MA012 Statistika II

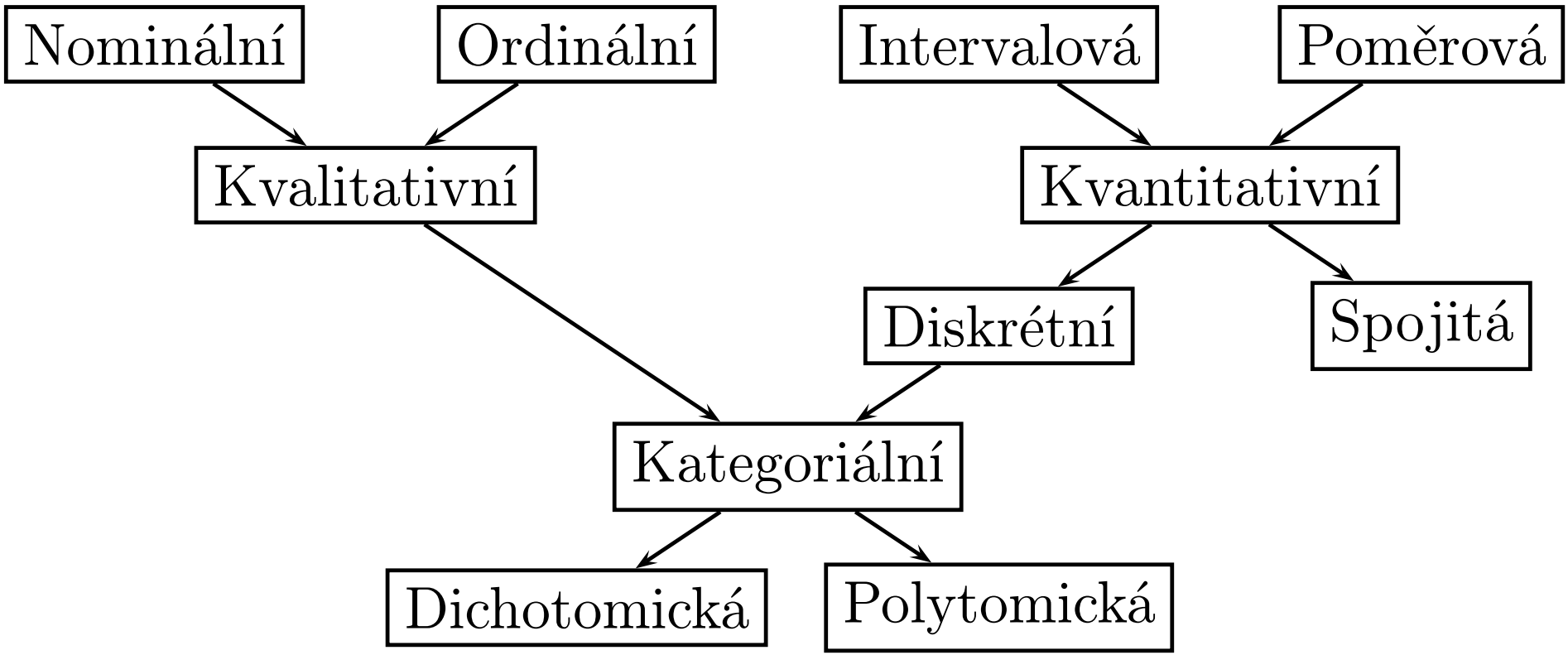
10. Konkrétní GLM modely

Ondřej Pokora (pokora@math.muni.cz)

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

(podzim 2015)

Typy náhodných veličin



Spojitá data s normálně rozdělenými chybami

Předpokládejme, že $Y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, n$.

Přirozeným parametrem je střední hodnota, $\theta_i = \mu_i$. Pro modelování střední hodnoty $\mathrm{E}Y_i = \mu_i$ používáme identickou linkovací funkci $g = \mathrm{id}$,

$$g(\mu_i) = \mu_i = \theta_i = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Takový GLM vede na klasický lineární regresní model.

Gaussova-Markovova věta zaručuje, že odhady parametrů $\widehat{\boldsymbol{\beta}}$ metodou nejmenších čtverců jsou BLUE (nejlepší lineární nestranný odhad, best unbiased linear estimator), tedy pro LRM s parametry $\widehat{\boldsymbol{\beta}}$ je dosažen nejmenší reziduální součet čtverců. Normalita dat zaručuje, že odhady $\widehat{\boldsymbol{\beta}}$ metodou nejmenších čtverců jsou stejné, jako odhady $\widehat{\boldsymbol{\beta}}_{ML}$ metodou maximální věrohodnosti.

Alternativní a binomická data

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

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

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

- **jednotlivá pozorování** U_i:

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

- **skupinově** – pomocí **absolutních četností** úspěchů Y_j v n_j pokusech:

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

kde $j = 1, \ldots, n$, $N = n_1 + \cdots + n_n$, a data můžeme zapsat formou tabulky

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

Alternativní a binomická data

- **skupinově** – pomocí **relativních četností** úspěchů $Z_j = Y_j/n_j$ v n_j pokusech:

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

kde $j = 1, \ldots, n$, $N = n_1 + \cdots + n_n$, a data lze zapsat do tabulky

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

Úkolem statistické analýzy je nalézt vztah mezi Z_i (resp. Y_i) a kovariátami $x_{i1}, \ldots, x_{ik}$. V GLM modelujeme pravděpodobnosti p_i pomocí linkovací funkce

$$g(p_i) = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Nejjednodušším modelem je **lineární model** s $g = \mathrm{id}$,

$$p_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Avšak tento model má řadu nevýhod, především je třeba zajistit, aby $\boldsymbol{x}_i'\boldsymbol{\beta}$ nabývala hodnot mezi 0 a 1, tedy je třeba přidat nějaké dodatečné podmínky.

