární vazba odezvy Y na regresoru x, vyjdeme z modelu

$$E(Y_i) = \mu_i = \lambda_i = \beta_1 + \beta_2 x_i.$$

Cílem je odhadnout neznámé parametry β_1 a β_2. Využijeme k tomu rovnice (20). Zřejmě je $k = 2$ dále $n = 9$ a matice

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{91} & x_{92}, \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_9, \end{pmatrix}.$$

Odhad parametrů zobecněného lineárního modelu metodou maximální věrohodnosti – příklad

Dále matice

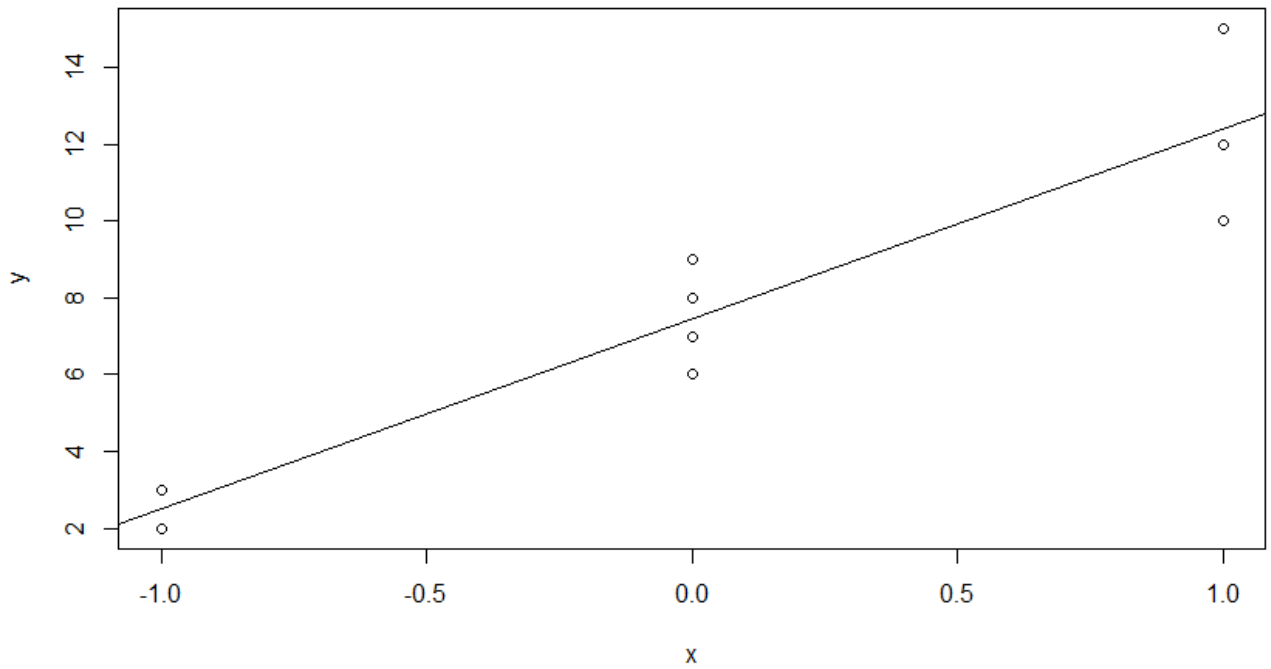
$$\boldsymbol{V} = \text{diag}\{D(Y_1), \ldots, D(Y_9)\} = \text{diag}\{\beta_1 + \beta_2 x_1, \ldots, \beta_1 + \beta_2 x_9\}$$

a

$$\boldsymbol{D}_h = \text{diag}(h'(\eta_1), \ldots, h'(\eta_9)) = \boldsymbol{I},$$

tedy $\boldsymbol{D}_h$ je jednotková matice, protože linkovací funkce je identita, $h(x) = x$ a $h'(x) = 1$. Můžeme sestavit rovnice (20), to je systém nelineární rovnic pro neznámé parametry β_1, β_2 a pro jeho řešení je nutné použít nějakou metodu numerické matematiky.

