

Bi7491 Regresní modelování

Logistický model

Co byste již měli vědět a umět?

- Vědět, jak se definuje lineární regresní model
- Vysvětlit předpoklady regresních modelů
- Umět použít v lineárním regresním modelu různé typy prediktorů
- Vědět, co je multikolinearita, jak ji zjistit a jak se s ní vypořádat
- Umět se vypořádat s chybějícími daty
- Vědět, co je interakce, jak ji poznat, a jak ji zohlednit v konstruovaném modelu
- Znat možnosti kauzálního působení různých faktorů, umět popsat rozdíl mezi zkreslující proměnnou a mediátorem, popisovat jednoduché vztahy pomocí modelových diagramů
- Znat základní pravidla pro zařazování proměnných do modelu
- Umět posoudit splnění modelových předpokladů pomocí grafických nástrojů

Co byste měli vědět a umět po dnešní hodině ?

- ➔ Znat užitečné veličiny pro měření vztahu/účinku: poměr rizik a poměr šancí
- ➔ Znat princip metody maximální věrohodnosti
- ➔ Vědět, co nového nám ve srovnání s klasickým lineárním modelem mohou poskytnout zobecněné lineární modely
- ➔ Umět nadefinovat logistický model a popsat jeho užití
- ➔ Znat základní metody pro ověření předpokladů modelu

Logistický model

Poměr rizik a poměr šancí

Motivace

➔ Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS).

Výsledky dány v tabulce:

SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

➔ Jak rozhodnete o závislosti uvedených veličin?

Můžete nějak kvantifikovat sílu vztahu mezi veličinami?



Motivace

- ➔ Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS).

Výsledky dány v tabulce:

SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

- ➔ Pomocí Pearsonova chí-kvadrát nebo Fisherova exaktního testu můžeme rozhodovat o závislosti/nezávislosti dvou sledovaných veličin. Testy ale neumožňují tento vztah kvantifikovat.
- ➔ **Má-li to smysl a chceme-li kvantifikovat (rozhodovat o těsnosti této závislosti) můžeme použít tzv. **relativní riziko** a **poměr šancí**.**

Relativní riziko = Relative risk

- ➔ Výpočet relativního rizika (RR) umožňuje srovnat pravděpodobnosti výskytu sledovaného jevu ve dvou různých skupinách.
- ➔ 1. skupina – **experimentální nebo skupina s expozicí určitému faktoru**
- ➔ 2. skupina – **kontrolní nebo skupina bez expozice**

$$RR = \frac{\text{Pravděpodobnost výskytu jevu v 1. skupině (experimentální)}}{\text{Pravděpodobnost výskytu jevu ve 2. skupině (kontrolní)}} = \frac{P_1}{P_0}$$

Sledovaný jev	Skupina		Celkem
	Experimentální	Kontrolní	
Ano	a	b	$a + b$
Ne	c	d	$c + d$
Celkem	$a + c$	$b + d$	n



$$RR = \frac{P_1}{P_0} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}$$

Příklad – relativní riziko

→ Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS).

Výsledky dány v tabulce:

SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

$$RR = \frac{P_1}{P_0} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{\frac{29}{29+7301}}{\frac{15}{15+11241}} = 2,97$$



Riziko výskytu SIDS u dětí matek ve věku do 25 je téměř třikrát vyšší než u dětí matek rodičů ve vyšším věku.

Riziko vs. „šance“ (odds)

- ➔ **Riziko** – odhad pravděpodobnosti vzniku onemocnění
- ➔ **Relativní riziko** – poměr dvou pravděpodobností
- ➔ **Šance** – poměr pravděpodobnosti výskytu jevu a výskytu opačného jevu

$$odds = \frac{P_1}{1 - P_1}$$

- ➔ nabývá hodnot mezi 0 a nekonečnem
- ➔ pokud kůň vyhraje s pravděpodobností 10%, jaká je jeho **šance** na výhru?




Poměr šancí = Odds ratio

- ➔ Poměr šancí (OR) je další charakteristikou, která umožňuje srovnat výskyt sledovaného jevu ve dvou různých skupinách.
- ➔ 1. skupina – **experimentální nebo skupina s expozicí určitému faktoru**
- ➔ 2. skupina – **kontrolní nebo skupina bez expozice**

$$OR = \frac{\frac{\text{Pravděpodobnost výskytu jevu v 1. skupině (experimentální)}}{1 - \text{Pravděpodobnost výskytu jevu v 1. skupině (experimentální)}}}{\frac{\text{Pravděpodobnost výskytu jevu ve 2. skupině (kontrolní)}}{1 - \text{Pravděpodobnost výskytu jevu ve 2. skupině (kontrolní)}}} = \frac{O_1}{O_0} = \frac{\frac{P_1}{1 - P_1}}{\frac{P_0}{1 - P_0}}$$

Sledovaný jev	Skupina		Celkem
	Experimentální	Kontrolní	
Ano	a	b	$a + b$
Ne	c	d	$c + d$
Celkem	$a + c$	$b + d$	n



$$OR = \frac{\frac{P_1}{1 - P_1}}{\frac{P_0}{1 - P_0}} = \frac{\frac{a}{c}}{\frac{b}{d}}$$

Příklad – odds ratio

→ Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS).

