M5VM05 Statistické modelování

10. Konkrétní GLM modely – I.

Jan Koláček (kolacek@math.muni.cz)

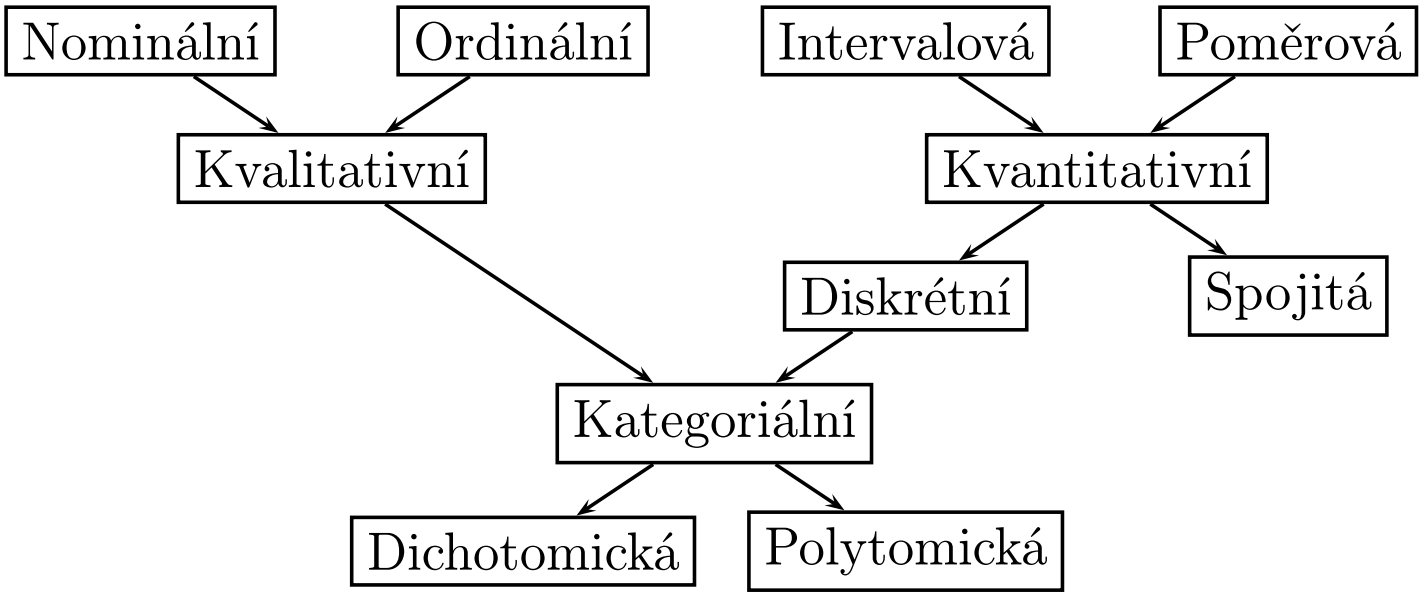
Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

Motivace

Na minulé přednášce jsme si uvedli obecnou definici zobecněného lineárního modelu a obecné konstrukce testů hypotéz o parametrech těchto modelů. Na této přednášce se již budeme zabývat zobecněnými lineárními modely pro konkrétní případy podle toho, jaké rozdělení má závisle proměnná Y.

Motivace

Typy veličin:



Modely pro alternativní a binomická data

Předpokládejme, že $U_i \sim A(\pi_i)$ $(i = 1, \ldots, N)$ nabývá pouze dvou hodnot 0 a 1,

$$P(U_i = u) = \begin{cases} \pi_i & u = 1 \\ 1 - \pi_i & u = 0 \\ 0 & \text{jinak} \end{cases} = \begin{cases} \pi_i^u (1 - \pi_i)^{1-u} & u = 0, 1 \\ 0 & \text{jinak.} \end{cases}$$

Předpokládejme, že náhodná veličina U_i závisí na k veličinách $x_{i1}, \ldots, x_{ik}$, tzv. **kovariáty**.
Data můžeme mít zadána různým způsobem:

- **jednotlivá pozorování** U_i:

hodnoty kovariát	pozorované binární veličiny
$x_{i1}, \ldots, x_{ik}$	U_i

Modely pro alternativní a binomická data

- **skupinově**, kdy známe **absolutní** četnosti úspěchů Y_j a celkový počet pokusů n_j, tedy máme k dispozici binomická data $Y_j \sim Bi(n_j, \pi_j)$

$$P(Y_j = y) = \begin{cases} \binom{n_j}{y} \pi_j^y (1-\pi_j)^{n_j - y} & y = 0, 1, \ldots, n_j \\ 0 & \text{jinak} \end{cases}$$

kde

$$j = 1, \ldots, n; \qquad N = n_1 + \cdots + n_n$$

a data můžeme zapsat formou tabulky

hodnota kovariát	počet úspěchů	počet pokusů
$x_{j1}, \ldots, x_{jk}$	Y_j	n_j

Modely pro alternativní a binomická data

- **skupinově**, kdy známe **relativní** četnost úspěchů $Z_j = \frac{Y_j}{n_j}$ a celkový počet pokusů n_j

$$P(Z_j = y) = \begin{cases} \binom{n_j}{n_j y} \pi_j^{n_j y} (1-\pi_j)^{n_j - n_j y} & y = 0, \frac{1}{n_j}, \ldots, 1 \\ 0 & \text{jinak} \end{cases}$$

kde

$$j = 1, \ldots, n; \qquad N = n_1 + \cdots + n_n$$

Data lze zapsat do tabulky

kovariáty	relativní úspěšnost	počet pokusů
$x_{j1}, \ldots, x_{jk}$	$Z_j = \frac{Y_j}{n_j}$	n_j