Odhad parametrů zobecněného lineárního modelu metodou maximální věrohodnosti – příklad



Obrázek: Poissonovská regrese – odhady parametrů získané metodou „Fisher Scoring“ po 3 iteracích jsou $\hat{\beta}_1 = 7{,}4516$ a $\hat{\beta}_2 = 4{,}9353$

Odhad parametrů zobecněného lineárního modelu metodou maximální věrohodnosti

Ukážeme speciální případ věrohodnostních rovnic pro situaci, kdy odezvy Y_i mají normální rozdělení $N(\mu_i, \sigma^2)$. Použijeme-li kanonickou linkovací funkci $\mu_i = \eta_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$, pak $\frac{\partial \mu_i}{\partial \eta_i} = 1$ a dosazením do věrohodnostních rovnic (18) ihned dostaneme

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\sigma^2} x_{ij} = 0, \; j = 1, \ldots, k.$$

Vynásobíme-li tuto rovnici rušivým parametrem σ^2, abychom jej eliminovali, dostaneme rovnici

$$\sum_{i=1}^{n} \left(y_i - \sum_{s=1}^{k} \beta_s x_{is} \right) x_{ij} = 0, \; j = 1, \ldots, k.$$

Její maticový tvar je $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$. Je tedy zřejmé, že systém věrohodnostních rovnic (16) v tomto speciálním případě přechází v systém normálních rovnic, které se používají v lineárním regresním modelu k odhadu parametru $\boldsymbol{\beta}$ metodou nejmenších čtverců.
Na uvedeném příkladu je také názorně vidět, že věrohodnostní rovnice (18) lze považovat za zobecnění normálních rovnic při přechodu od klasického lineárního regresního modelu ke zobecněnému lineárnímu modelu.

Statistická inference v zobecněných lineárních modelech

Statistická inference o parametrech $\beta_1, \ldots, \beta_k$, z níž se při analýze zpracovávaného datového souboru vychází, je založena na vlastnostech odhadů získaných metodou maximální věrohodnosti.
ro statistickou analýzu je důležitý výsledek, že v případě, kdy rovnice věrohodnosti mají řešení $\widehat{\boldsymbol{\beta}}$, které je konzistentním odhadem parametru $\boldsymbol{\beta}$, má náhodný vektor

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

asymptoticky normální rozdělení $N_k(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\beta})^{-1})$, kde

$$\boldsymbol{J}(\boldsymbol{\beta}) = -E\left(\frac{\partial^2 l(\boldsymbol{\beta}; Y)}{\partial \beta_i \partial \beta_j}\right)_{\substack{i=1,\ldots,k \\ j=1,\ldots,k}} . \tag{21}$$

je **Fisherova informační matice** (zobecnění Fisherovy informační míry na vektorový parametr) a používá se k výpočtu asymptotické varianční matice odhadovaných parametrů.

Statistická inference v zobecněných lineárních modelech

V praktických situacích se tedy vychází z předpokladů asymptotické normality odhadů $\widehat{\boldsymbol{\beta}}$ a neznámý parametr $\boldsymbol{\beta}$ se ve varianční matici $\boldsymbol{J}(\boldsymbol{\beta})^{-1}$ nahrazuje jeho maximálně věrohodným odhadem. Z uvedených výsledků pak plyne, že statistika

$$W = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \boldsymbol{J}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

má asymptoticky χ^2 rozdělení o k stupních volnosti. Statistika W se nazývá Waldova statistika. Lze ji použít k testování nulové hypotézy $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ alternativou je, že nulová hypotéza neplatí, $\boldsymbol{\beta}_0$ je daný vektor.

Pomocí Taylorovy aproximace věrohodnostní funkce $L(\boldsymbol{\beta}; \boldsymbol{Y})$ v bodě $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ lze ukázat, že statistika

$$D^* = 2(L(\widehat{\boldsymbol{\beta}}) - L(\boldsymbol{\beta}))$$

je asymptoticky ekvivalentní se statistikou W a má tedy také asymptoticky χ^2 rozdělení o k stupních volnosti, když $\boldsymbol{\beta}$ je skutečná hodnota tohoto parametru. Tato statistika se nazývá **deviance**, používá pro testování adekvátnosti modelu, lze ji stejně jako Waldovu statistiku použít pro testování nulové hypotézy $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, respektive pro testování vhodnosti redukovaného modelu.

