Bi7491 Regresní modelování

Dodatky ke zobecněným lineárním modelům

Co byste měli vědět a umět po dnešní hodině ?

- Chápat princip analýzy deviance
- Umět nadefinovat Poissonův model a popsat jeho užití
- Umět vysvětlit pojem overdispersion – čím je způsobena a jak ji poznat a řešit
- Znát základní možnosti modelování ordinálních výsledků

Lékařská fakulta Masarykovy univerzity
Institut biostatistiky a analýz

Dodatky ke zobecněným lineárním modelům

Analýza deviance
ve zobecněných lineárních modelech

Modely a submodely

- Modelování - **y** nahrazujeme $\hat{\boldsymbol{\mu}}$ prostřednictvím odhadu $\hat{\boldsymbol{\beta}}$
- Jak moc se vzájemně liší?

- Model s *n* parametry
  MAXIMÁLNÍ MODEL (plný, saturovaný)
  → veškerá variabilita do systematické složky
- Model s *k* parametry
  ZKOUMANÝ MODEL
- když vyloučíme některý prediktor (*m* < *k* parametrů)
  SUBMODEL
- Model s 1 parametrem (konstantou – průměrem)
  NULOVÝ MODEL → veškerá variabilita do náhodné složky

vždy stejný typ rozdělení, stejná linkovací funkce

Lékařská fakulta Masarykovy univerzity
Institut biostatistiky a analýz

Deviance

- představuje odchylku zkoumaného modelu od „dokonalého" maximálního modelu

$$D = 2[\underbrace{l(\mathbf{y};\mathbf{y})}_{\text{log-věrohodnost maximálního modelu}} - \underbrace{l(\hat{\boldsymbol{\mu}};\mathbf{y})}_{\text{log-věrohodnost zkoumaného modelu}}]$$

- analogie s analýzou rozptylu – zde formulovaná pomocí změny ve věrohodnosti
- umožňuje test odchylky od maximálního modelu

Testování submodelů

$$\Delta D = 2[l(\hat{\boldsymbol{\mu}};\mathbf{y}) - l(\hat{\boldsymbol{\mu}}_{SUB};\mathbf{y})]$$

rozdíl deviancí **log-věrohodnost zkoumaného modelu** **log-věrohodnost submodelu**

- Deviance je velmi užitečná při srovnání dvou modelů z nichž jeden je podmodelem (submodelen) druhého
- Je-li **Δ*D* > $\chi^2_{1-\alpha}$(*k*-*m*)**, kde *m* (*k*) je počet odhadovaných parametrů submodelu (zkoumaného modelu), pak je submodel nevhodný – přehnaně zjednodušující

Test významnosti celého modelu

vs. maximální model

- srovnání maximálního (plného) modelu se **zkoumaným modelem – REZIDUÁLNÍ DEVIANCE** (odpovídá reziduálnímu součtu čtverců)
- ***Nechybí nám nějaký významný efekt?***

$$D > \chi^2_{1-\alpha}(\text{počet pozorování – počet parametrů})$$

NĚCO V MODELU CHYBÍ...

Software uvádí příslušnou statistiku

Je ale asymptotická – slouží spíš pro orientační kontrolu !!!