Výsledky dány v tabulce:

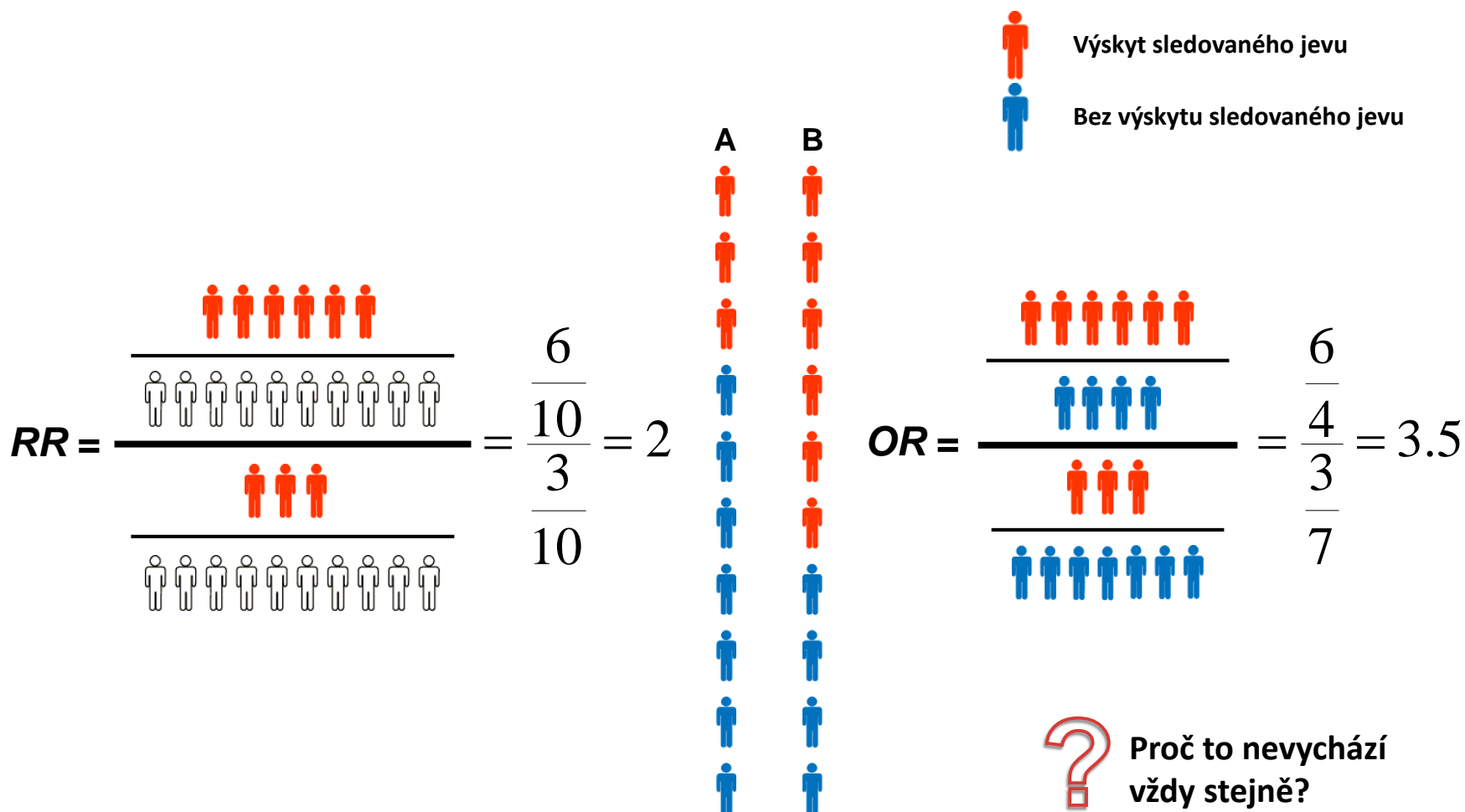
SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

$$OR = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{\frac{29}{7301}}{\frac{15}{11241}} = 2,98$$



„Šance“ na výskyt SIDS u dětí matek ve věku do 25 je téměř třikrát vyšší než u dětí matek rodičích ve vyšším věku.

Grafické srovnání *RR* a *OR*



Komentáře k RR, OR

- hodnota relativního rizika leží mezi 0 a $1/P_0$
- pro běžné jevy nelze pozorovat vysoké hodnoty relativního rizika pokud je riziko v kontrolní skupině 66%, maximální RR je 1,5
- OR je obtížnější interpretovat
- může být vhodné konvertovat na RR, musíme ale znát riziko v kontrolní skupině

$$RR = \frac{OR}{1 - P_0(1 - OR)} \quad OR = \frac{RR(1 - P_0)}{1 - P_0RR}$$

- nevychází stejně, ale oba jsou validní ukazatele účinku

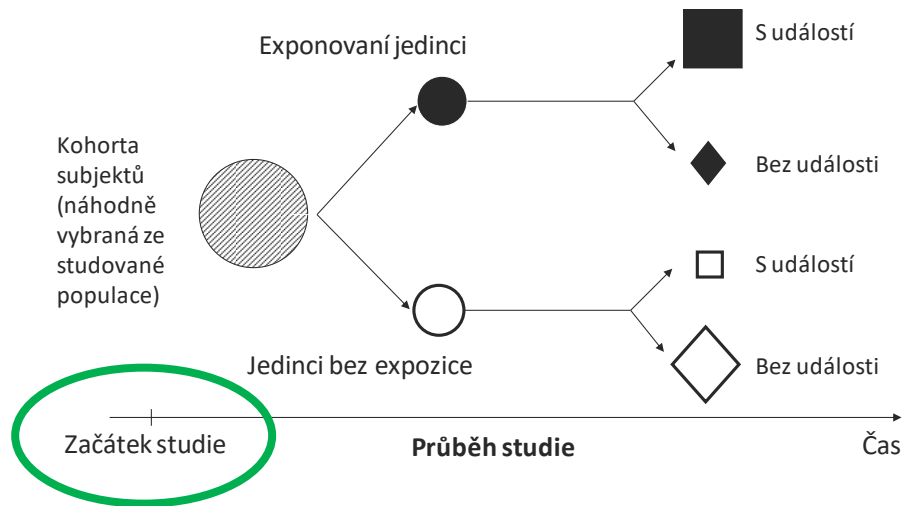


kdy spolu obě veličiny RR a OR splývají?