Statistická inference v zobecněných lineárních modelech

Pro testování nulové hypotézy, že zvolený model dobře vysvětluje data, se používá srovnání zvoleného modelu s modelem maximálním, který se též nazývá **saturovaným modelem**. Uvažujme daný, pevně zvolený zobecněný lineární model s pevně danou linkovací funkcí g a s pevně daným rozdělením odezvy, které je exponenciálního typu. Pak saturovaným modelem příslušným k danému uvažovanému modelu je zobecněný lineární model, který má stejnou linkovací funkci a stejné rozdělení odezvy jako uvažovaný model a vektor jeho parametrů $\boldsymbol{\beta}_{\max}$ má n složek. Pro saturovaný model zřejmě platí, že odhad střední hodnoty $\widehat{\mu}_i$ je roven Y_i tedy $\widehat{\mu}_i = Y_i,\ i = 1, \ldots, n$, a to znamená, že saturovaný model úplně vysvětluje data. Statistika

$$D^* = 2(L(\widehat{\boldsymbol{\beta}}_{\max}) - L(\boldsymbol{\beta}_{\max}))$$

má asymptoticky χ^2 rozdělení o n stupních volnosti.

Jestliže platí nulová hypotéza, že model s k-rozměrným parametrem $\boldsymbol{\beta}$ vysvětluje data stejně dobře jako model saturovaný, platí $L(\boldsymbol{\beta}_{\max}) - L(\boldsymbol{\beta}) \doteq 0$. Označíme $\widehat{\boldsymbol{\beta}}_{\max}$ odhad $\boldsymbol{\beta}_{\max}$ v saturovaném modelu a podobně $\widehat{\boldsymbol{\beta}}$ odhad $\boldsymbol{\beta}$ v uvažovaném modelu. Můžeme zavést statistiku

$$D = 2[L(\widehat{\boldsymbol{\beta}}_{\max}) - L(\boldsymbol{\beta}_{\max})) - (L(\widehat{\boldsymbol{\beta}}) - L(\boldsymbol{\beta})) + (L(\boldsymbol{\beta}_{\max}) - L(\boldsymbol{\beta})].$$

Statistická inference v zobecněných lineárních modelech

Protože první člen na pravé straně uvedeného výrazu má asymptotické rozdělení $\chi^2(n)$, druhý má asymptotické rozdělení $\chi^2(k)$ a třetí je za platnosti nulové hypotézy přibližně roven nule, lze ukázat, že za platnosti nulové hypotézy má statistika

$$D = 2(L(\widehat{\boldsymbol{\beta}}_{\text{max}}) - L(\widehat{\boldsymbol{\beta}}))$$

asymptoticky rozdělení $\chi^2(n-k)$. Statistika D je vhodnou testovací statistikou pro ověření nulové hypotézy, že uvažovaný model popisuje data stejně dobře jako saturovaný model a také se nazývá deviancí. V některých situacích (např. ve výstupní sestavě programového systému R) se pro devianci D používá název **reziduální deviance**

Statistická inference v zobecněných lineárních modelech

Konečně v řadě statistických analýz bývá často potřeba testovat hypotézu, že daný model s k parametry lze redukovat na submodel s menším počtem q parametrů. Označíme-li v této situaci D_k devianci pro model s k parametry a D_q devianci pro model s q parametry, lze ukázat, že rozdíl

$$\Delta D = D_q - D_k$$

má asymptoticky rozdělení χ^2 s $k - q$ stupni volnosti. Odtud plyne, že nulovou hypotézu, že model s k parametry lze redukovat na model s q parametry, zamítneme na hladině významnosti α, když $\Delta D > \chi^2_{1-\alpha}(k - q)$. V situaci, kdy redukovaný model obsahuje pouze jeden parametr, tedy $q = 1$ (např. když lineární prediktor $\eta = \beta_1$), nazývá se deviance D_q ve výstupních programech systému R **nulovou deviancí**.

Statistická inference v zobecněných lineárních modelech

V případech, kdy je potřeba posoudit shodu modelu s daty a eliminovat vliv počtu parametrů uvažovaných modelů, lze pro srovnání využít statistiku AIC (z anglického Akaike information criterion), která je založena na logaritmické věrohodnostní funkci a je definována vztahem

$$\text{AIC} = -2L(\widehat{\boldsymbol{\beta}}) + 2k, \tag{22}$$

kde k je počet odhadovaných parametrů modelu. Statistika AIC je součástí výstupní sestavy programů ve výpočetním prostředí R v modulech, které umožňují provádět vyhodnocení zobecněných lineárních modelů.

Ještě poznamenejme, že jiná míra pro testování shody modelu s daty je Pearsonova statistika

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)},$$

která má asymptoticky rozdělení $\chi^2(n-k)$.