Test významnosti celého modelu

vs. nulový model

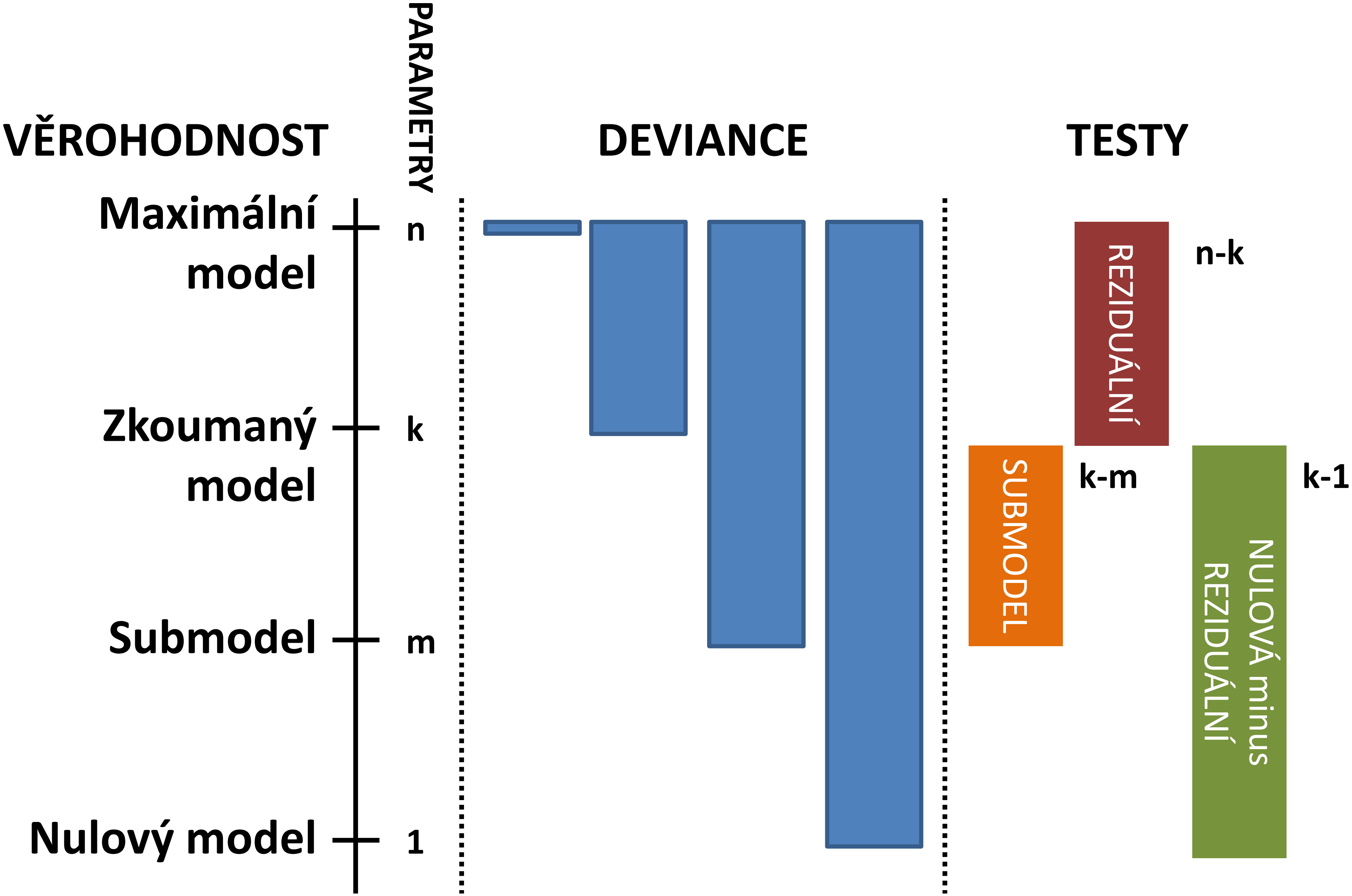
- srovnání **zkoumaného modelu** s **nulovým modelem**
  NULOVÁ DEVIANCE – REZIDUÁLNÍ DEVIANCE
- ***Vysvětluje vůbec zkoumaný model nějakou informaci?***

$$\Delta D > \chi^2_{1-\alpha}(\text{počet parametrů} - 1)$$

⇒ **MODEL NĚCO VYSVĚTLUJE**

Software uvádí příslušnou statistiku

Je ale asymptotická – slouží spíš pro orientační kontrolu !!!