Základní typy epidemiologických studií

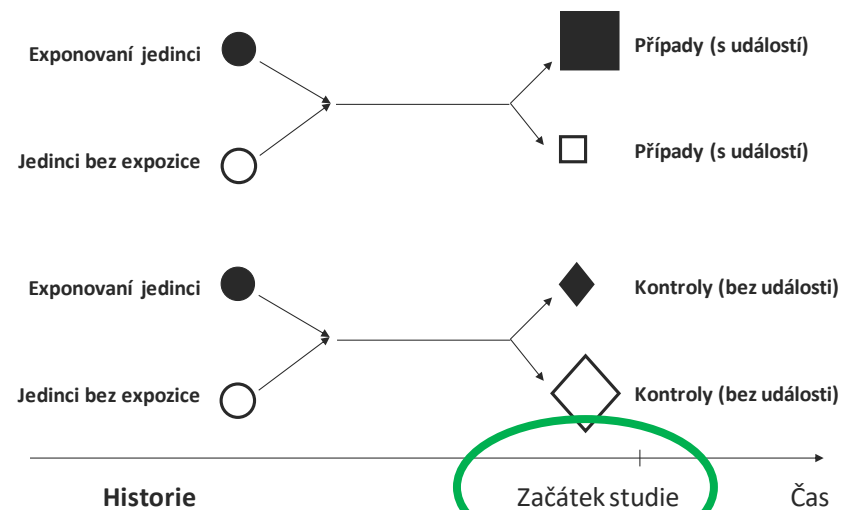
➔ Kohortová studie

➔ U některých subjektů je rizikový faktor přítomen a u jiných ne → sledujeme v čase, zda se vyskytne událost.



➔ Studie případů a kontrol

➔ U některých subjektů se událost vyskytla a u jiných ne → zpětně hodnotíme, zda se liší s ohledem na nějaký rizikový faktor.



Použití *RR* a *OR*

- ➔ **Kohortová studie** – u některých subjektů je rizikový faktor přítomen a u jiných ne → sledujeme, zda se vyskytne událost.
- ➔ Zjištěná pravděpodobnost výskytu události v kontrolní skupině je reprezentativní, neboť prospektivně zařazujeme všechny pacienty
→ **korektní použití *RR*.**

- ➔ **Studie případů a kontrol** – u některých subjektů se událost vyskytla a u jiných ne → zpětně hodnotíme, zda se liší s ohledem na nějaký rizikový faktor.
- ➔ Zjištěná pravděpodobnost výskytu události v kontrolní skupině není reprezentativní, neboť ji ovlivňujeme zpětným výběrem skupin subjektů.
→ **nekorektní použití *RR*.**
→ **korektní použití *OR*.**

Poměr šancí – další příklad

Protektivní účinek hormonální antikoncepce na riziko zhoubného nádoru vaječníku

TABLE 2. Oral contraceptive use characteristics, including estrogen and progestin dose, among ovarian cancer cases and controls, Delaware Valley area, May 1994 to June 1998

Variable	Cases (n = 767)	Controls (n = 1,367)	Crude OR*	95% CI*	Adjusted OR†	95% CI
Oral contraceptive use						
Never	341	426	1.0		1.0	
Ever	426	940	0.6	0.5, 0.7	0.6	0.5, 0.8

Ovarian cancer	OC user		Total
	Yes	No	
Yes	426	341	767
No	940	426	1366
Total	1366	767	2133

$$OR = \frac{\frac{P_1}{1 - P_1}}{\frac{P_0}{1 - P_0}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{\frac{426}{940}}{\frac{341}{426}} = 0,6$$

„Šance“ na zhoubný nádor ovarií se snižuje o 40%

Zdroj: Risk of Ovarian Cancer in Relation to Estrogen and Progestin Dose and Use Characteristics of Oral Contraceptives, Ness a kol. 2000

Logistický model

Metoda maximální věrohodnosti

Metoda maximální věrohodnosti

- Autorem je R. A. Fisher (1922). Anglicky „maximum likelihood estimation“.
- Máme n nezávislých stejně rozdělených pozorování (i.i.d.) z rozdělení s hustotou $f(x; \theta)$.
- **Sdružená hustota** odpovídající n pozorovaným hodnotám x_1, x_2, \dots, x_n je:

$$f(x_1, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i; \theta)$$

- Sdružená hustota vyjadřuje (za předpokladu, že známe θ), jak moc je pravděpodobné, že pozorované hodnoty pochází z rozdělení s hustotou $f(x; \theta)$
- **Pointa metody maximální věrohodnosti:** Dívat se na sdruženou hustotu jako na funkci θ a vybrat θ takové, aby výraz $f(x_1, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i; \theta)$ byl co největší (maximum).