Probitový model

Probitový model používá tzv. **probitovou** linkovací funkci g, což je kvantilová funkce Φ^{-1} standardizovaného normálního rozdělení. GLM je tedy ve tvaru

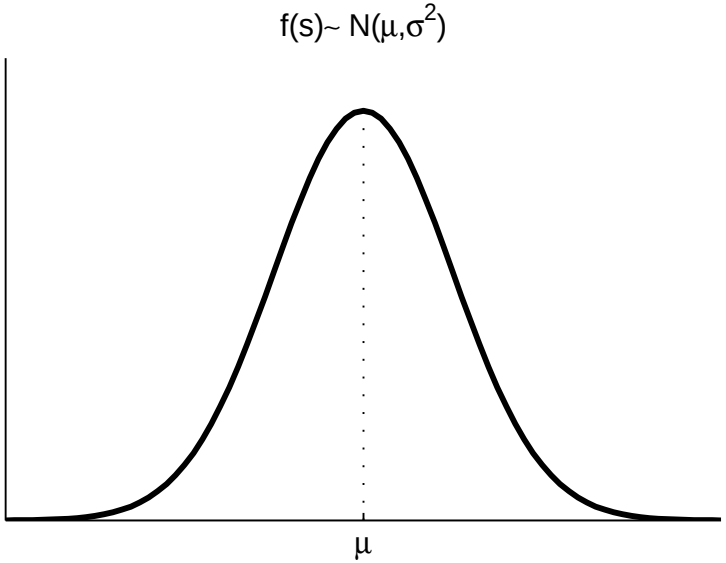
$$g(p_i) = \Phi^{-1}(p_i) = u_{p_i} = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

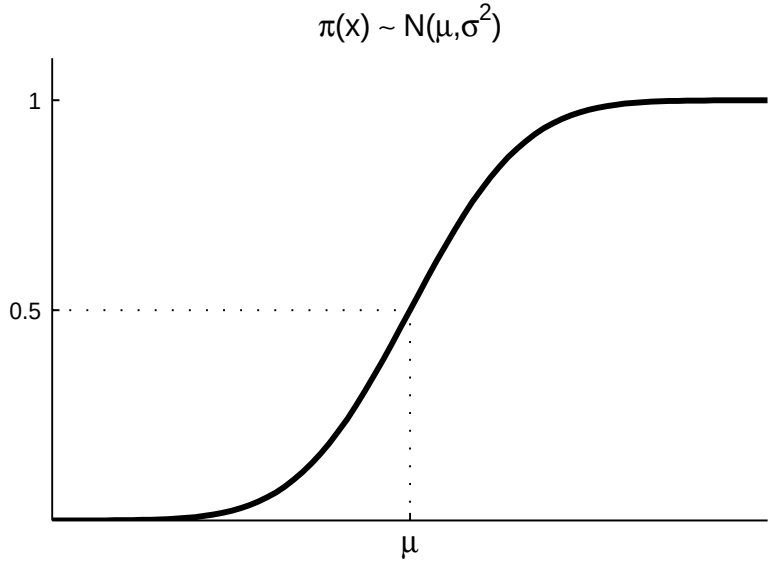
To znamená, že pravděpodobnost úspěchu je modelována vztahem

$$p_i = g^{-1}(\eta_i) = \Phi(\eta_i) = \Phi\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right),$$

kde Φ je distribuční funkce standardizovaného normálního rozdělení.

V případě $\eta_i = \beta_0 + \beta_1 x_i$ potom dostáváme $p_i = \Phi\left(\beta_0 + \beta_1 x_i\right) = F(x_i)$,
kde F je distribuční funkce rozdělení $N(-\beta_0/\beta_1,\ \beta_1^{-2})$.





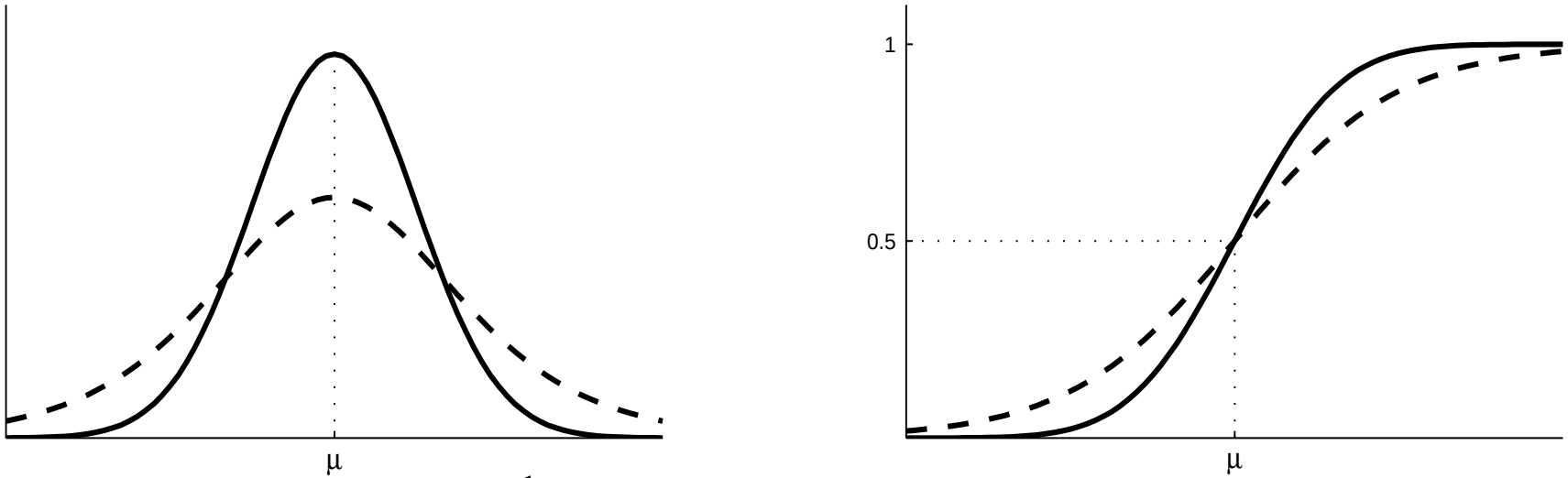
Logistický (logitový) model – logistická regrese

