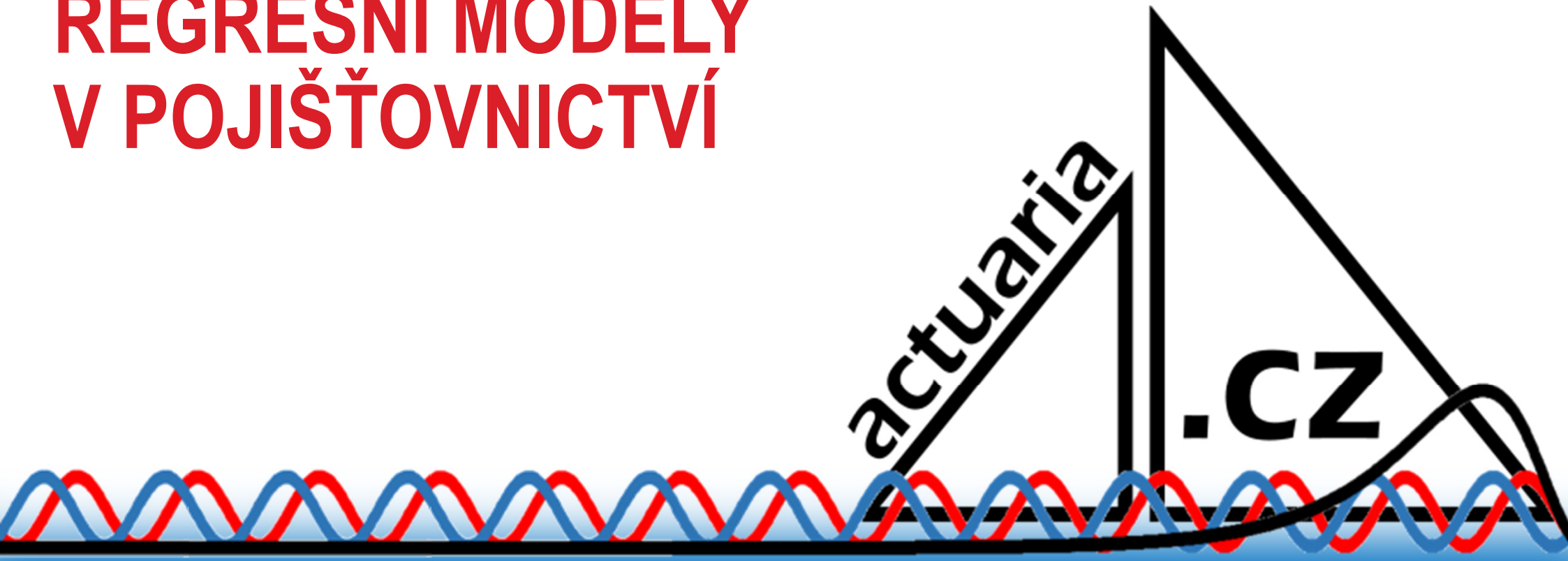
REGRESNÍ MODELY V POJIŠŤOVNICTVÍ



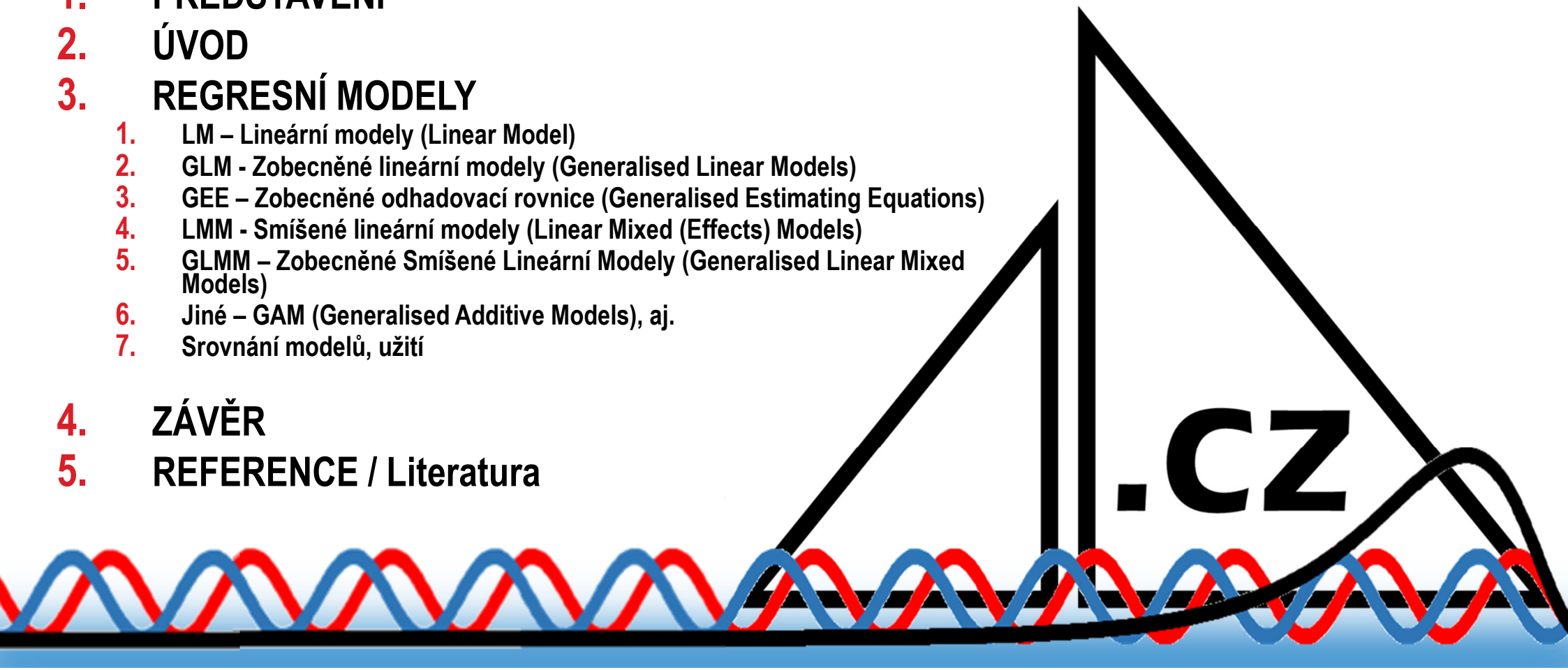
Seminář z aktuárských věd

2. prosince 2016

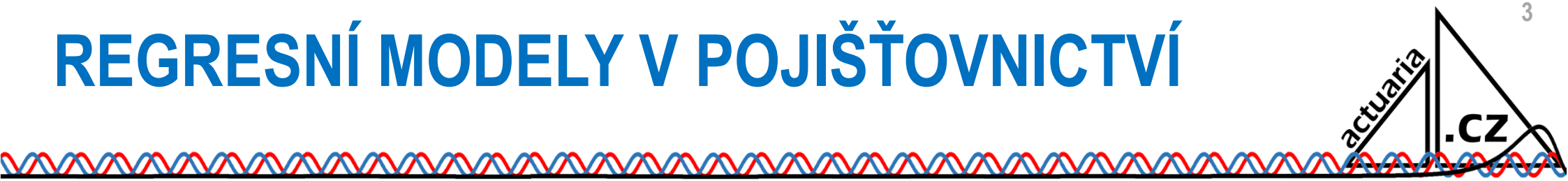
Kateřina Vlčková

REGRESNÍ MODELY V POJIŠŤOVNICTVÍ

1. PŘEDSTAVENÍ
2. ÚVOD
3. REGRESNÍ MODELY
    1. LM – Lineární modely (Linear Model)
    2. GLM - Zobecněné lineární modely (Generalised Linear Models)
    3. GEE – Zobecněné odhadovací rovnice (Generalised Estimating Equations)
    4. LMM - Smíšené lineární modely (Linear Mixed (Effects) Models)
    5. GLMM – Zobecněné Smíšené Lineární Modely (Generalised Linear Mixed Models)
    6. Jiné – GAM (Generalised Additive Models), aj.
    7. Srovnání modelů, užití
4. ZÁVĚR
5. REFERENCE / Literatura



REGRESNÍ MODELY V POJIŠŤOVNICTVÍ



▲ CO JE ««MODEL»»?