Věrohodnostní funkce

→ Zavádíme tzv. **věrohodnostní funkci** („likelihood function“):

$$L(\theta | x_1, \dots, x_n) = f(x_1, \dots, x_n | \theta)$$

→ Maximálně věrohodný odhad, značíme ho $\hat{\theta}_{MLE}$, je číslo, které maximalizuje věrohodnostní funkci, tedy

$$\hat{\theta}_{MLE} = \arg \max_{\theta \in \Theta} L(\theta | x_1, \dots, x_n)$$

→ Výpočetně se jedná o řešení rovnice (rovníc):

$$dL(\theta | x_1, \dots, x_n) / d\theta = 0$$

→ Musíme si ještě ověřit, že se jedná o maximum – např. pomocí druhých derivací.

Logaritmus věrohodnostní funkce

→ Často je výhodnější (hlavně výpočetně jednodušší) maximalizovat logaritmus věrohodnostní funkce:

$$l(\theta | x_1, \dots, x_n) = \ln L(\theta | x_1, \dots, x_n) =$$



Logaritmus věrohodnostní funkce

→ Často je výhodnější (hlavně výpočetně jednodušší) maximalizovat logaritmus věrohodnostní funkce:

$$l(\theta | x_1, \dots, x_n) = \ln L(\theta | x_1, \dots, x_n) = \ln \prod_{i=1}^n f(x_i; \theta) = \sum_{i=1}^n \ln f(x_i; \theta)$$

Příklad

ML odhad parametru μ normálního rozdělení

→ Máme n i.i.d. pozorování z normálního rozdělení: x_1, x_2, \dots, x_n .

→ Sdružená hustota má tvar:

$$f(x_1, \dots, x_n | \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - \mu)^2 / 2\sigma^2}$$

→ Logaritmus věrohodnostní funkce má tvar:

$$\ln L(\mu, \sigma^2 | x_1, \dots, x_n) = ?$$

→ Parciální derivace logaritmu věrohodnostní funkce mají tvar:

$$\partial \ln L / \partial \mu = ?$$

$$\partial \ln L / \partial \sigma^2 = ?$$

Příklad

ML odhad parametru μ normálního rozdělení

→ Máme n i.i.d. pozorování z normálního rozdělení: x_1, x_2, \dots, x_n .

→ Sdružená hustota má tvar:

$$f(x_1, \dots, x_n | \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - \mu)^2 / 2\sigma^2}$$

→ Logaritmus věrohodnostní funkce má tvar:

$$\ln L(\mu, \sigma^2 | x_1, \dots, x_n) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

→ Parciální derivace logaritmu věrohodnostní funkce mají tvar:

$$\partial \ln L / \partial \mu = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0$$

$$\partial \ln L / \partial \sigma^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

Příklad

ML odhad parametru μ normálního rozdělení

→ Výsledkem jsou následující odhady:

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

$$\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Logistický model

Zobecněný lineární model

Zobecněný lineární model

- zobecnění mnoha existujících statistických modelů
- John A. Nelder (1924-2010), Robert W.M. Wedderburn (1947-1975)

J. R. Statist. Soc. A,
(1972), **135**, Part 3, p. 370

370

Generalized Linear Models

By J. A. NELDER and R. W. M. WEDDERBURN

Rothamsted Experimental Station, Harpenden, Herts

SUMMARY

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These generalized linear models are illustrated by examples relating to four distributions; the Normal, Binomial (probit analysis, etc.), Poisson (contingency tables) and gamma (variance components).

The implications of the approach in designing statistics courses are discussed.

Zobecněný lineární model

→ zobecnění mnoha existujících statistických modelů

→ John A. Nelder (1924-2010), Robert W.M. Wedderburn (1947-1975)

Klasický model nezávislé normální

1. Náhodná část $EY_i = \mu_i \quad DY_i = \sigma^2$ Předpoklad rozložení

2. Systematická část $\eta_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij}$ Lineární prediktor

3. Spojení mezi náhodnou a systematickou částí $\mu_i = \eta_i$ Linkovací funkce

Zobecněný lineární model

- zobecnění mnoha existujících statistických modelů
- John A. Nelder (1924-2010), Robert W.M. Wedderburn (1947-1975)

Klasický model nezávislé normální

- Náhodná část** $EY_i = \mu_i \quad DY_i = \sigma^2$ **Předpoklad rozložení**
Třída exponenciálních rozdělení
- Systematická část** $\eta_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij}$ **Lineární prediktor**
- Spojení mezi náhodnou a systematickou částí** $\mu_i = \eta_i$ **Linkovací funkce**
Vhodné funkce

ODHAD PARAMETRŮ METODOU MAXIMÁLNÍ VĚROHODNOSTI

Čemu se budeme věnovat?

→ Logistická regrese

→ binomické (alternativní) rozdělení výsledku, linkovací funkce $\text{logit}(\cdot)$

→ Poissonova regrese

→ Poissonovo rozdělení výsledku, linkovací funkce $\ln(\cdot)$

→ později o zobecněných lineárních modelech

→ viz předmět **M7222 Zobecněné lineární modely**

Logistický model

Logistická regrese

Binomické rozdělení

→ Diskrétní rozdělení, které **popisuje počet výskytů sledované události** (ve formě nastala/nenastala) **v sérii n nezávislých experimentů**, kdy v každém experimentu **je stejná pravděpodobnost výskytu události** a je $p = \pi$.