Logistický model používá tzv. **logitovou** linkovací funkci g, což je kvantilová funkce logistického rozdělení pravděpodobnosti. GLM je ve tvaru

$$g(p_i) = \ln \frac{p_i}{1-p_i} = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Pravděpodobnost úspěchu, resp. neúspěchu, je modelována vztahem

$$p_i = g^{-1}(\eta_i) = \frac{1}{1+\exp(-\boldsymbol{x}_i'\boldsymbol{\beta})}, \qquad 1-p_i = \frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}.$$



Obr. Porovnání funkcí g^{-1} v probitovém (plně) a logitovém (čárkovaně) GLM

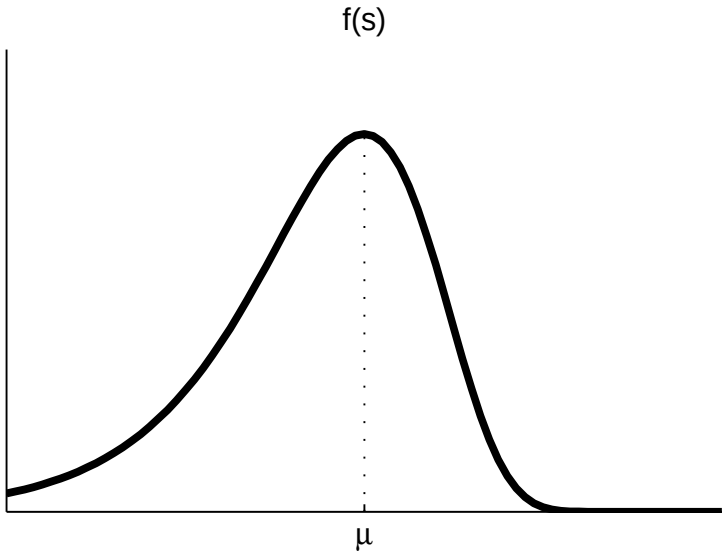
CLogLog (komplementární LogLog) model

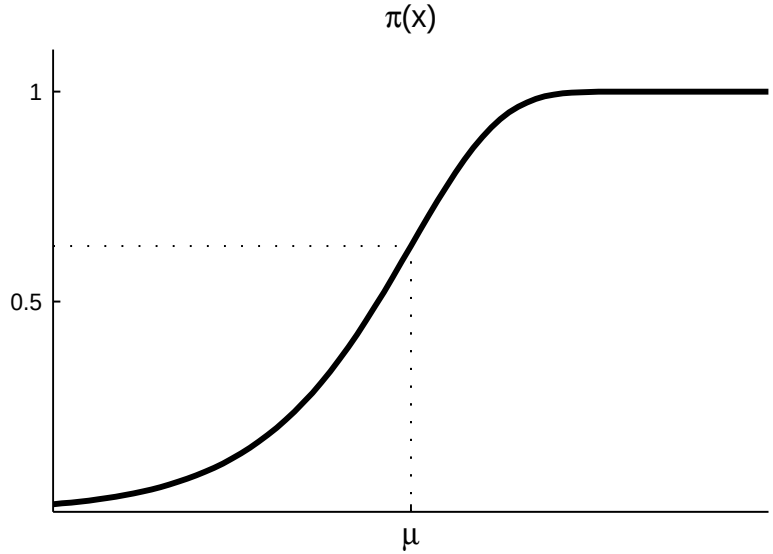
CLogLog model používá jako linkovací funkci g kvantilovou funkci logaritmického Weibullova *(extreme-minimal-value)* rozdělení pravděpodobnosti. GLM je ve tvaru

$$g(p_i) = \ln\left[-\ln(1-p_i)\right] = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Pravděpodobnost úspěchu je tedy modelována vztahem

$$p_i = g^{-1}(\eta_i) = 1 - \exp\left[-\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right].$$