VĚROHODNOST
PARAMETRY
DEVIANCE
TESTY
Maximální model
n
Zkoumaný model
k
Submodel
m
Nulový model
1
REZIDUÁLNÍ
n-k
SUBMODEL
k-m
NULOVÁ minus REZIDUÁLNÍ
k-1

Akaikeovo informační kritérium

(Akaike information criterion, AIC)

$$AIC = -2l(\hat{\boldsymbol{\mu}};\mathbf{y}) + 2k$$

AIC= – 2 maximum logaritmované věrohodnosti + 2 počet parametrů modelu

- Čím je hodnota AIC **menší**, tím je model lepší.
- AIC **penalizuje modely s velkým počtem parametrů**
- užití brání „přeučení“ modelu (takový model by dobře neodpovídal novému vzorku)

716 IEEE TRANSACTIONS ON AUTOMATIC CONTROL, VOL. AC-19, NO. 6, DECEMBER 1974

A New Look at the Statistical Model Identification

HIROTUGU AKAIKE, MEMBER, IEEE

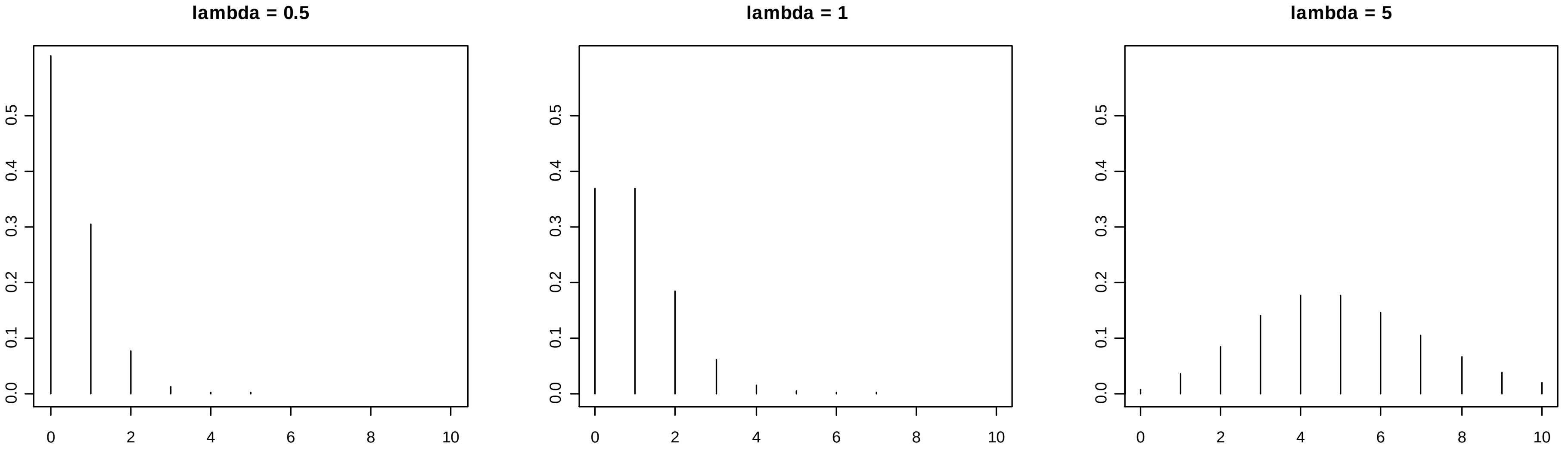
Dodatky ke zobecněným lineárním modelům

Poissonova regrese

Poissonovo rozdělení

- Diskrétní rozdělení, které **popisuje počet výskytů sledované události na danou jednotku** (času, plochy, objemu), když se tyto události vyskytují vzájemně **nezávisle** s konstantní intenzitou (**jediný** parametr λ).
- Jedná se o zobecnění binomického rozdělení pro $n \to \infty$ a $p \to 0$.
- Pravděpodobnostní funkce: $P(X = x) = p_X(x;\lambda) = \dfrac{\lambda^x e^{-\lambda}}{x!}, x \geq 0$
- Střední hodnota, rozptyl: $EX = \lambda, DX = \lambda$
- **Příklady**: průměrný výskyt mutací bakterií na 1 Petriho misku, počet krvinek v poli mikroskopu, počet žížal vyskytujících se na 1 m^2, počet pooperačních komplikací během určitého časového intervalu po výkonu.