Cíl

Hlavním úkolem statistické analýzy je nalézt vztah mezi Z_i, (tj. i Y_i) a $x_{i1}, \ldots, x_{ik}$, tj. funkci

$$\pi_i = \pi(\mathbf{x}_i) = \pi(x_{i1}, \ldots, x_{ik}).$$

Protože chceme použít GLM modely, modelujeme pravděpodobnosti π_i pomocí linkovacích funkcí

$$g(\pi_i) = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}.$$

Nejjednodušším modelem je **lineární model**

$$\pi_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}.$$

Avšak tento model má řadu nevýhod, především je třeba zajistit, aby $\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}$ nabývala hodnot mezi 0 a 1, tedy je třeba přidat nějaké dodatečné podmínky. Proto, abychom tuto podmínku dodrželi, využijeme nějakou **distribuční funkci**

$$F(t) = \int_{-\infty}^{t} f(s)ds \qquad f(s) \geq 0 \qquad \int_{-\infty}^{\infty} f(s)ds = 1$$

s odpovídající **hustotou** $f(s)$, která se v tomto případě nazývá **toleranční funkce** (toleranční distribuce).

Modely dávka – odpověď

Typickým příkladem těchto modelů je vztah mezi dávkou toxické látky a odezvy (kladná-přežití, záporná-smrt) jedince na tuto dávku. Odezvy bývají obvykle udávány jako procenta kladné odezvy (quantal responses).

Symetrické modely

Jestliže uvažujeme toleranční distribuci jako **rovnoměrně spojitou** na nějakém intervalu (a,b), tj

$$f_0(s) \sim Rs(a,b) \qquad f(s) = \begin{cases} \frac{1}{b-a} & s \in (a,b) \\ 0 & \text{jinak} \end{cases}$$

pak

$$\pi_0(x) = F_0(x) = \int_a^x f(s)ds = \frac{x-a}{b-a} \quad \text{pro} \quad x \in (a,b)$$

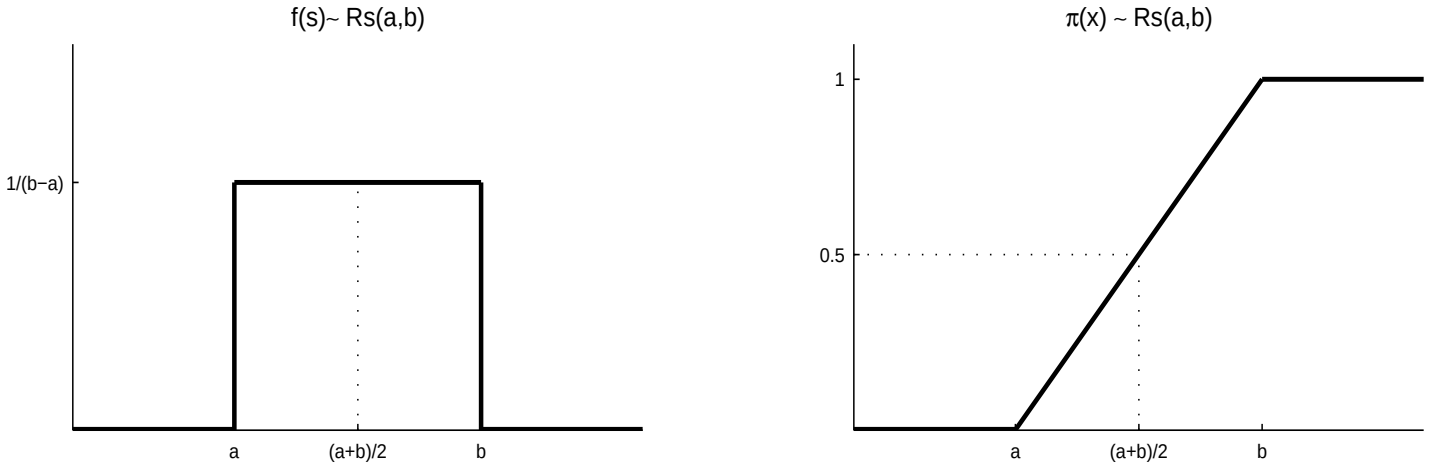
a tento model je **lineárním modelem**

$$\pi_0(x) = \frac{x-a}{b-a} = \beta_0 + \beta_1 x \qquad \text{tj. } \beta_0 = -\frac{a}{b-a},\ \beta_1 = \frac{1}{b-a} > 0$$

s identickou linkovací funkcí

$$g_0(\pi) = \pi.$$

Symetrické modely



Obrázek : Rovnoměrné rozdělení na (a, b).