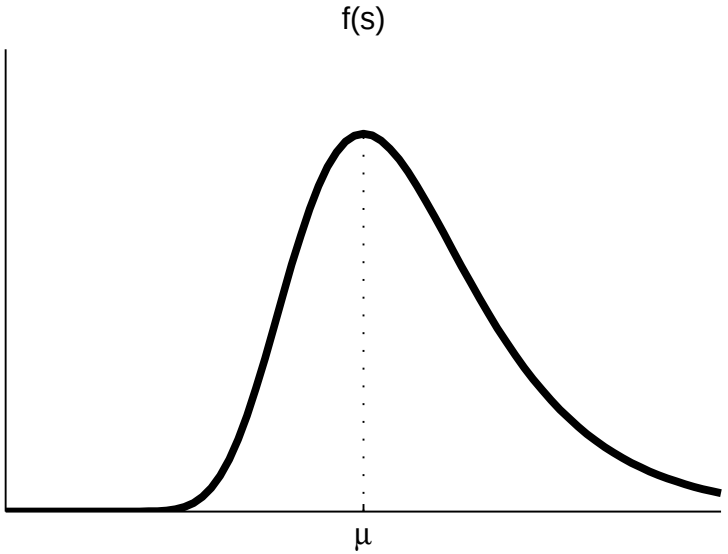
LogLog model

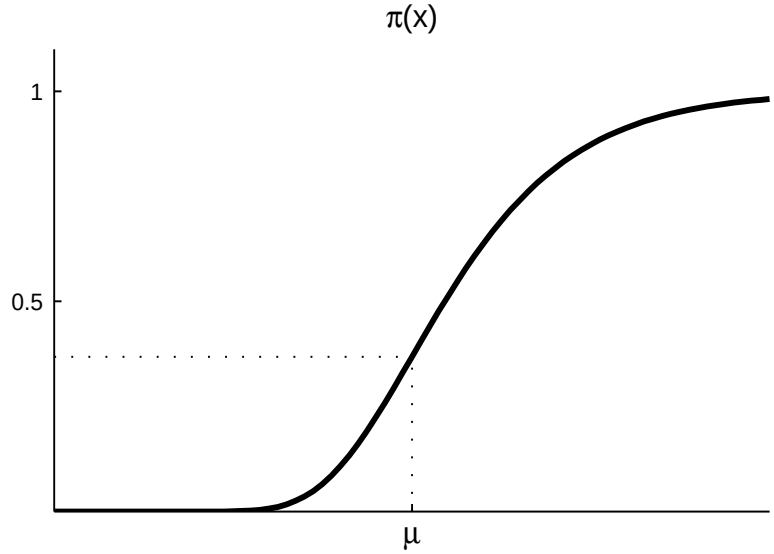
LogLog model používá jako linkovací funkci g kvantilovou funkci zobecného Gumbelova *(extreme-maximal-value)* rozdělení pravděpodobnosti. GLM je ve tvaru

$$g(p_i) = -\ln(-\ln p) = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Pravděpodobnost úspěchu je tedy modelována vztahem

$$p_i = g^{-1}(\eta_i) = \exp\left[-\exp(-\boldsymbol{x}_i'\boldsymbol{\beta})\right].$$





Logistická regrese a šance

Definice 1 (Šance)

Podíl pravděpodobnosti úspěchu a pravděpodobnosti neúspěchu v binomickém (příp. alternativním) rozdělení pravděpodobnosti se nazývá **šance (odds)**,

$$odds = \frac{P(Y=1)}{P(Y=0)} = \frac{p}{1-p}.$$

Logaritmus $\ln odds$ je v finančnictví někdy nazýván **skóre**.

V **logistické regresi** se používá logitová linkovací funkce $g(p) = \ln \frac{p}{1-p}$.
Pravděpodobnost úspěchu, resp. neúspěchu, je pak rovna

$$p = g^{-1}(\eta) = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}, \qquad 1-p = \frac{1}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})},$$

a šance vychází jako exponenciála lineárního prediktoru,

$$odds = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})} \; \frac{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1} = \exp(\boldsymbol{x}'\boldsymbol{\beta}) = \mathrm{e}^{\eta}, \qquad \ln odds = \eta = \boldsymbol{x}'\boldsymbol{\beta}.$$

Logistická regrese a podíl šancí

Pokud v logistické regresí modelujeme binární veličinu $Y \in \{0,\ 1\}$ pomocí taktéž binární kovariáty $x \in \{0,\ 1\}$, měříme asociovanost těchto dvou veličin pomocí statistiky OR:

Definice 2 (Podíl šancí)

Podíl šancí (odds ratio, podíl rizik) je podílem podmíněných šancí,

$$OR = \frac{odds(x=1)}{odds(x=0)} = \frac{\frac{P(Y=1\,|\,x=1)}{P(Y=0\,|\,x=1)}}{\frac{P(Y=1\,|\,x=0)}{P(Y=0\,|\,x=0)}} = \frac{P(Y=1\,|\,x=1)\ P(Y=0\,|\,x=0)}{P(Y=0\,|\,x=1)\ P(Y=1\,|\,x=0)}.$$

- $OR \approx 1$: nekorelovanost
- $OR > 1$: korelovanost, souhlasná asociovanost
- $OR < 1$: korelovanost, nesouhlasná asociovanost

Dosazením šancí dostáváme výpočetní tvar

$$OR = \frac{odds(x=1)}{odds(x=0)} = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})\,|_{x=1}}{\exp(\boldsymbol{x}'\boldsymbol{\beta})\,|_{x=0}}.$$

Logistická regrese a podíl šancí

V případě modelu logistické regrese veličiny Y pomocí binární proměnné $x \in \{0,\ 1\}$ s lineárním prediktorem tvaru $\eta = \beta_0 + \beta_1 x$, tzn. v GLM

$$Y \sim Bi(n,\, p), \qquad \ln \frac{p}{1-p} = \beta_0 + \beta_1 x,$$

dostáváme

$$OR = \frac{\exp(\beta_0 + \beta_1 x)\,|_{x=1}}{\exp(\beta_0 + \beta_1 x)\,|_{x=0}} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \mathrm{e}^{\beta_1},$$

tedy lineární koeficient β_1 binární kovariáty x v modelu logistické regrese je rovný logaritmu podílu šancí, $\beta_1 = \ln OR$.