→ Pravděpodobnostní funkce:

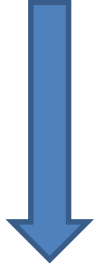
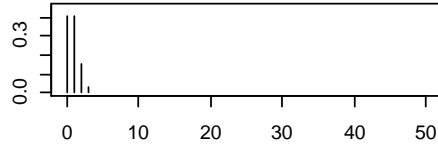
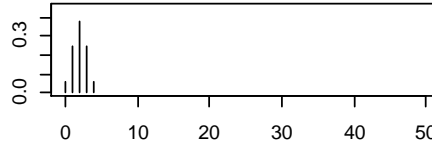
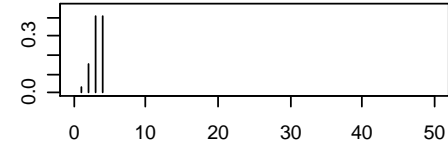
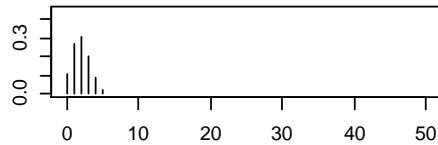
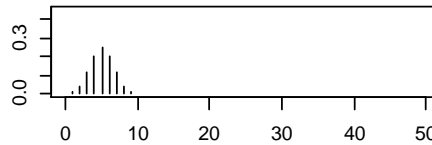
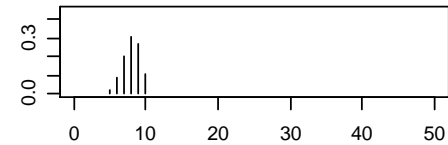
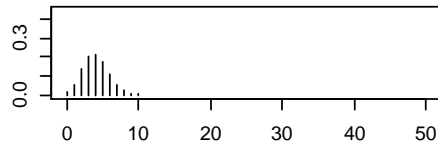
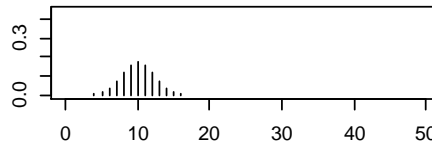
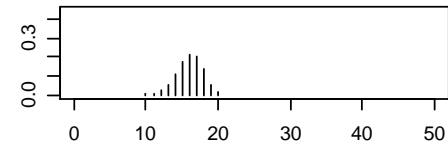
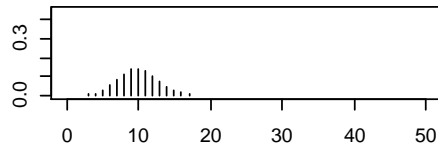
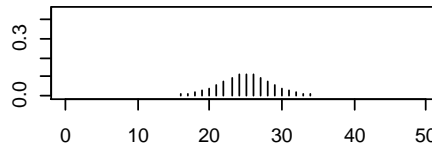
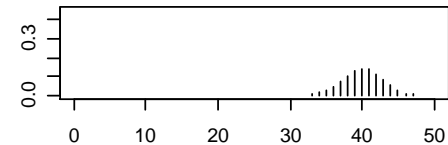
$$P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

→ Střední hodnota

$$E(X) = n\pi$$

→ Rozptyl

$$D(X) = n\pi(1 - \pi)$$

π  n  $n = 4 \quad \pi = 0.2$  $n = 4 \quad \pi = 0.5$  $n = 4 \quad \pi = 0.8$  $n = 10 \quad \pi = 0.2$  $n = 10 \quad \pi = 0.5$  $n = 10 \quad \pi = 0.8$  $n = 20 \quad \pi = 0.2$  $n = 20 \quad \pi = 0.5$  $n = 20 \quad \pi = 0.8$  $n = 50 \quad \pi = 0.2$  $n = 50 \quad \pi = 0.5$  $n = 50 \quad \pi = 0.8$ 

Formulace logistického modelu

- Uvažujeme binární výsledek, který chceme vztáhnout ke známým vysvětlujícím proměnným – modelujeme pomocí alternativního (binomického) rozdělení

$$Y_i \sim A(\pi_i)$$

$$i = 1, \dots, n$$

Formulace logistického modelu

Normální lineární regresní model:

$$EY_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$
$$i = 1, \dots, n$$

Logistický regresní model – modelujeme **pravděpodobnost události pro i-tý subjekt**:

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

linkovací funkce

lineární prediktor (označujeme písmenem η – „éta“)