Model

- napodobenina předmětu postrádající některé původní vlastnosti
- objekt s charakteristickými vlastnostmi sloužící pro vytváření podobných objektů
- kategorie výrobků se společnými parametry

Matematický model

- Matematický model je abstraktní model používající matematický zápis k popisu chování soustavy (systému).

Co je MODEL ?

George E. P. Box

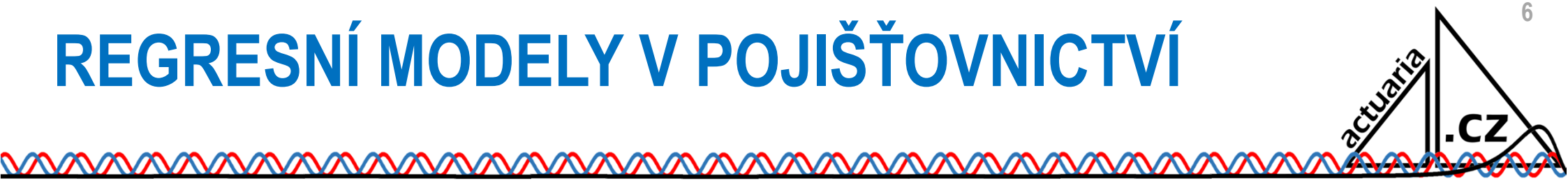
- "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

1987 , Empirical Model-Building and Response Surfaces

- „Essentially, all models are wrong, but some are useful.„

- George Edward Pelham Box,
  (*18. 10. 1919 – †28. 3. 2013)

REGRESNÍ MODELY V POJIŠŤOVNICTVÍ



▲ CO JE ««REGRESE »»?

Co je REGRESE?

- Do statistiky zavedl pojem **REGRESE** britský učenec Francis Galton kolem roku 1880, a to jako „**regres(i) k průměru“.** Tím označil fakt, že např. synové vysokých rodičů jsou sice v průměru (statisticky) vyšší než průměrná populace, zároveň ale individuálně nedosahují extrémních hodnot předchozí generace. Jako kdyby se jedinci postupně "vraceli k průměru". Podobně je tomu i s jinými vlastnostmi, nejen u lidí.

REGRESNÍ ANALÝZA

- Regresní analýza je označení statistických metod, které umožňují odhadovat hodnotu jisté náhodné veličiny na základě znalosti jiných veličin

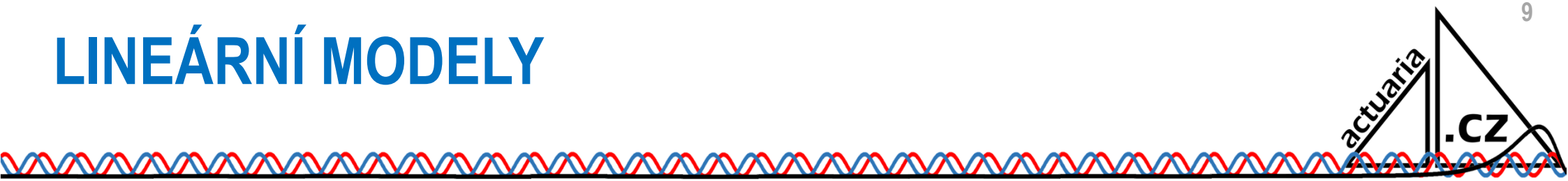
REGRESNÍ MODELY

Lineární modely

Linear Models

LM

LINEÁRNÍ MODELY



▲ CO JE «« LINEÁRNÍ »» ?