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 $\mathrm{E}Y_i = \mathrm{D}Y_i = \lambda_i$.

Y_i = počet výskytů sledovaného jevu v určitém časovém intervalu, na dané ploše, v daném objemu, apod. 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.

Modely pro Poissonovská data

Logaritmicko-lineární model používá logaritmickou linkovací funkci g. GLM je ve tvaru

$$g(\lambda_i) = \ln \lambda_i = \eta_i = \boldsymbol{x}_i' \boldsymbol{\beta}.$$

Střední počet událostí je tedy modelován exponenciálně,

$$\lambda_i = g^{-1}(\eta_i) = \exp(\boldsymbol{x}_i' \boldsymbol{\beta}).$$

Odmocninový model používá linkovací funkci g ve tvaru odmocniny, GLM je ve tvaru

$$g(\lambda_i) = \sqrt{\lambda_i} = \eta_i = \boldsymbol{x}_i' \boldsymbol{\beta}.$$

Střední počet událostí je tedy modelován kvadraticky,

$$\lambda_i = g^{-1}(\eta_i) = (\boldsymbol{x}_i' \boldsymbol{\beta})^2.$$

Další modely

Pro data s **exponenciálním a gama** rozdělením pravděpodobnosti, tj. pro spojitá data s nekonstantní chybou a konstantním poměrem střední hodnoty a směrodatné odchylky, se obvykle používá **inverzní** linkovací funkce g. GLM je ve tvaru

$$g(\mu_i) = \frac{1}{\mu_i} = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Střední hodnota je tedy modelována vztahem

$$\mu_i = g^{-1}(\eta_i) = \frac{1}{\boldsymbol{x}_i'\boldsymbol{\beta}}.$$

Pro data s **inverzním Gaussovým** rozdělením pravděpodobnosti se obvykle používá **inverzní kvadratická** linkovací funkce g. GLM je ve tvaru

$$g(\mu_i) = \frac{1}{\mu_i^2} = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Střední hodnota je tedy modelována vztahem

$$\mu_i = g^{-1}(\eta_i) = \frac{1}{\sqrt{\boldsymbol{x}_i'\boldsymbol{\beta}}}.$$

Zobecněné lineární modely v R

Obecná funkce pro řešení GLM v *R* je `glm`.

```
model <-glm (formula, family, data)
```

family	family (link = ...)
gaussian	identity, log, inverse
binomial	logit, probit, cloglog, log, cauchit
poisson	log, sqrt, identity
Gamma	inverse, log, identity
inverse.gaussian	1/mu^2, inverse, log, identity

```
v <- summary (model)
```

S výsledky se pracuje analogicky jako s výsledky funkce `lm` pro LRM.