Linkovací funkce

$$\text{logit}(p) = \ln \frac{p}{1-p}$$

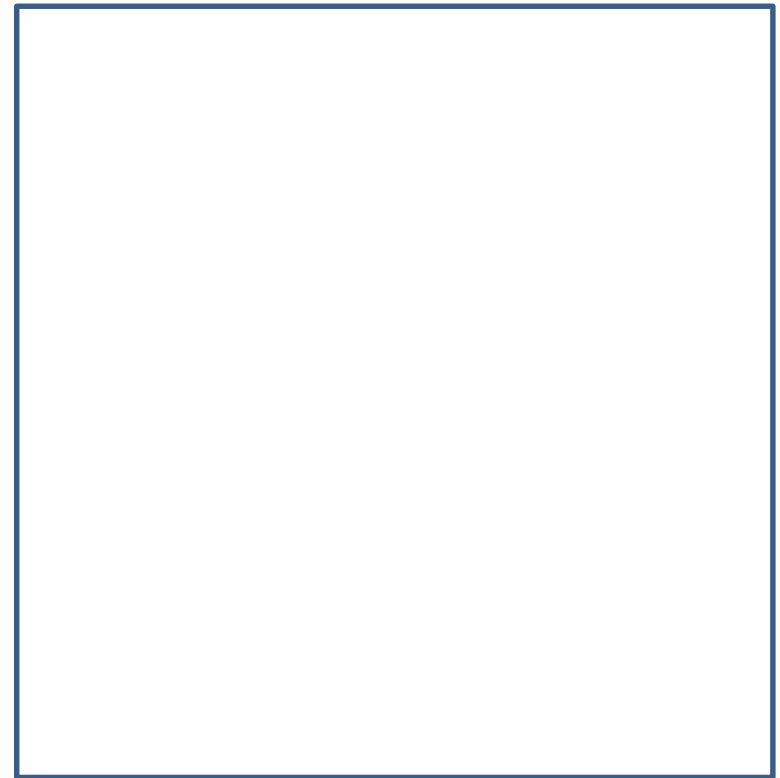
$$\text{logit}^{-1}(\eta) = \text{expit}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

logit(p)



p

expit(η)

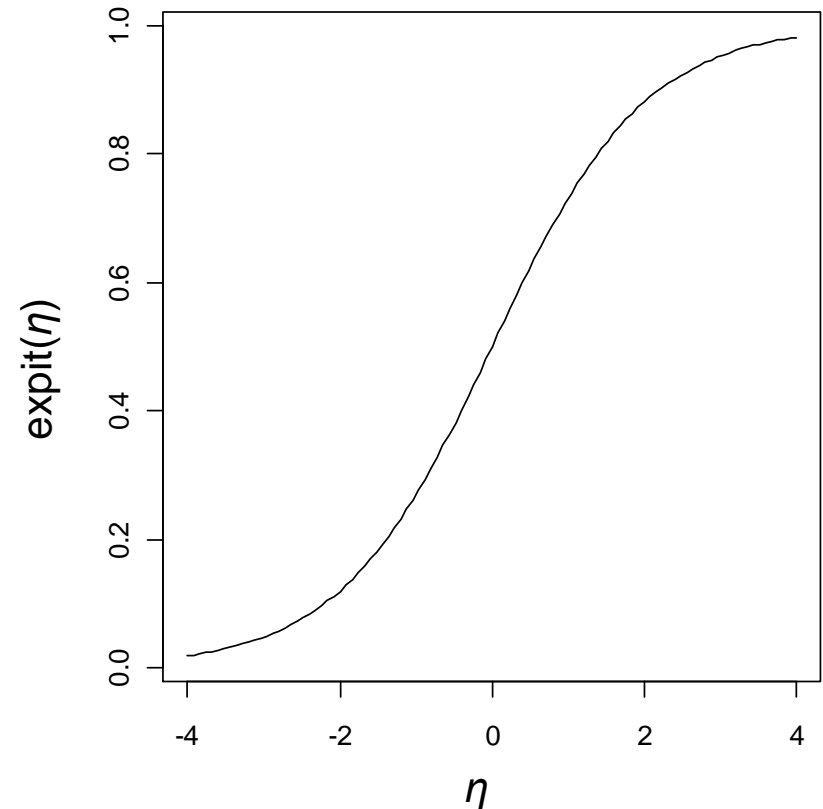
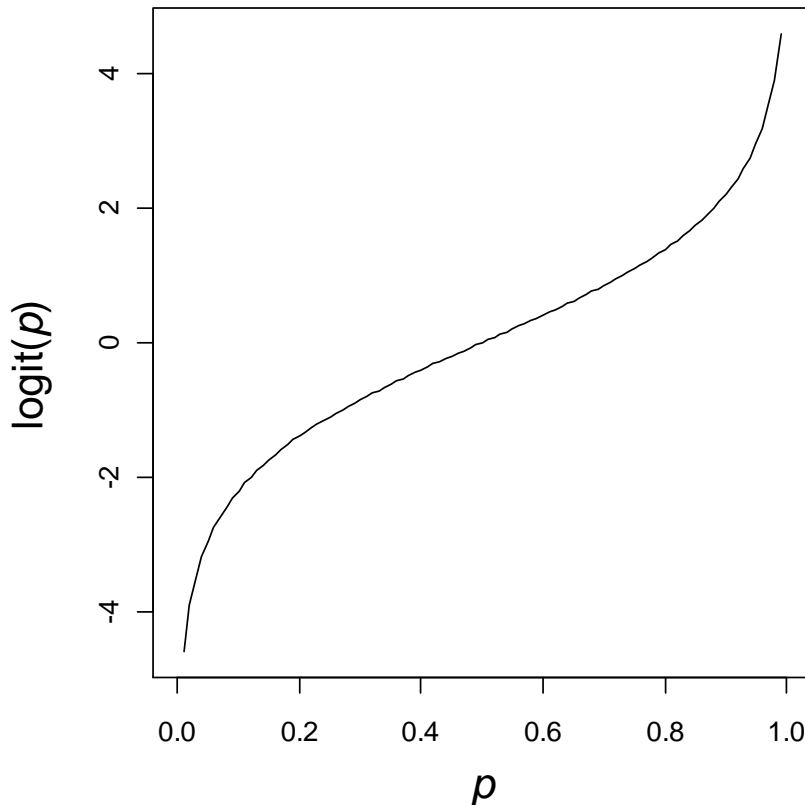