Probitový model

Další možností je vzít **normální hustotu** jako **toleranční funkci**. V tomto případě mluvíme o tzv. **probitovém modelu**:

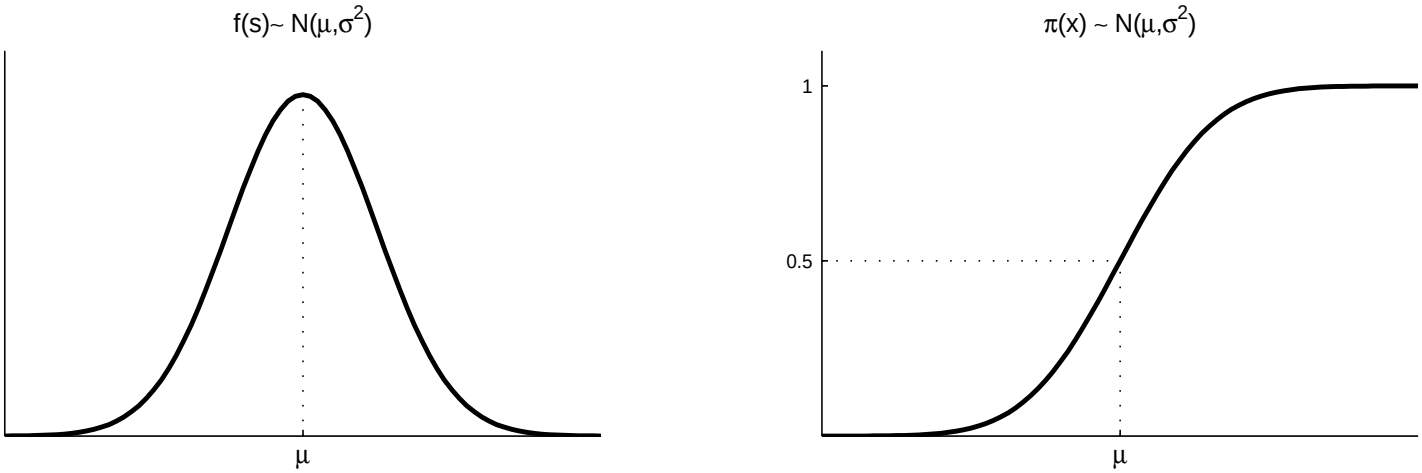
$$\pi_1(x) = F_1(x) = \int_{-\infty}^{x} f_1(s)ds = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2} ds = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

kde Φ je distribuční funkce standardizovaného normálního rozdělení. Pak tzv. **probitovou linkovací funkcí** je kvantilová funkce normálního rozdělení

$$g_1(\pi) = \Phi^{-1}(\pi) = \frac{x-\mu}{\sigma} = \beta_0 + \beta_1 x \qquad \text{tj.} \qquad \beta_0 = -\frac{\mu}{\sigma} \quad \beta_1 = \frac{1}{\sigma} > 0.$$

Hodnota mediánu $x = \mu$ se nazývá **mediánová smrtící dávka** (*median lethal dose* - LD50) a odpovídá dávce, při které polovina jedinců má kladnou a polovina zápornou odezvu.

Probitový model



Obrázek : Normální rozdělení $N(\mu, \sigma^2)$.

Logistický model

Jiným velmi podobným modelem je tzv. **logistický model**, kde toleranční funkce je hustota **logistického rozdělení**

$$f_2(s) = \frac{1}{\sigma} \frac{\exp\left(\frac{s-\mu}{\sigma}\right)}{\left[1+\exp\left(\frac{s-\mu}{\sigma}\right)\right]^2} = \frac{1}{\sigma} \frac{\exp\left(-\frac{s-\mu}{\sigma}\right)}{\left[1+\exp\left(-\frac{s-\mu}{\sigma}\right)\right]^2},$$

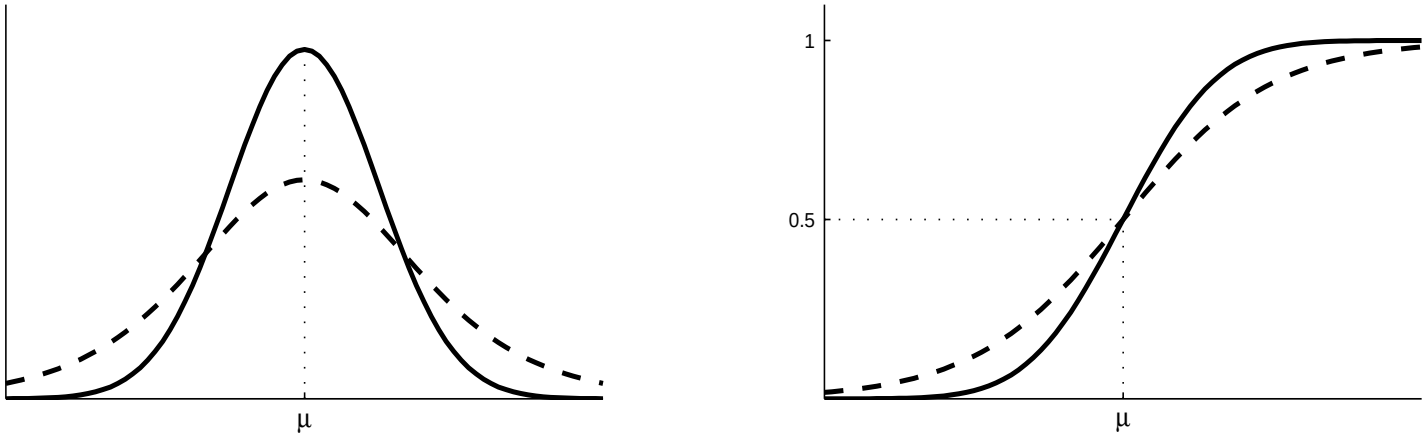
takže

$$\pi_2(x) = F_2(x) = \int_{-\infty}^{x} \frac{1}{\sigma} \frac{\exp\left(\frac{s-\mu}{\sigma}\right)}{\left[1+\exp\left(\frac{s-\mu}{\sigma}\right)\right]^2} \, ds = \frac{\exp\left(\frac{x-\mu}{\sigma}\right)}{1+\exp\left(\frac{x-\mu}{\sigma}\right)} = \frac{1}{1+\exp\left(-\frac{x-\mu}{\sigma}\right)}$$