Statistická inference v zobecněných lineárních modelech

Uvedené výsledky vycházely z předpokladu, že rušivý parametr je známý. Pokud bychom předpokládali, že rušivý parametr není známý, je možné jej odhadnout metodou maximální věrohodnosti a příslušné testy modifikovat. K jednoduchému odhadu funkce rušivého parametru $a(\phi)$ dospějeme využitím Pearsonovy statistiky, která má v situaci, kdy uvažujeme rušivý parametr, asymptoticky rozdělení $a(\phi)\chi^2(n-k)$. Tedy $\chi^2/a(\phi)$ má rozdělení χ^2. Protože $E(\chi^2) \cong n-k$, lze statistiky χ^2 použít k odhadu rušivého parametru. Výsledkem je odhad $a(\phi)$ tvaru

$$\widehat{a}(\phi) = \frac{\chi^2}{n-k} = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_i)^2/V(\widehat{\mu}_i).$$

Statistická inference v zobecněných lineárních modelech

V případě, kdy je potřeba testovat hypotézu, že daný model s k parametry lze redukovat na submodel s menším počtem q parametrů lze ukázat, že rozdíl

$$\Delta D = D_q - D_k$$

má asymptoticky rozdělení $a\chi^2$ s $k - q$ stupni volnosti. Odtud plyne, že nulovou hypotézu, že model s k parametry lze redukovat na model s q parametry, zamítneme na hladině významnosti α, když $\frac{1}{a}\Delta D > \chi^2_{1-\alpha}(k - q)$. V některých situacích je jednodušší využít pro test této nulové hypotézy statistiku

$$F = \frac{D_q - D_k}{D_k} \cdot \frac{n - k}{k - q},$$

která má asymptoticky Fisher-Snedecorovo F rozdělení o $k - q$ a $n - k$ stupních volnosti. Tím se eliminuje vliv rušivého parametru.

Statistické inference pro binomický model

Budeme předpokládat, že $Y_1, \ldots, Y_n$ jsou nezávislé náhodné veličiny, $y_1, \ldots, y_n$ jejich realizace a dále předpokládáme, že Y_i má binomické rozdělení $Bi(n_i, \pi_i)$, přičemž

$$\mu_i = E(Y_i) = n_i \pi_i = n_i h(\eta_i) = n_i h(\beta_1 x_{i1} + \cdots + \beta_k x_{ik}), i = 1, \ldots, n, \qquad (23)$$

kde h odpovídá volbě linkovací funkce g. Pak logaritmická věrohodnostní funkce je tvaru

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \ln \pi_i - y_i \ln(1 - \pi_i) + n_i \ln(1 - \pi_i) + \ln \binom{n_i}{y_i} \right].$$

V saturovaném modelu je y_i odhadem μ_i a tedy odhadem π_i je v saturovaném modelu relativní četnost $\frac{y_i}{n_i}$. Dále v uvažovaném modelu (23) lze parametry π_i odhadnout metodou maximální věrohodnosti (řešením rovnice (19) při vhodně zvolené linkovací funkci). Když položíme $\widehat{y}_i = n_i \widehat{\pi}_i$, kde $\widehat{\pi}_i = h(\widehat{\eta}_i) = n_i h(\widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik})$ jsou maximálně věrohodné odhady parametrů π_i v modelu (23), můžeme pomocí logaritmické věrohodnostní funkce $L(\boldsymbol{\beta})$ zapsat devianci D^* ve tvaru

$$D = 2[L(\widehat{\boldsymbol{\beta}}_{\max}) - L(\widehat{\boldsymbol{\beta}})] = 2 \sum_{i=1}^{n} \left[y_i \ln \left(\frac{y_i}{\widehat{y}_i} \right) + (n_i - y_i) \ln \left(\frac{n_i - y_i}{n_i - \widehat{y}_i} \right) \right].$$

Statistické inference pro binomický model

Pro kanonickou linkovací funkci dostaneme **logistický regresní model**

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

nebo ekvivalentně

$$\pi_i = \frac{\exp\{\beta_1 x_{i1} + \cdots + \beta_k x_{ik}\}}{1 + \exp\{\beta_1 x_{i1} + \cdots + \beta_k x_{ik}\}}.$$