η

Linkovací funkce

$$\text{logit}(p) = \ln \frac{p}{1-p}$$

$$\text{logit}^{-1}(\eta) = \text{expit}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$



Interpretace koeficientů

Subjekt 1:

$$\text{logit}(p_1) = \beta_0$$

$$\ln \frac{p_1}{1-p_1} = \beta_0$$

$$\frac{p_1}{1-p_1} = \exp(\beta_0)$$

Subjekt 2:

$$\text{logit}(p_2) = \beta_0 + \beta_1$$

$$\ln \frac{p_2}{1-p_2} = \beta_0 + \beta_1$$

$$\frac{p_2}{1-p_2} = \exp(\beta_0 + \beta_1)$$

**Parametr
asociovaný
s nějakým
binárním
prediktorem**

Odds ratio (poměr šancí) na nějakou událost:

$$OR(2,1) = \frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \frac{\exp(\beta_0) \exp(\beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

Exp(odhad parametru) PŘEDSTAVUJE ODDS RATIO SPOJENÉ S DANÝM PREDIKTOREM

Interpretace koeficientů – příklad 1

Rizikové faktory pro trombózu spojenou s chemoterapií

Table 2. Predictors of venous thromboembolism in the derivation cohort by multivariate logistic regression analysis

Patient characteristic	β	Odds ratio* (95% CI)
Site of cancer		
Very high risk (stomach, pancreas)	1.46	4.3 (1.2-15.6)
High risk (lung, lymphoma, gynecologic, genitourinary excluding prostate)	0.43	1.5 (0.9-2.7)
Low risk (breast, colorectal, head and neck)	0.0	1.0 (reference)
Prechemotherapy platelet count $350 \times 10^9/L$ or more	0.60	1.8 (1.1-3.2)
Hemoglobin level less than 100 g/L or use of red cell growth factors	0.89	2.4 (1.4-4.2)
Prechemotherapy leukocyte count more than $11 \times 10^9/L$	0.77	2.2 (1.2-4)
BMI 35 kg/m^2 or more	0.90	2.5 (1.3-4.7)

*Odds ratios are adjusted for stage.

Zdroj: Development and validation of a predictive model for chemotherapy-associated thrombosis, Khorana a kol. 2008

Interpretace koeficientů – příklad 2

Rizikové faktory pro pooperační plicní komplikace

Table 6

Results from the multiple logistic regression analysis, relating the probability of postoperative pulmonary complications to various covariates.

Covariate	χ^2 (df)	P	Odds-ratio (CI)
Type of operation	33.7 (2)	<0.0001	
Abdominal vs. orthopaedic			10.18 (3.69–28.12)
Gynaecological vs. orthopaedic			1.95 (0.51–7.39)
Special risk group			
Pancuronium and a TOF-ratio below 0.7	11.9 (1)	0.0006	5.08 (2.12–12.15)
Duration of anaesthesia			
dichotomized, above vs. below 200 min	10.4 (1)	0.0013	3.32 (1.60–6.88)
Age			
linear effect for each 10 years	19.7 (1)	<0.0001	1.63 (1.29–2.06)

Zdroj: Residual neuromuscular block is a risk factor for postoperative pulmonary complications, Berg a kol. 1997

Logistický model

Ověření správnosti modelu

Správnost modelu

Pozorované hodnoty výsledku $\mathbf{y}, \mathbf{y}' = (y_1, y_2, \dots, y_n)$

Predikované hodnoty výsledku $\hat{\mathbf{y}}, \hat{\mathbf{y}}' = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$

- 1. Pozorované a predikované hodnoty od sebe nejsou příliš vzdáleny**
 - celková shoda
- 2. Žádné z pozorování k celkové vzdálenosti extrémně nepřispívá**
 - individuální komponenty celkové statistiky, další metody
 - analýza reziduí

Celková shoda

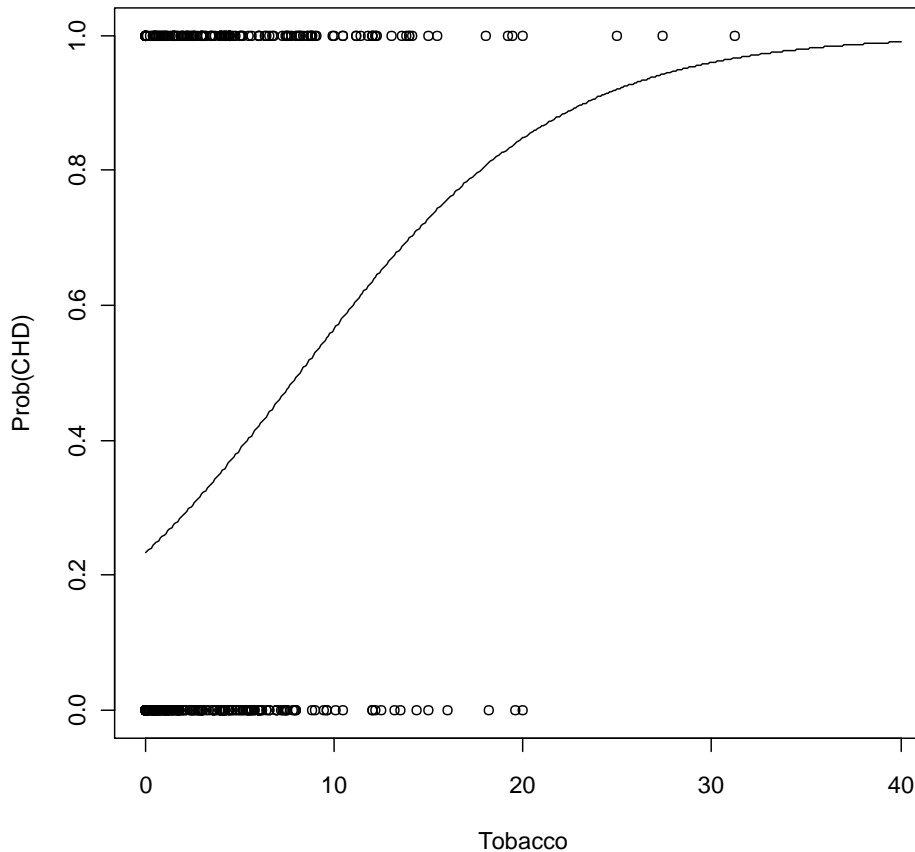
(overall goodness of fit)