Lékařská fakulta Masarykovy univerzity
Institut biostatistiky a analýz



lambda = 0.5
lambda = 1
lambda = 5
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10

Formulace Poissonova modelu

- Uvažujeme výsledek vyjádřený počtem (událostí, objektů), který chceme vztáhnout ke známým vysvětlujícím proměnným – modelujeme pomocí Poissonova rozdělení

$$Y_i \sim Po(\lambda_i)$$

$$i = 1, \ldots, n$$

Formulace Poissonova modelu

Normální lineární regresní model:

$$EY_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

$$i = 1, \ldots, n$$

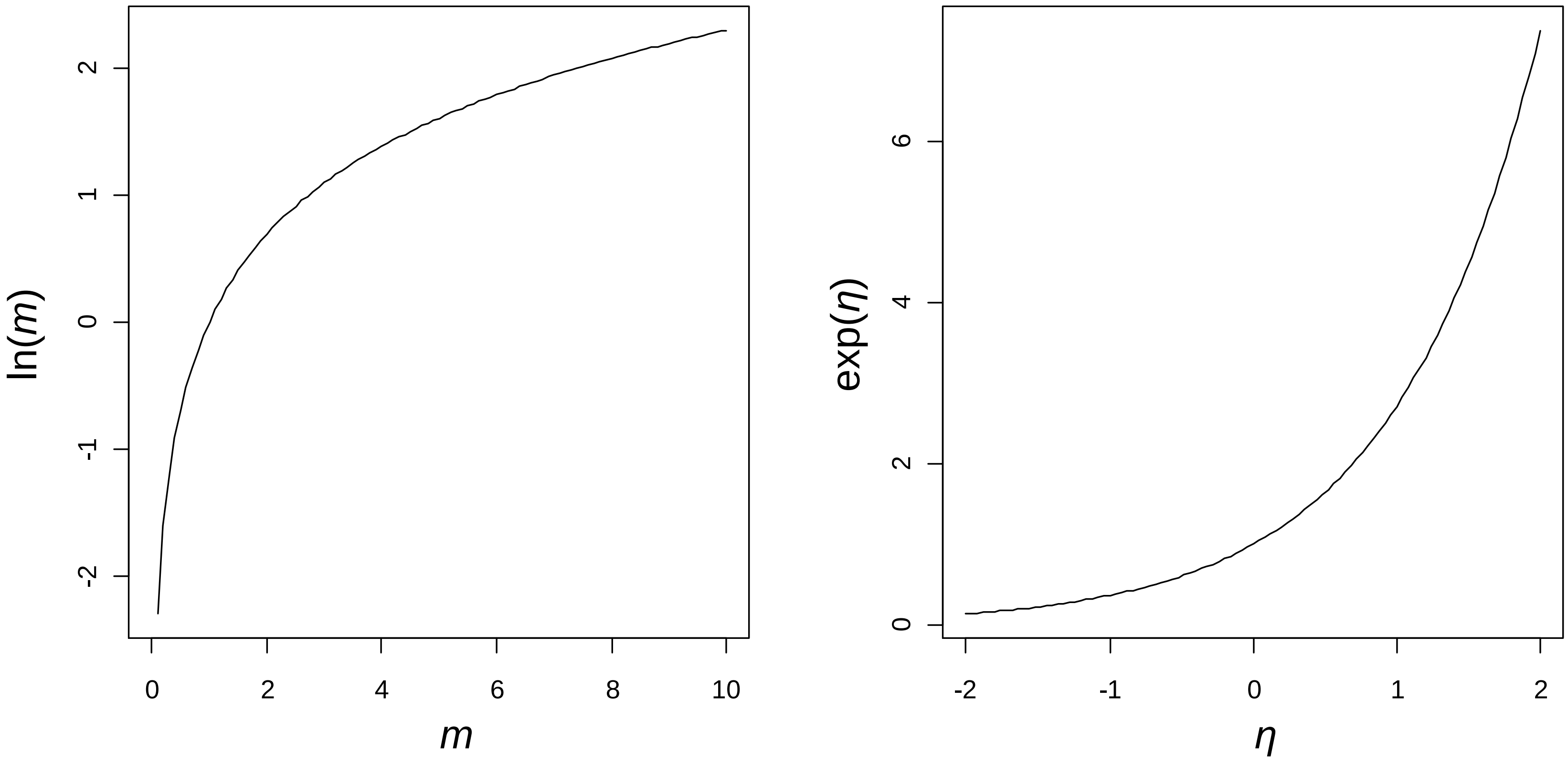
Poissonův regresní model – modelujeme **očekávaný počet událostí:**