Dále pro probitovou linkovací funkci dostaneme **probitový model**

$$\Phi_{-1}(\pi_i) = \eta_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

nebo ekvivalentně

$$\pi_i = \Phi(\beta_1 x_{i1} + \cdots + \beta_k x_{ik}),$$

kde Φ je distribuční funkce standardizovaného normálního rozdělení $N(0,1)$. Použijeme log-log linkovací funkci, dostaneme **model extrémního typu**

$$\pi_i = 1 - \exp[-\exp(\beta_1 x_{i1} + \cdots + \beta_k x_{ik})].$$

Statistické inference pro binomický model – příklad

V rámci marketingového průzkumu byl zjišťován počet zájemců o koupi nového vozu. Průzkum byl proveden u 481 domácností, v každé byl zjišťován měsíční příjem domácnosti na jednoho člena domácnosti, značíme jej X a dále byla zjišťována odpověď na otázku, zda mají členové domácnosti v tříletém horizontu zájem o koupi nového vozu. Cílem bylo modelovat počet zájemců (domácností) o koupi nového vozu v závislosti na příjmu X. Příslušná data jsou uvedena v tabulce. V prvním sloupci tabulky je uveden zaokrouhlený příjem na tisíce Kč. Dále ve druhém sloupci tabulky je uveden počet domácností s daným zaokrouhleným příjmem a ve třetím sloupci tabulky je uveden počet domácností, které projevily o koupi nového vozu ve tříletém horizontu zájem.

Příjem x_i *[v tisících Kč]*	*Počet domácností* n_i	*Počet zájemců* y_i
16	59	6
17	60	13
18	62	18
19	56	28
20	63	52
21	59	53
22	62	61
23	60	60

Tabulka: Zaokrouhlený příjem na jednoho člena domácnosti x_i, celkové počty domácností n_i odpovídající příjmové skupině x_i a odpovídající počty domácností se zájmem o koupi nového vozu

Statistické inference pro binomický model – příklad

Počet zájemců byl modelován pomocí zobecněného lineárního modelu s logitovou, probitovou a komplementární log-log linkovací funkcí v závislosti na zaokrouhleném měsíčním příjmu, který připadá na jednoho člena domácnosti (proměnná x). Pro predikci pravděpodobnosti $\pi(x)$, že při daném příjmu x bude mít domácnost o koupi nového vozu zájem byl použit lineární prediktor $\eta = \beta_1 + \beta_2 x$. Logistický model byl tvaru

$$\pi(x) = \frac{\exp\{\beta_1 + \beta_2 x\}}{1 + \exp\{\beta_1 + \beta_2 x\}}.$$