s tzv. **logit linkovací funkcí**

$$g_2(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \frac{x-\mu}{\sigma} = \beta_0 + \beta_1 x \qquad \text{tj.} \qquad \beta_0 = -\frac{\mu}{\sigma} \quad \beta_1 = \frac{1}{\sigma} > 0.$$

Logistický model



Obrázek : Srovnání probitového a logistického (- - -) modelu při stejných parametrech μ a σ.

CLogLog model

Asymetrické (extremální) modely

Pokud za toleranční funkci zvolíme **Log-Weibullovo rozdělení** (*extreme-minimal-value distribution*) ve tvaru

$$f_3(s) = \frac{1}{\sigma} \exp\left(\frac{s-\mu}{\sigma}\right) \exp\left[-\exp\left(\frac{s-\mu}{\sigma}\right)\right],$$

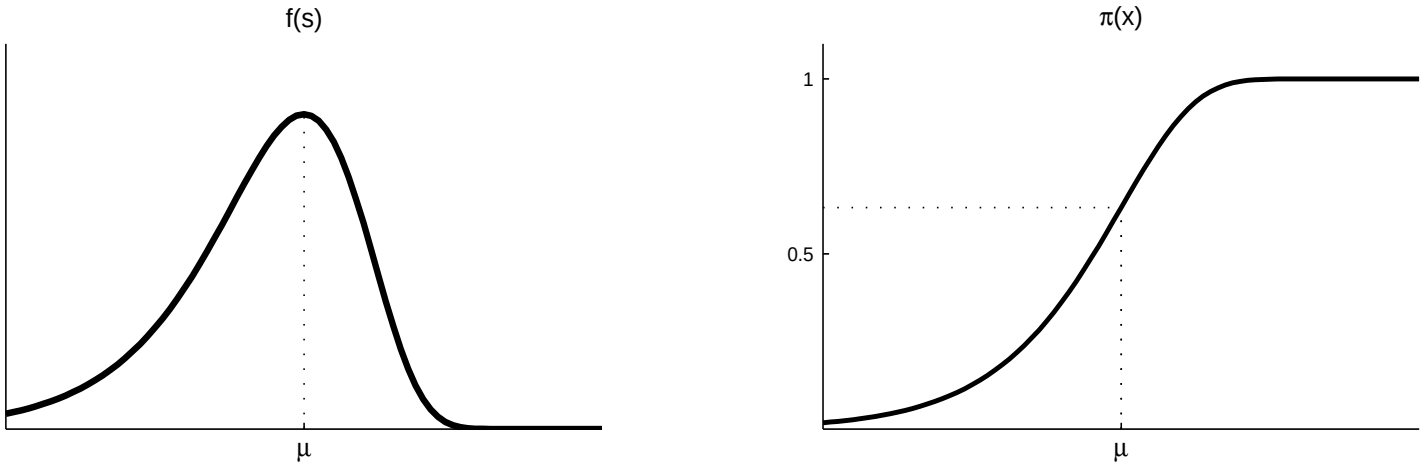
pak

$$\pi_3(x) = \int_{-\infty}^{x} \frac{1}{\sigma} \exp\left(\frac{s-\mu}{\sigma}\right) \exp\left[-\exp\left(\frac{s-\mu}{\sigma}\right)\right] \, ds = 1 - \exp\left[-\exp\left(\frac{x-\mu}{\sigma}\right)\right]$$

s tzv. **komplementární log-log linkovací funkcí**

$$g_3(\pi) = \log[-\log(1-\pi)] = \frac{x-\mu}{\sigma} = \beta_0 + \beta_1 x \quad \text{tj.} \quad \beta_0 = -\frac{\mu}{\sigma} \quad \beta_1 = \frac{1}{\sigma} > 0.$$

CLogLog model



Obrázek : Log-Weibullovo rozdělení.

LogLog model

Pokud jako toleranční funkci zvolíme **zobecněné Gumbelovo rozdělení** (*extreme-maximal-value distribution*) ve tvaru

$$f_4(s) = \frac{1}{\sigma} \exp\left(-\frac{s-\mu}{\sigma}\right) \exp\left[-\exp\left(-\frac{s-\mu}{\sigma}\right)\right],$$