Příklad

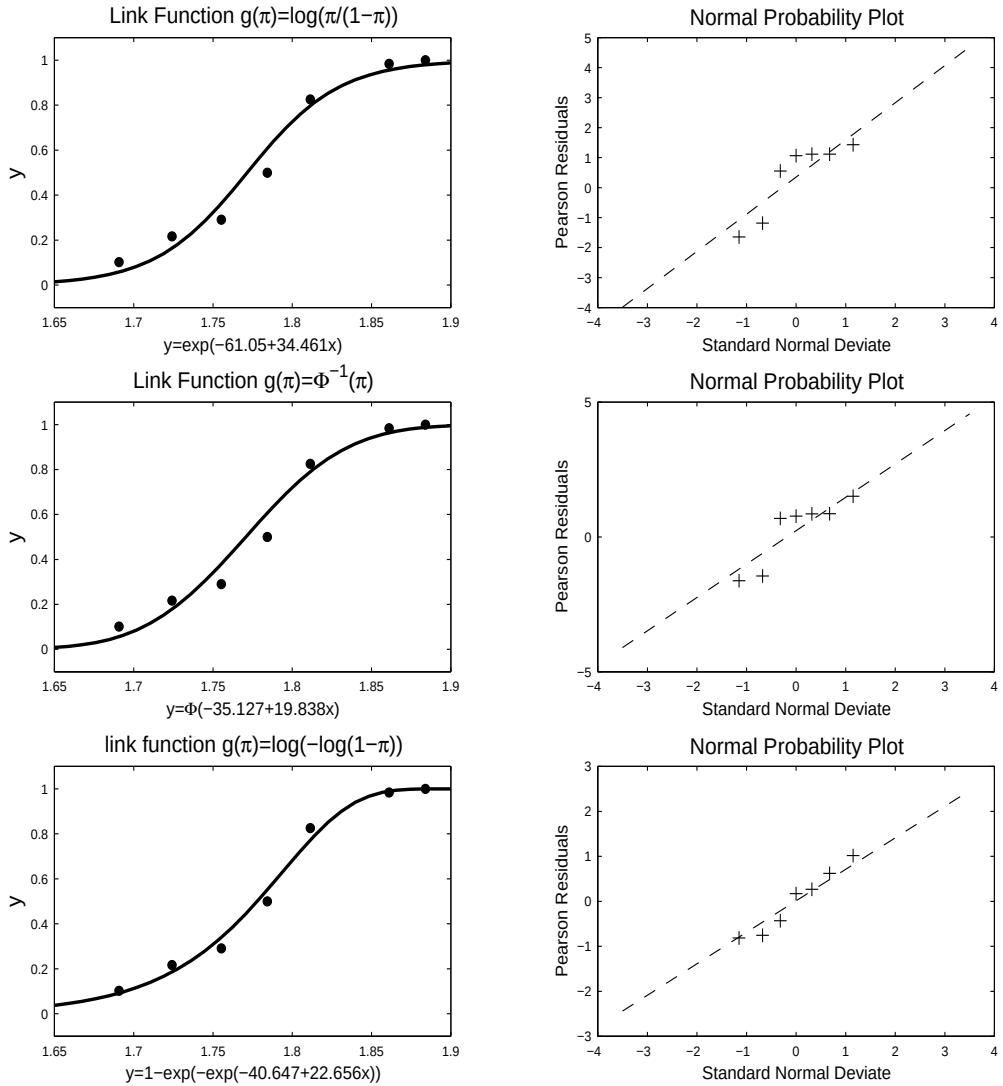
Příklad 1

V souboru `beetle.csv` *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.



Obrázek: Modely pro úmrtnost *Potemníka skladištního*.

Příklad

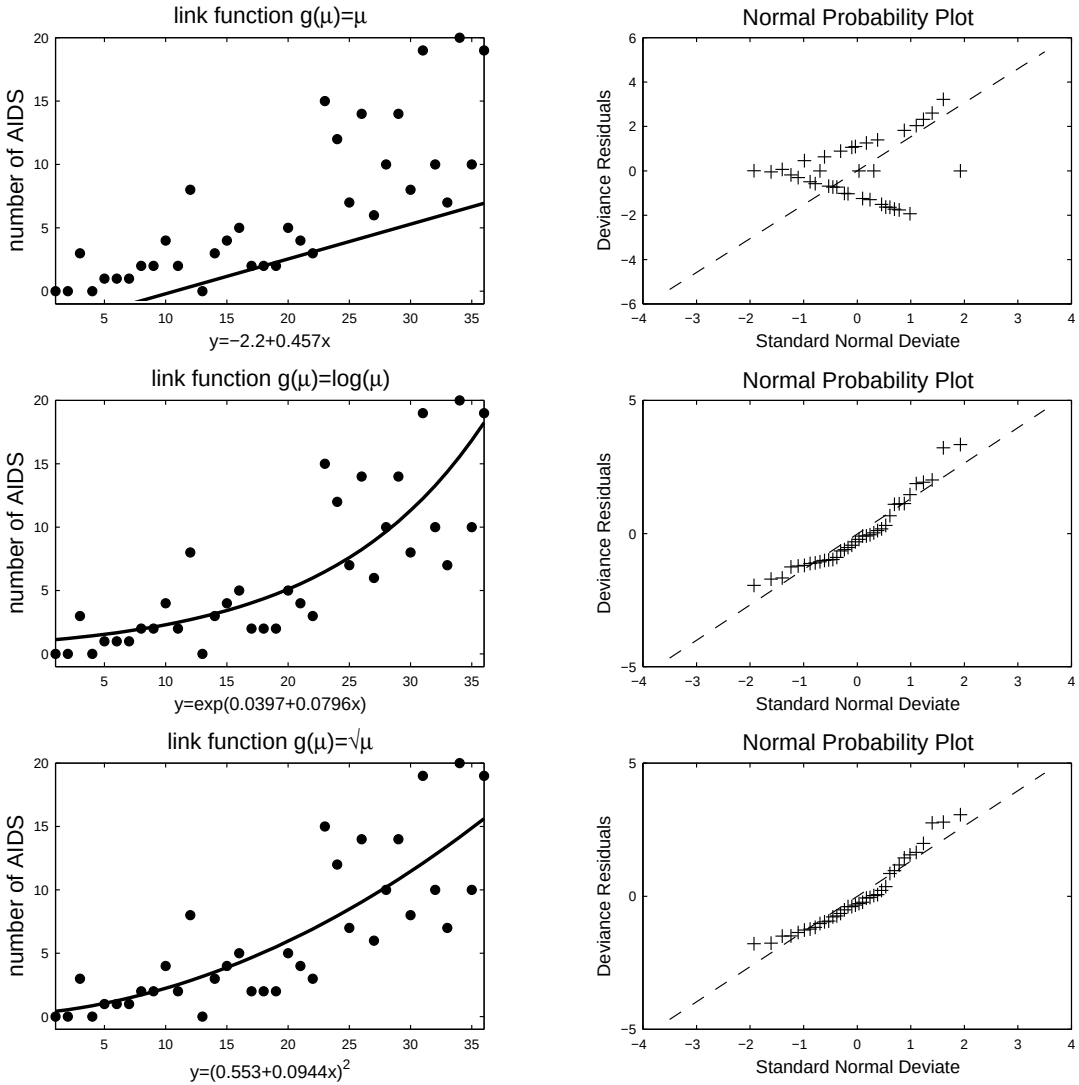
Příklad 2

V souboru `aids.csv` *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.