LINEÁRNÍ MODELY

CO JE «« LINEÁRNÍ »»

- Pomocí X ... regresorů / vysvětlujících proměnných (resp. jejich lineární kombinací $\eta = X\beta$) modelujeme $\mathbb{E}Y$

LINEÁRNÍ MODELY $\boldsymbol{Y = X\beta + \varepsilon}$

- **LM (lineární model):**
  - $\boldsymbol{Y = X\beta + \varepsilon}$,
  - $\boldsymbol{\varepsilon}$ …chybové členy jsou **nezávislé** náhodné veličiny,
    t.ž. $\mathbb{E}\varepsilon = \mathbf{0}, \boldsymbol{var}\varepsilon = \boldsymbol{\sigma^2}$

- **Obvykle předpokládáme Normální lineární model:**
  - $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2 I)$ tj. $\boldsymbol{Y}|\boldsymbol{X} \sim N(X\beta, \sigma^2 I)$
  - Pozn:
    - Pomocí X regresorů / vysvětlujících proměnných (resp. jejich lineární kombinací $= X\beta$) modelujeme $\mathbb{E}Y$
    - odezvy Y jsou homoskedastické (mají stejný rozptyl)

LINEÁRNÍ MODELY $\boldsymbol{Y = X\beta + \varepsilon}$

- **(N)LM (Normální lineární model):**
  - $\boldsymbol{Y = X\beta + \varepsilon}$, nezávislé $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 I)$
- **V Normálním LM platí:**
  - $\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{X\beta}, \sigma^2 \mathbb{I})$
  - $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}_0, \sigma^2 (\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1})$ - odhad
  - $\widehat{\boldsymbol{Y}} \sim \mathrm{N}_n(\boldsymbol{X\beta}_0, \sigma^2 \mathbb{H})$ - odhad
  - $\boldsymbol{u} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 (\mathbb{I} - \mathbb{H}))$ rezidua
  - $SS_e/\sigma \sim {\chi_{n-p}}^2$
  - SS_e a $\widehat{\boldsymbol{\beta}}$ jsou nezávislé

LINEÁRNÍ MODELY $\boldsymbol{Y = X\beta + \varepsilon}$

- **V Normálním LM platí:**
  - $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}\ \ , \sigma^2(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1})$

- $\widehat{\boldsymbol{\beta}}$ je **nestranný** odhad $\boldsymbol{\beta}$ (tj. $\mathbb{E}\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$)
- $\widehat{\boldsymbol{\beta}}$ je MLE – **maximálně věrohodný** - odhad $\boldsymbol{\beta}$
- $\widehat{\boldsymbol{\beta}}$ je **konzistentní** odhad $\boldsymbol{\beta}$

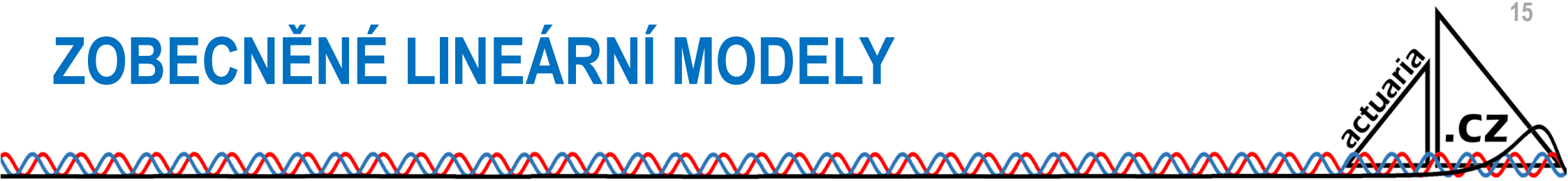
REGRESNÍ MODELY

Zobecněné lineární modely

Generalised Linear Models

GLM

ZOBECNĚNÉ LINEÁRNÍ MODELY



▲ CO JE «« ZOBECNĚNÝ»» ?

ZOBECNĚNÉ LINEÁRNÍ MODELY - GLM

- **Lineární model**
  - $\boldsymbol{Y = X\beta + \varepsilon,}$ nezávislost, $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2 I)$
  - Lineární model má svá omezení: Co nám vadí ?
    - Normalita → EDF – **Rozdělení exponenciálního typu**
    - „linearita“ → **linková / spojovací funkce**
  - Zobecnění lineárního modelu
  - Nelder, John A; Wedderburn, Robert W (1972). "Generalized linear models". Journal of the Royal Statistical Society, Series A. Royal Statistical Society. 135 (3): 370–384.
    - Zobecnění – definování společných předpokladů pro metody, které si již dříve používaly: *Logistická regrese, Poissonovská regrese, Gamma regrese, atd.*