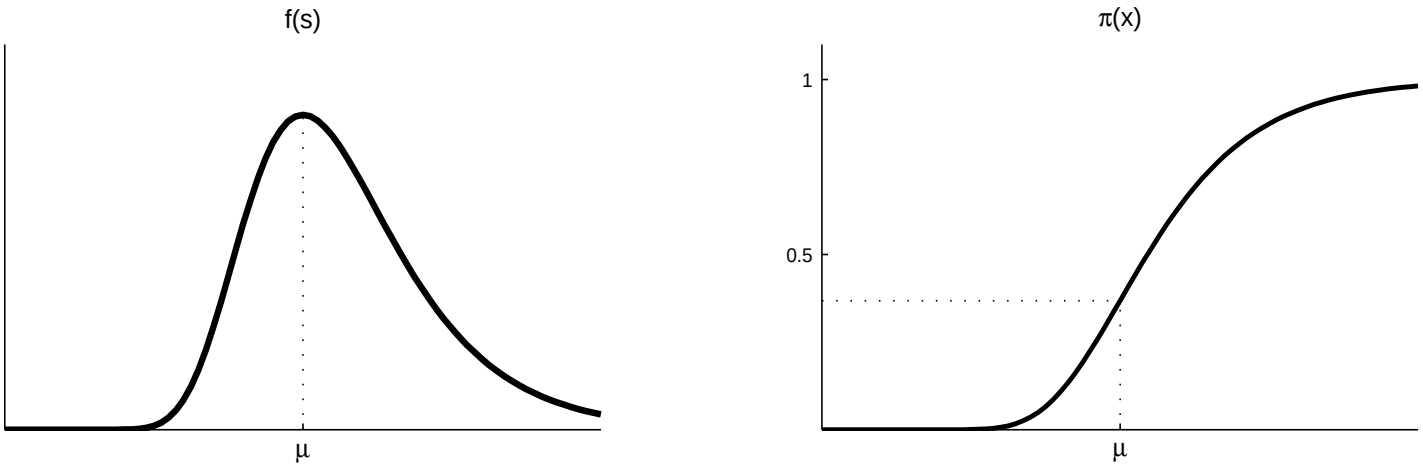
dostaneme

$$\pi_4(x) = \int_{-\infty}^{x} \frac{1}{\sigma} \exp\left[-\frac{s-\mu}{\sigma} - \exp\left(-\frac{s-\mu}{\sigma}\right)\right] ds = \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right]$$

s tzv. **log-log linkovací funkcí**

$$g_3(\pi) = -\log[-\log(\pi)] = \frac{x-\mu}{\sigma} = \beta_0 + \beta_1 x \quad \text{tj.} \quad \beta_0 = -\frac{\mu}{\sigma} \quad \beta_1 = \frac{1}{\sigma} > 0.$$

LogLog model



Obrázek : Zobecněné Gumbelovo rozdělení.

Logistická regrese

Nejčastěji se používá **logit** linkovací funkce

$$g_2(\pi) = \log\left(\frac{\pi}{1-\pi}\right).$$

Zajímá nás vztah pravděpodobností úspěchu či neúspěchu k hodnotám regresorů (kovariát) $\mathbf{x} = (x_1, \ldots, x_k)^{\mathrm{T}}$, tj.

$$P(Y = 1|x_1, \ldots, x_k) = \pi(\mathbf{x}) = \frac{\exp\{\eta(\mathbf{x})\}}{1 + \exp\{\eta(\mathbf{x})\}} = \frac{1}{1 + \exp\{-\eta(\mathbf{x})\}}$$

a

$$P(Y = 0|x_1, \ldots, x_k) = 1 - \pi(\mathbf{x}) = \frac{1}{1 + \exp\{\eta(\mathbf{x})\}} = \frac{\exp\{-\eta(\mathbf{x})\}}{1 + \exp\{-\eta(\mathbf{x})\}}$$

Předpokládejme, že lineární prediktor je roven

$$\eta(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}.$$

Logistická regrese

Všimněme se nejprve, že podíl

$$odds(x_1, \ldots, x_k) = \frac{P(Y = 1|x_1, \ldots, x_k)}{P(Y = 0|x_1, \ldots, x_k)} = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})$$

má bezprostřední interpretaci. Porovnává pravděpodobnost jedničky (tj. výskyt sledovaného jevu při daných hodnotách kovariát) a nuly (nevýskyt sledovaného jevu při daných hodnotách kovariát). Anglickému označení **odds** odpovídá české označení **šance**.
Pro $k = 1$ jsou šance

$$odds(0) = exp(\beta_0),$$

$$odds(1) = exp(\beta_0 + \beta_1).$$

Poměr šancí (anglicky *odds ratio*) pro binární x je pak

$$OR = \frac{odds(1)}{odds(0)} = \exp(\beta_1),$$

takže parametr β_1 je roven logaritmu poměru šancí.