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

Neškálová deviance

Předpokládáme, že náhodný výběr $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ se řídí GLM modelem, tj.

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \theta_i) = \exp\left\{ \sum_{i=1}^{n} \frac{y_i \theta_i - \gamma(\theta_i)}{\psi_i(\phi)} + d(y_i, \phi) \right\}.$$

Předpokládejme dále, že hustota je ve škálové formě, tj.

$$\psi_i(\phi) = \frac{\phi}{\omega_i} > 0,$$

se známými váhami $\omega_i > 0$ a jedním neznámým rušivým parametrem *(scale factor)* $\phi > 0$.

Škálová deviance je rovna

$$\begin{aligned} D &= 2\left[\ell(\widehat{\boldsymbol{\beta}}_{max}; \boldsymbol{Y}) - \ell(\widehat{\boldsymbol{\beta}}; \boldsymbol{Y})\right] \\ &= \frac{1}{\phi} \underbrace{2 \sum_{i=1}^{n} \omega_i \left[Y_i(\hat{\theta}_{i,max} - \hat{\theta}_i) - \gamma(\hat{\theta}_{i,max}) + \gamma(\hat{\theta}_i) \right]}_{D^*} = \frac{D^*}{\phi} \end{aligned}$$

a D^* nazýváme **neškálovou deviancí (unscaled deviance)**.

Odhady rušivého parametru

Protože platí

$$D = \frac{D^*}{\phi} \overset{as.}{\sim} \chi^2(n-k) \quad \Rightarrow \quad \mathrm{E}D = \frac{\mathrm{E}D^*}{\phi} \approx n-k,$$

lze rušivý parametr ϕ odhadnout pomocí statistiky

$$\hat{\phi}_{D^*} = \frac{D^*}{n-k}.$$

Pro tzv. zobecněnou Pearsonovu statistiku χ^2 platí

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\gamma''(\hat{\mu}_i)} \overset{as.}{\sim} \chi^2(n-k).$$

Odtud lze získat jiný odhad rušivého parametru,

$$\hat{\phi}_{\chi^2} = \frac{\chi^2}{n-k}.$$

Overdispersion, underdispersion

Připomeňme, že v normálním rozdělení je rušivým parametrem rozptyl, $\phi = \sigma^2$ a v gama rozdělení $\phi = 1/k$.

V prostředí R je pro řešení problémů s neadekvátně odhadnutým rozptylem k dispozici modifikovaná volba pro třídu exponenciálního rozdělení. V případě binomického rozdělení máme možnost volby

```
family = quasibinomial
```

a pro Poissonovo rozdělení

```
family = quasipoisson.
```

Nejedná se přiotm o nové rozdělení exponenciálního typu, ale o změnu ve výpočtu druhého momentu, pro jehož odhad se použije jednoduchý momentový odhad disperzního parametru ϕ. Výsledná korekce rozptylu je pak důležitá při testování hypotéz, neboť zohledňuje vyšší či nižší variabilitu v datech a zabraňuje tak nadbytku či nedostatku falešně pozitivních výsledků testů hypotéz o parametrech modelu.