$$\boxed{\ln(m_i)} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

$$i = 1, \ldots, n$$

Lékařská fakulta Masarykovy univerzity
Institut biostatistiky a analýz

Linkovací funkce



Interpretace koeficientů - příklad

Subjekt 1:

$$\ln(m_1) = \beta_0$$
$$m_1 = \exp(\beta_0)$$

Subjekt 2:

$$\ln(m_2) = \beta_0 + \beta_1$$
$$m_2 = \exp(\beta_0 + \beta_1)$$

Parametr asociovaný s nějakým binárním prediktorem

Risk ratio (relativní riziko) nějaké události:

$$RR(2,1) = \frac{m_2}{m_1} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \frac{\exp(\beta_0)\exp(\beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

Exp(odhad parametru) PŘEDSTAVUJE RELATIVNÍ RIZIKO SPOJENÉ S DANÝM PREDIKTOREM

Model incidence (míry)

$$\textbf{Incidence} = \frac{\textbf{počet nových případů}}{\textbf{součet „osoboroků" v riziku}}$$

- Popsaný model lze využít pro modelování incidence onemocnění (výskytu událostí apod.)
- Nezbytné, pokud se pro jednotlivá pozorování liší např. doba sledování
- Do modelu je nezbytné uvést jmenovatele – součet osoboroků v riziku (person-years at risk), označ. d_i
- V rámci softwarových nástrojů se specifikuje jako tzv. offset:

$$\ln\left(\frac{m_i}{d_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, i = 1, \ldots, n$$

$$\ln(m_i) - \ln(d_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, i = 1, \ldots, n$$

$$\ln(m_i) = \ln(d_i) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, i = 1, \ldots, n$$

Model incidence (míry)

$$\textbf{Incidence} = \frac{\textbf{počet nových případů}}{\textbf{součet „osoboroků" v riziku}}$$

- Popsaný model lze využít pro modelování incidence onemocnění (výskytu událostí apod.)
- Nezbytné, pokud se pro jednotlivá pozorování liší např. doba sledování
- Do modelu je nezbytné uvést jmenovatele – součet osoboroků v riziku (person-years at risk), označ. d_i
- V rámci softwarových nástrojů se specifikuje jako tzv. offset
- **Interpretace exp(β) – poměr incidencí**

Lékařská fakulta Masarykovy univerzity
Institut biostatistiky a analýz

Ověření splnění předpokladů

1. **Linkovací funkce** – ln(.)
2. **Správnost lineárního prediktoru** – netřeba přidávat další proměnné, transformovat proměnné, nebo přidat interakce mezi proměnnými
3. **Správnost předpokládaného rozptylu výsledků** – dáno vzorcem pro Poissonovo rozdělení

Obdobně jako u logistické regrese

- analýza reziduí a vlivu
- analýza deviance

Lékařská fakulta Masarykovy univerzity
Institut biostatistiky a analýz

Dodatky ke zobecněným lineárním modelům

„Nadměrný rozptyl“ - *overdispersion*

Probíraná rozdělení v GLM

- prozatím jsme se věnovali logistické a poissonově regresi...