➔ Nulová hypotéza:

Skutečná regresní funkce je rovna modelové funkci

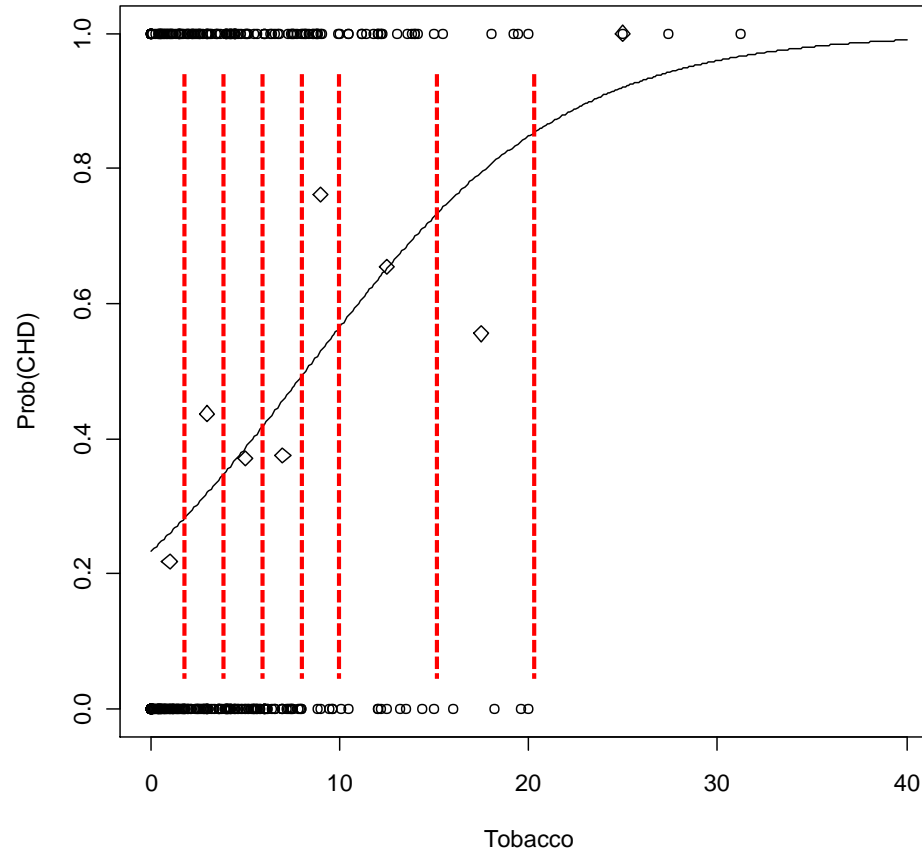
Predikce pravděpodobnosti události

$$p_i = \text{expit}(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})$$



Predikce pravděpodobnosti události

Je vhodné srovnávat se seskupenými pozorováními:



To lze popsat formálně a statisticky testovat...

Ověření splnění předpokladů analýza reziduí: Pearsonova rezidua

➔ Pearsonova rezidua

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}$$

Ověření splnění předpokladů analýza reziduí: Devianční rezidua

→ Devianční rezidua

$$d_i = -\sqrt{2|\ln(1 - \hat{\pi}_j)|} \quad \text{pro } y_j = 0$$

$$d_i = \sqrt{2|\ln(\hat{\pi}_j)|} \quad \text{pro } y_j = 1$$

Ověření splnění předpokladů

Další běžné diagnostiky

→ Leverage (pákové body)

- určení potenciální vlivnosti daného bodu (pozorování, které je daleko od ostatních a může tak ovlivnit výsledek regrese)
- závisí na váze pozorování a vzdálenosti pozorování prediktoru od průměru
- váha pozorování s predikovanou hodnotou pravděpodobnosti blízkou 0 nebo 1 je nízká

→ Cookova vzdálenost

- shrnuje informaci z reziduí a z leverage
- deleční diagnostika – určení vlivu daného pozorování na výsledek
- ukazuje, nakolik přítomnost daného pozorování ovlivňuje odhady koeficientů

Logistický model

Závěr

Co byste měli vědět a umět po dnešní hodině ?

- ➔ Znat užitečné veličiny pro měření vztahu/účinku: poměr rizik a poměr šancí
- ➔ Znat princip metody maximální věrohodnosti
- ➔ Vědět, co nového nám ve srovnání s klasickým lineárním modelem mohou poskytnout zobecněné lineární modely
- ➔ Umět nadefinovat logistický model a popsat jeho užití
- ➔ Znat základní metody pro ověření předpokladů modelu