Příklad

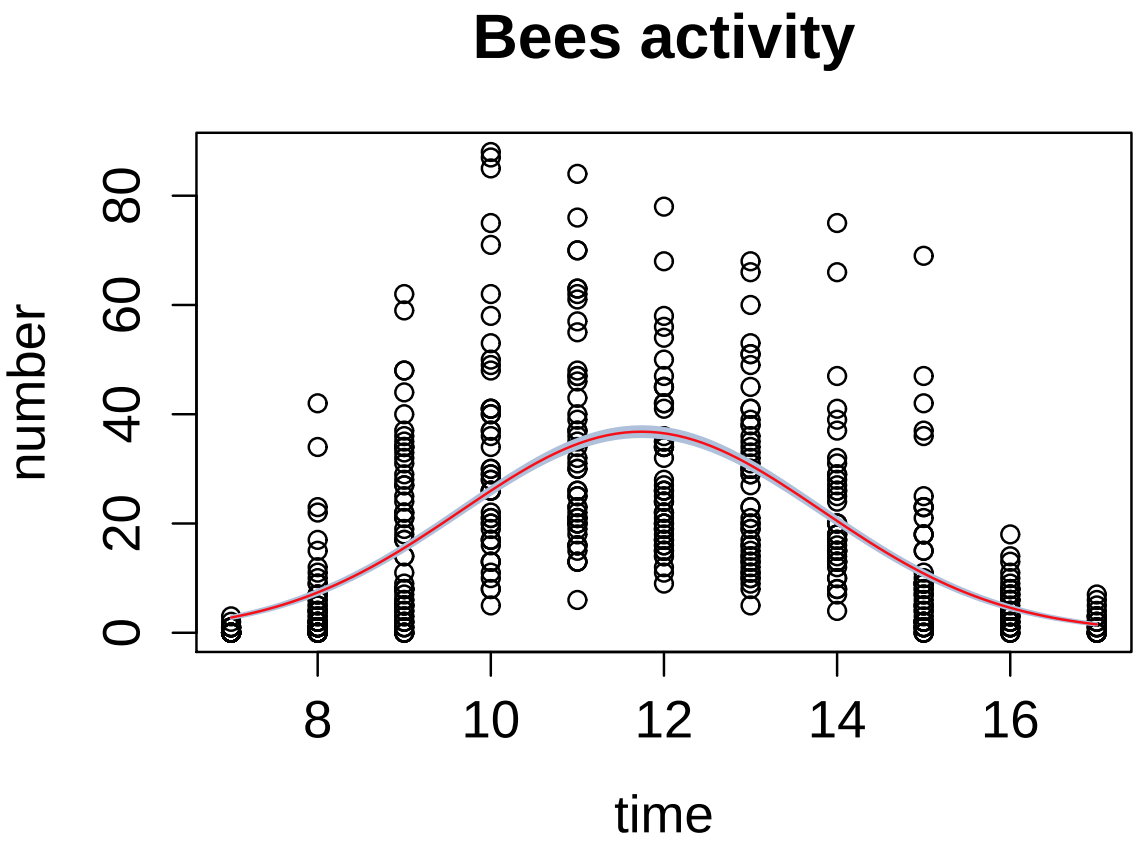
Příklad 3

V souboru `bees.csv` *jsou uvedeny údaje o aktivitě včel v závislosti na čase. Jednou z důležitých charakteristik při zkoumání včelí aktivity je počet včel, které opustí úl kvůli práci ve vnějším prostředí. Studie se zabývala měřením této veličiny během několika slunečných dní v závislosti na čase během dne. Datový soubor obsahuje tyto proměnné*

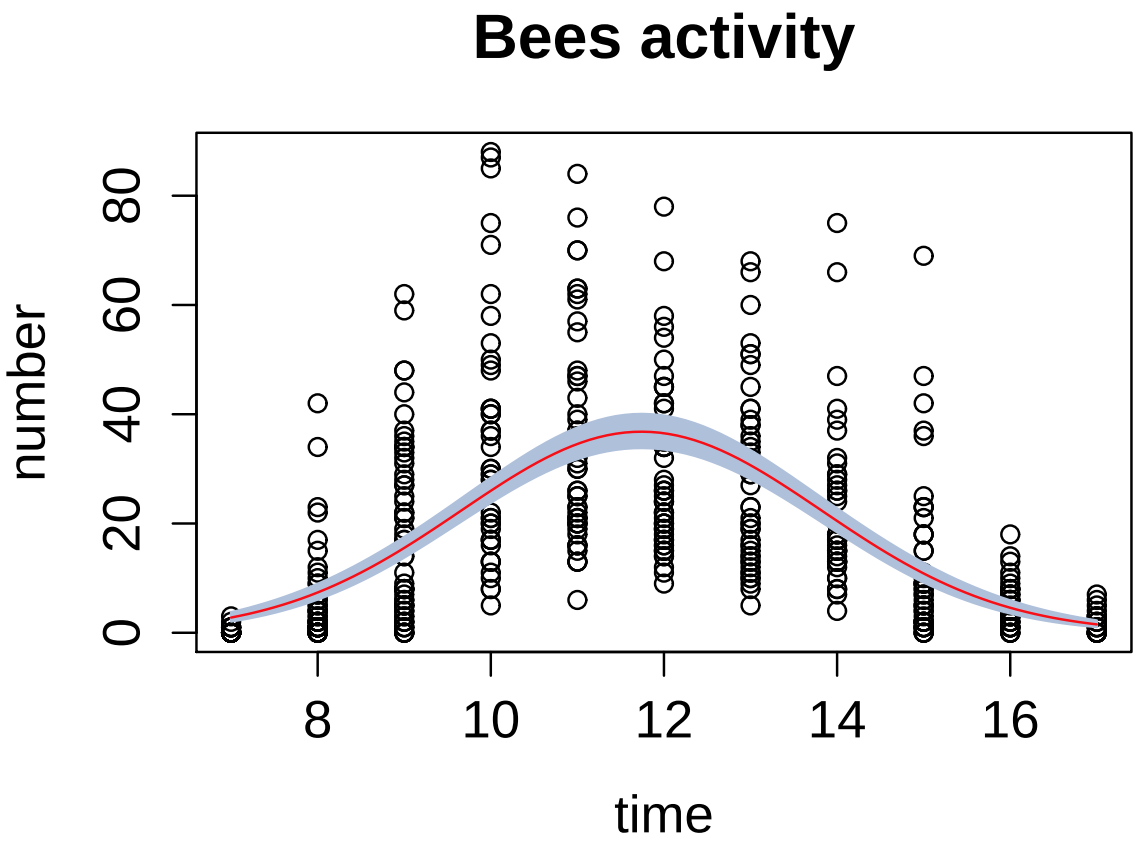
`number`	*počet včel, které opustily úl*
`time`	*čas, kdy byl tento údaj zaznamenán*

Modelujte závislost počtu včel, které opustí úl, na čase během dne.

Řešení. Pro modelování závislosti použijeme poissonovský model s kanonickou linkovací funkcí. Do modelu vstupuje jediná vysvětlující proměnná `time` a přidáme také její druhou mocninu.
Hodnota reziduální deviance (4 879,3) je nepoměrně vyšší než počet stupňů volnosti (501). Je zřejmé, že došlo k „overdispersion“ a v jazyce R je třeba volit `family=quasipoisson`. Použití této volby neovlivňuje odhady koeficientů, ale mění jejich odhady variability, což se projeví např. v intervalu spolehlivosti.



Obrázek: Odhad regresní funkce **bez** vyrovnání se s problematikou velkého rozptylu.



Obrázek: Odhad regresní funkce **s** vyrovnáním se s problematikou velkého rozptylu.