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* = θ**.

- Pravděpodobnostní funkce:

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

- Střední hodnota

$$\boxed{E(X) = n\theta}$$

- Rozptyl

$$\boxed{D(X) = n\theta(1-\theta)}$$

Poissonovo rozdělení

- Diskrétní rozdělení, které **popisuje počet výskytů sledované události na danou jednotku** (času, plochy, objemu), když se tyto události vyskytují vzájemně **nezávisle** s konstantní intenzitou (**jediný** parametr λ).
- Jedná se o zobecnění binomického rozdělení pro $n \to \infty$ a $p \to 0$.
- Pravděpodobnostní funkce: $P(X = x) = p_X(x;\lambda) = \dfrac{\lambda^x e^{-\lambda}}{x!}, x \geq 0$
- Střední hodnota, rozptyl: $\boxed{EX = \lambda,\ DX = \lambda}$
- **Příklady**: průměrný výskyt mutací bakterií na 1 Petriho misku, počet krvinek v poli mikroskopu, počet žížal vyskytujících se na 1 m², počet pooperačních komplikací během určitého časového intervalu po výkonu.

Střední hodnota a rozptyl

- prozatím jsme se věnovali logistické a poissonově regresi...
- v těchto rozděleních jsou spjaty střední hodnota a rozptyl:
  - v Poissonově rozdělení platí

    je li střední hodnota 1,5, je rozptyl rovněž 1,5

    (návštěv na urgentním příjmu za hodinu, moučných červů v dl mouky,...)
  - v Binomickém rozdělení platí

    je li střední hodnota 1,5, je rozptyl 0,75

    (v situaci, kdy např. odhadujeme počet chlapců mezi třemi potomky)

Overdispersion v praxi

- v praxi rozdělení výsledků nemusí přesně odpovídat předpokladům

- **DŮVOD**
  - výsledky nejsou vzájemně zcela nezávislé
    (více měření u jednoho pacienta/lékaře/laboratoře, autokorelace
    v časových řadách, ...)
  - naše naměřené a zkoumané prediktory kompletně nespecifikují výsledek

- **INDIKACE**
  - velmi vysoká reziduální variabilita (vysoká významnost testu)

- **ŘEŠENÍ**
  - přidat více prediktorů (pokud ale ten důležitý byl změřen)
  - odhadnout a využít zvlášť disperzní parametr
    `family=quasibinomial / quasipoisson`

Dodatky ke zobecněným lineárním modelům

Multinomiální modely

Ordinální výsledek

- kategorie lze seřadit, ale jen obtížně k nim lze přiřadit číselnou hodnotu (např. stadium choroby)

- opět vycházíme z lineárního prediktoru

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

- modelovat binomicky? modelovat nějaké skóre? ani jedno nemusí být vhodné

Ordinální výsledek

- Příklad: modelujeme stadium fibrózy jater (Y_i = 0,1,2,3) pomocí tří krevních markerů
- kumulativní pravděpodobnosti (odshora) jednotlivých stadií:

$q_{i,3} = p_{i,3}$	pravděpodobnost kategorie 3	cut-off mezi 2 a 3
$q_{i,2} = p_{i,2} + p_{i,3}$	pravděpodobnost kategorie 2 a více	cut-off mezi 1 a 2
$q_{i,1} = p_{i,1} + p_{i,2} + p_{i,3}$	pravděpodobnost kategorie 1 a více	cut-off mezi 0 a 1

- kdybychom použili logistickou regresi pro spojené kategorie 2 a více

$$\eta_i = \mathrm{logit}(q_{i,2}) = \ln\left(\frac{q_{i,2}}{1-q_{i,2}}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

Model proporcionálních šancí

Předpoklad: Vliv proměnné nezávisí na volbě cut-off (!)