Příklad

Příklad 1

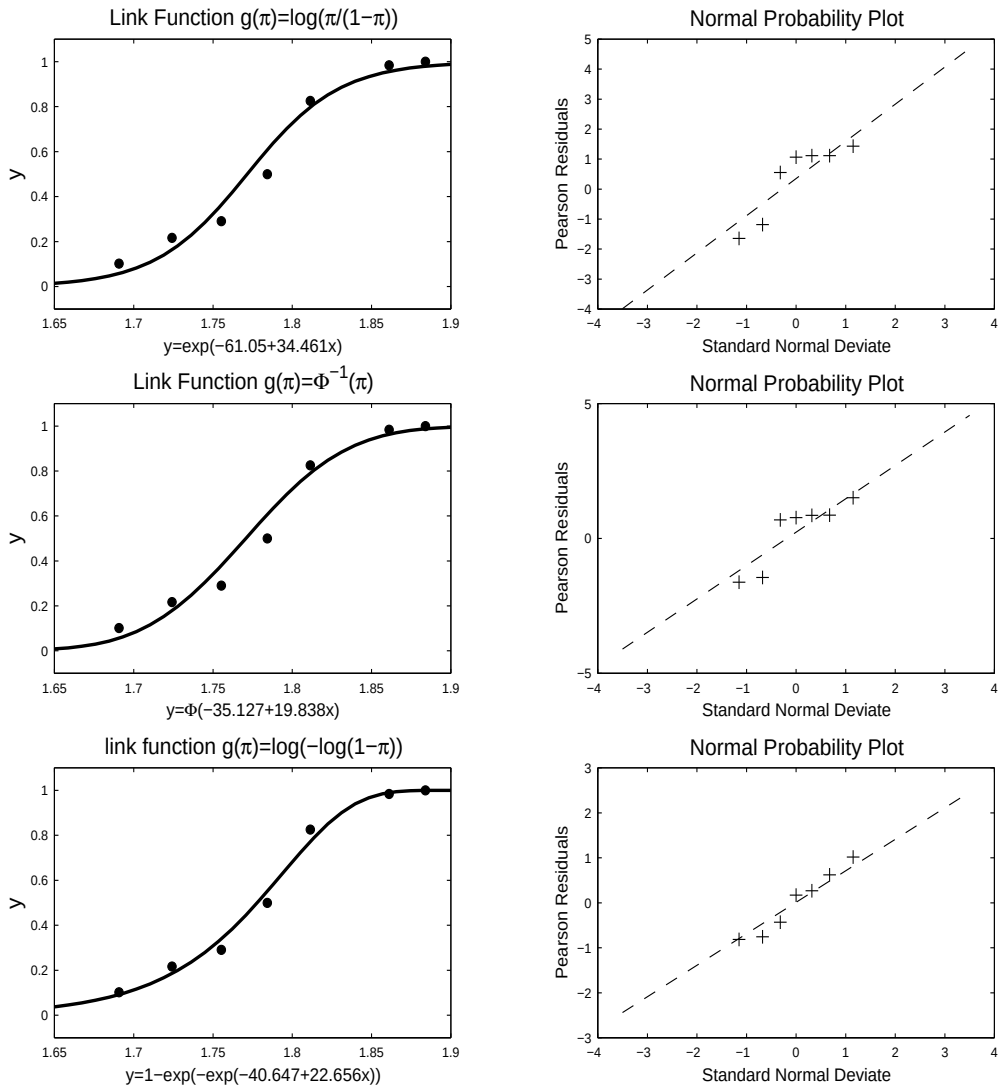
V souboru „`beetle.RData`“ jsou uvedeny údaje o úmrtnosti Potemníka skladištního (Tribolium confusum) v reakci na sirouhlík CS_2. Datový soubor obsahuje tyto proměnné

`dose`	*množství sirouhlíku* (mg/l)
`population`	*počet kusů ve zkoumaném vzorku*
`killed`	*počet mrtvých kusů ve zkoumaném vzorku*

Modelujte závislost úmrtnosti na množství CS_2.

Řešení. Pro modelování závislosti použijeme logistický model, probitový model a model s komplementární log-log linkovací funkcí.

Příklad



Obrázek : Modely pro úmrtnost *Potemníka skladištního*.

Modely pro poissonovská data

Předpokládejme, že náhodný výběr rozsahu n je z Poissonova rozdělení, tj

$$Y_i \sim Po(\lambda_i), \quad P(Y_i = y) = \begin{cases} \frac{\lambda^y e^{-\lambda_i}}{y!} = & \lambda_i > 0;\ y = 0, 1, 2, \ldots \\ 0 & \text{jinak} \end{cases}$$

přičemž

$$EY_i = DY_i = \lambda_i.$$

Y_i . . . **počet výskytu sledovaného jevu v určitém časovém intervalu** (na ploše velikosti t apod.).
Jestliže jsou splněny následující podmínky

a) jev může nastat v kterémkoliv časovém okamžiku,

b) počet výskytů jevu během časového intervalu závisí jen na jeho délce a ne na jeho počátku ani na tom, kolikrát jev nastoupil před jeho počátkem,

c) pravděpodobnost, že jev nastoupí více než jednou v intervalu délky t, konverguje k nule rychleji než t,

d) λ je **střední hodnota počtu výskytů jevu za časovou jednotku**

pak uvedená náhodná veličina má rozdělení $Po(\lambda)$.

Modely pro poissonovská data

Náhodnou veličinou, která má Poissonovo rozdělení, je tedy např.

- počet vadných výrobků ve velké sérii, jestliže pravděpodobnost vyrobení vadného výrobku je velmi malá
- počet těžkých dopravních úrazů za den v určitém městě
- počet zákazníků v prodejně během nějakého časového intervalu
- počet částic v jednotce plochy nebo objemu, např. počet částic v zorném poli mikroskopu
- počet telefonních volání v časovém intervalu t
- počet létavic pozorovaných během intervalu délky t