Probitový model byl tvaru

$$\pi(x) = \Phi(\beta_1 + \beta_2 x),$$

a model extrémního typu byl tvaru

$$\pi_i = 1 - \exp[-\exp(\beta_1 x_{i1} + \beta_2 x_{i2})].$$

Statistické inference pro binomický model – příklad

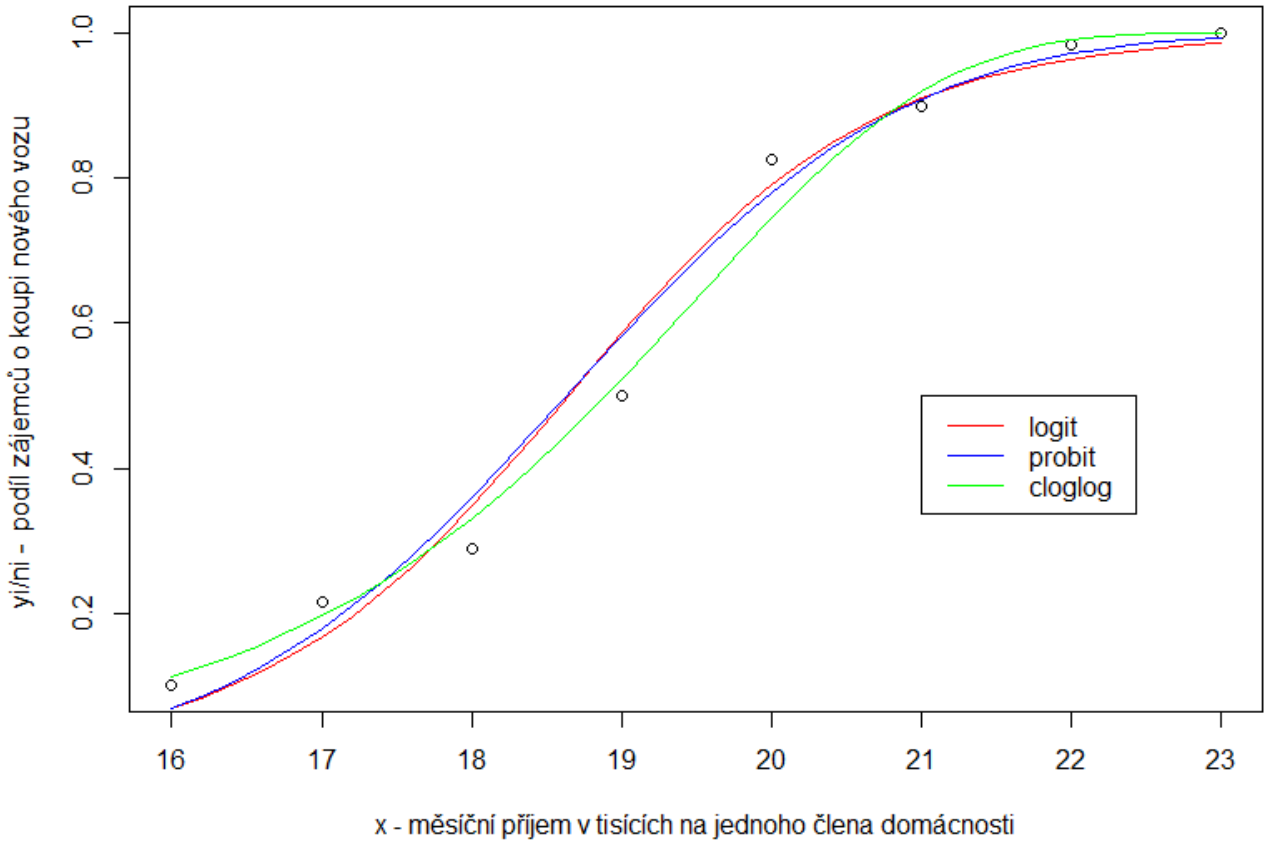
Odhadovaná veličina	*Logit model Odhad*	*Probit model Odhad*	*Model extrémního typu Odhad*
$\widehat{\beta}_1$	$-18{,}23^{***}$	$-10{,}49^{***}$	$-11{,}89^{***}$
Sm.odchylka $\widehat{\beta}_1$	1,582	0,816	1,002
$\widehat{\beta}_2$	$0{,}98^{***}$	$0{,}56^{***}$	$0{,}61^{***}$
Sm. odchylka $\widehat{\beta}_2$	0,084	0,043	0,051
D(rezid. deviance)	7,56 (6 st. v.)	6,48 (6 st. v.)	3,72 (6 st. v.)
Nulová deviance	284,20 (7 st. v.)	284,20 (7 st. v.)	284,20 (7 st. v.)
AIC	37,76	36,68	33,924

Tabulka: Výsledky statistických vyhodnocení. Odhady parametrů, jejich směrodatné chyby, nulová a reziduální deviance, *AIC* kritérium. *** u hodnoty parametru značí jeho statistickou významnost na hladině významnosti nižší než 0,001.

Statistické inference pro binomický model – příklad

V daném příkladě je počet hodnot zaokrouhlených příjmů $n = 8$ a počet odhadovaných parametrů je $k = 2$. V případě, že uvedený model dobře postihuje statistickou vazbu pravděpodobnosti $\pi(x)$ na příjmu x, má deviance D přibližně rozdělení $\chi^2(6)$ (stupně volnosti jsou uvedeny v závorce u příslušné hodnoty vypočtené statistiky). Protože 95% kvantil rozdělení $\chi^2(6)$ je 12,59, žádný z uvedených modelů nelze zamítnout. Nejmenší reziduální devianci vykazuje model extrémního typu. Rovněž oba parametry β_1 a β_2 jsou v každém uvažovaném modelu statisticky vysoce významné, *** u jejich hodnoty značí statistickou významnost na hladině významnosti nižší než 0,001.

Statistické inference pro binomický model – příklad



Obrázek: Odhady pravděpodobnosti $\pi(x)$, že domácnost uvažuje o koupi nového vozu v závislosti na příjmu x získané pro logistický model (červeně), probitový model (modře) a model extrémního typu (zeleně).

Použité zdroje

- AGRESTI, A., 2002. *Categorical Data Analysis.* John Wiley & Sons.
- ANDĚL, J., 1978. *Matematická statistika.* Praha: SNTL.
- ANDĚL, J., 2003. *Statistické metody.* Praha: Matfyzpress.
- ANDĚL, J., 2005. *Základy matematické statistiky*. 1. vyd. Praha: Matfyzpress.
- DOBSON, A., 2008. *An Introduction to Generalized Linear Models.* London: Chapman & Hall.