- Model proporcionálních **šancí** pro kumulativní logit

  j = 1,2,3

$$\eta_{i,j} = \mathrm{logit}(q_{i,j}) = \ln\left(\frac{q_{i,j}}{1-q_{i,j}}\right) = \ln\left(\frac{\mathrm{P}(Y_i \geq j)}{1-\mathrm{P}(Y_i \geq j)}\right)$$

$$= \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

- v měřítku **pravděpodobností**

$$q_{i,j} = q_j(x_i) = \frac{\exp(\alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})}{1+\exp(\alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})}$$

Proporcionalita

- Model proporcionálních **šancí** pro kumulativní logit

  j = 1,2,3

$$\eta_{i,j} = \mathrm{logit}(q_{i,j}) = \ln\left(\frac{q_{i,j}}{1-q_{i,j}}\right) = \ln\left(\frac{\mathrm{P}(Y_i \geq j)}{1-\mathrm{P}(Y_i \geq j)}\right)$$

$$= \underbrace{\alpha_j}_{\text{intercept pro každý cut-off zvlášť}} + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}_{\text{poměry šancí společné}}$$

Proporcionalita

- Model proporcionálních **šancí** pro kumulativní logit

  j = 1,2,3

$$\eta_{i,j} = \text{logit}(q_{i,j}) = \ln\left(\frac{q_{i,j}}{1-q_{i,j}}\right) = \ln\left(\frac{\text{P}(Y_i \geq j)}{1-\text{P}(Y_i \geq j)}\right)$$

$$= \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

- odhadnuté pravděpodobnosti

$$p_{i,3} = q_{i,3}$$

$$p_{i,2} = q_{i,2} - q_{i,3}$$

$$p_{i,1} = q_{i,1} - q_{i,2}$$

$$p_{i,0} = 1 - q_{i,1}$$

Interpretace

Výsledek modelování: závislost stadia fibrózy na krevních markerech

Byla provedena log2 transformace markerů – odhadujeme účinek zdvojnásobení jejich hodnot

Marker	Effect of Doubling	Effect of 1 SD on log-scale	P-Value
ha	1.48 (1.07, 2.04)	1.52	0.019
p3np	2.28 (1.37, 3.79)	1.69	0.0016
ykl40	1.72 (1.24, 2.39)	1.46	0.0011

- všechny markery jsou spojeny se stadiem choroby
- zdvojnásobení hodnoty markeru ykl40 dává o 72% vyšší šanci fibrózy vyššího stadia

Andersen & Skovgaard, 2010

Lékařská fakulta Masarykovy univerzity
Institut biostatistiky a analýz

Logistický a Poissonův model

Závěr

Co byste měli vědět a umět po dnešní hodině ?

- Chápat princip analýzy deviance
- Umět nadefinovat Poissonův model a popsat jeho užití
- Znát interpretaci probíraných modelů a jejich koeficientů
- Umět vysvětlit pojem overdispersion – čím je způsobena a jak ji poznat a řešit
- Znát základní možnosti modelování ordinálních výsledků

Dodatky ke zobecněným lineárním modelům

Cvičení

V adresáři naleznete článek:

Lee a kol.: Predicting Mortality Among Patients Hospitalised for Heart Failure

Úkoly:

1. Co představuje závisle proměnnou (výsledek)?
2. Jaký model byl využit pro modelování vztahu mezi prediktory a výsledkem?
3. Jaká byla modelovací strategie pro výběr prediktorů?
4. Byla ověřena celková shoda mezi pozorovanými a predikovanými odhady rizika (kalibrace)? Jak?
5. Najděte některé odlišnosti ve výsledcích univariátní a multivariátní analýzy?
6. Jak interpretujete výsledky multivariátního modelu (tabulka 3)?