Modely pro poissonovská data

Předpokládejme, že Y_i závisí na k veličinách $x_{i1}, \ldots, x_{ik}$, a úkolem je najít vztah mezi nimi, tj. hledáme funkci

$$\lambda_i = \lambda(\mathbf{x}_i) = \lambda(x_{i1}, \ldots, x_{ik}).$$

GLM modely $\Rightarrow$ modelujeme pravděpodobnosti λ_i pomocí linkovacích funkcí

$$g(\lambda_i) = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}.$$

Nejjednodušším je lineární model

$$\lambda_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}.$$

Tento model má řadu nevýhod, především je třeba zajistit, aby $\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}$ nabývala pouze kladných hodnot. Nejčastěji se volí tyto dvě možnosti:

log-lineární model:

$$EY_i = \mu_i = \lambda_i = \exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}) \quad \Rightarrow \quad g_1(\mu_i) = \eta_i = g_1(\lambda_i) = \log(\lambda_i) = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}$$

odmocninový model:

$$EY_i = \mu_i = \lambda_i = \left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)^2 \quad \Rightarrow \quad g_2(\mu_i) = \eta_i = g_2(\lambda_i) = \sqrt{\lambda_i} = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}$$

Modelování binomických dat pomocí poissonovského modelu

Pomocí Poissonova rozdělení $Po(\lambda)$ lze dobře aproximovat binomické rozdělení $Bi(n, \pi)$ za podmínek

$$n \to \infty \quad \& \quad \pi \to 0 \quad \& \quad n\pi \to \lambda < \infty,$$

obvykle se doporučuje $n > 30$ a $\pi < 0,1$.
Chceme-li tedy aproximovat binomické rozdělení $Bi(n_i, \pi_i)$ pomocí Poissonova rozdělení $Po(\lambda_i = n_i\pi_i)$ a přitom použijeme logaritmickou linkovací funkci, platí

$$\lambda_i = n_i\pi_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta}) \Rightarrow \log(\lambda_i) = \underbrace{\log(n_i)}_{\text{tzv. „offset“}} + \log(\pi_i) = \mathbf{x}_i^T\boldsymbol{\beta}.$$

Příklad

Příklad 2

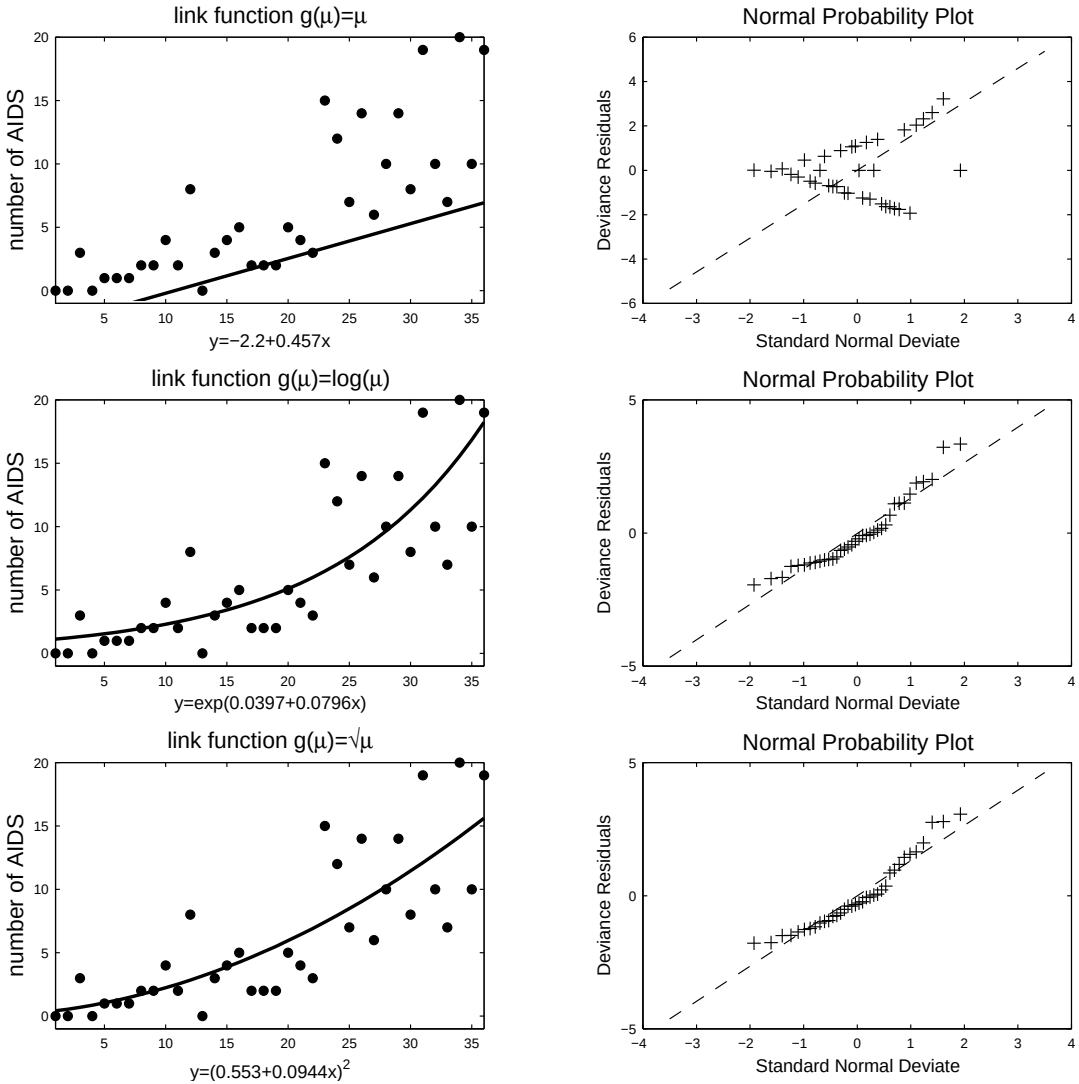
V souboru „`aids.RData`“ jsou uvedeny údaje o počtech nových případů AIDS ve Velké Británii za období prosinec 1982 až listopad 1985. Datový soubor obsahuje tyto proměnné

`month`	*měsíc*
`year`	*rok*
`number`	*počet nových případů AIDS*

Modelujte závislost počtu nových případů AIDS na čase.

Řešení. Pro modelování závislosti použijeme lineární model, log-lineární model a odmocninový model.

Příklad



Obrázek : Modely pro výskyt nových onemocnění AIDS ve Velké Británii.

Úlohy k procvičení

Příklad 1

V souboru „`heart.RData`“ jsou uvedena data o přítomnosti infarktu myokardu v závislosti na věku pacienta. Datový soubor obsahuje tyto proměnné:

`age`	*věk pacienta (roky)*
`chd`	*indikátor infarktu (1 – nastal, 0 – nenastal)*

Pro modelování závislosti použijte logistický model, probitový model a model s komplementární log-log linkovací funkcí. Výsledky vykreslete do obrázku.

Úlohy k procvičení

Příklad 2

V souboru „`nemocnice.RData`“ jsou uvedeny údaje o zotavení pacientů v závislosti na závažnosti onemocnění a nemocnici, ve které se léčili. Datový soubor obsahuje tyto proměnné:

`Infection_Severity`	*vážnost onemocnění*
`Treatment_Outcome`	*indikátor uzdravení (1 – zdravý, 0 – smrt)*
`Hospital`	*typ nemocnice (1, 2, 3)*

Pro modelování závislosti nalezněte vhodný logistický model. Výsledky vykreslete do obrázku.

Úlohy k procvičení

Příklad 3

V souboru „`cancer.RData`“ jsou uvedeny údaje o počtu onemocnění rakovinou kůže u žen v závislosti na věku a oblasti v USA, ve které pacientky žily. Datový soubor obsahuje tyto proměnné:

`Cases`	*počet onemocnění*
`Town`	*město (0 – Minneapolis (Minnesota), 1 – Dallas (Texas))*
`Age`	*věková skupina pacientky*
`Population`	*celkový počet žen dané věkové skupiny v příslušném městě*

Pro modelování závislosti nalezněte vhodný logistický model. Výsledky vykreslete do obrázku. Porovnejte pravděpodobnost vzniku onemocnění u 60-ti leté pacientky žijící v Minneapolisu s pravděpodobností pro stejně starou pacientku žijící v Dallasu.

[Minneapolis: 0.00117, Dallas: 0.00276.]

Úlohy k procvičení

Příklad 4

V souboru „`car_income.RData`“ jsou uvedeny údaje o koupi nového auta během posledních 12-ti měsíců v závislosti na příjmu domácnosti a stáří původního auta. Datový soubor obsahuje tyto proměnné:

`purchase`	*indikátor nákupu nového auta (1 – ano, 0 – ne)*
`income`	*roční příjem domácnosti (v tis. dolarů)*
`age`	*stáří původního auta (roky)*

Nejprve vykreslete závislosti proměnné `purchase` na ostatních. Pro modelování závislosti nalezněte vhodný logistický model. Jsou všechny proměnné statisticky významné? Znovu modelujte s použitím proměnné `age` jako `factor`. Opět sledujte statistickou významnost `age`. Vyzkoušejte tuto proměnnou zakomponovat do modelu jako `factor` s méně úrovněmi. Výsledky vykreslete do obrázku.

Úlohy k procvičení I

Příklad 5

V souboru „`druhy.RData`“ jsou k dispozici data, která se týkají dlouhodobého zemědělského experimentu. Bylo sledováno 90 pozemků (pastvin) o rozloze 25 m × 25 m, lišících se v biomase, pH půdy a druhové bohatosti (počet rostlinných druhů na celém pozemku). Je dobře známo, že s rostoucí biomasou dochází k poklesu druhové bohatosti. Ale zůstává otázka, zda rychlost poklesu nesouvisí s úrovní pH v půdě. Proto byly jednotlivé pozemky klasifikovány podle hodnoty pH v půdě do tří úrovní (nízká, střední a vysoká úroveň) a do experimentu bylo vybráno vždy po 30 pozemcích pro každou úroveň. Spojitá veličina `Biomass` je dlouhodobým průměrem naměřených červnových hodnot biomasy. Datový soubor obsahuje tyto proměnné:

`pH`	úroveň pH v půdě (`low` – nízká, `mid` – střední, `high` – vysoká)
`Biomass`	množství biomasy
`species`	počet rostlinných druhů

Úlohy k procvičení II

Příklad 5

Nejprve vykreslete závislosti proměnné `species` *na ostatních. Pro modelování závislosti nalezněte vhodný poissonovský model. Vyzkoušejte postupně logaritmickou, identickou a odmocninovou linkovací funkci. Jsou všechny proměnné statisticky významné? Pokud ne, zkuste modely zjednodušit a pomocí analýzy deviace rozhodněte, zda takové zjednodušení je možné. Získané výsledné modely vykreslete do obrázku. Pomocí všech modelů odhadněte počet rostlinných druhů na pozemku s hodnotou biomasy 9 a střední úrovní pH v půdě.*

[Odhady počtu druhů pro log link: 8,895, identity link: 4,513, sqrt link: 7,414.]