ROZDĚLENÍ EXPONENCIÁLNÍHO TYPU

- Exponential Dispersion Family , EDF
- Rodina exponenciálních rozdělení
- Rozdělení s hustotou ve tvaru

$$f(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \qquad y \in \mathbb{R}$$

- $Y \in EDF$

ROZDĚLENÍ EXPONENCIÁLNÍHO TYPU

- EDF…hustota

$$\boldsymbol{f(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}}, \qquad y \in \mathbb{R}$$

- ϕ…disperzní (škálový) parametr, neznámý…společný pro celý model
- θ… kanonický parametr, neznámý…θ_i… závisí na pozorování
- a, b, c… funkce, známé
  - b… kumulantová funkce, dvakrát spojitě diferencovatelná
  - a… obvykle: $a(\phi) = \varphi$ nebo $a(\phi) = \varphi/w$, w… váhy

ROZDĚLENÍ EXPONENCIÁLNÍHO TYPU

- Normální rozdělení
- Gama rozdělení
- Inverzní Gaussovo rozdělení
- Poissonovo rozdělení
- Binomické rozdělení
- Negativně binomické rozdělení
- Geometrické rozdělení
- Alternativní rozdělení
- Exponenciální rozdělení
- Chí-kvadrát rozdělení
- Weilbullovo rozdělení
- Paretovo rozdělení

ROZDĚLENÍ EXPONENCIÁLNÍHO TYPU

- Y_i …náhodná veličina , $Y_i \in EDF$
  - Pro její momenty platí (b… dvakrát spojitě diferencovatelná funkce):
- $\mathbb{E}Y_i = b'(\theta_i) = \mu$
- $var\ Y_i = a(\phi)b''(\theta_i) = \varphi\ b''(\theta_i)$ $\quad (= \frac{\varphi}{w}\ b''(\theta_i)\ )$
- $V(\mu)$ … rozptylová (varianční) funkce: $\quad \boldsymbol{V(\mu) \equiv b''\left[(b')^{-1}(\mu)\right]}$
- $var\ Y_i = a(\phi)V(\mu) = \varphi\ V(\mu)$

ROZPTYLOVÁ FUNKCE $V(\mu)$

- $V(\mu)$ … rozptylová funkce *(variance function)*
- definovaná vtahem: $V(\mu) = b''[(b')^{-1}(\mu)] = \varphi\, V(\mu)$
- určuje vztah mezi střední hodnotou a rozptylem
- $\boldsymbol{var\, Y = a(\phi)\, V(\mu) = \varphi\, V(\mu)}$
- Jednoznačně identifikuje konkrétní rozdělení z EDF

EDF – SPOJITÁ ROZDĚLENÍ

- Normální:
  - $V(\mu) = 1$
- Gama:
  - $V(\mu) = \mu^2$
- Inverzní Gaussovo:
  - $V(\mu) = \mu^3$

EDF – DISKRÉTNÍ ROZDĚLENÍ

- Poissovovo:
  - $V(\mu) = \mu$
- Binomické:
  - $V(\mu) = n\,\pi\,(1-\pi)$
- Negativně binomické:
  - $V(\mu) = \mu\,(1-\mu\,\kappa)$

ROZDĚLENÍ EXPONENCIÁLNÍHO TYPU

Rozdělení	θ	$b(\theta)$	$a(\phi)=\varphi$	$\mu = EY$	$V(\mu) = varY/\varphi$
$\mathrm{B}(n,\pi)$	$\ln \frac{\pi}{1-\pi}$	$n \ln\left(1+e^{\theta}\right)$	1	$n\pi$	$n\pi(1-\pi)$
$\mathrm{P}(\mu)$	$\ln \mu$	e^{θ}	1	μ	μ
$\mathrm{N}\left(\mu,\sigma^2\right)$	μ	$\frac{1}{2}\theta^2$	σ^2	μ	1
$\mathrm{G}(\mu,\nu)$	$-\frac{1}{\mu}$	$-\ln(-\theta)$	$\frac{1}{\nu}$	μ	μ^2
$\mathrm{IG}\left(\mu,\sigma^2\right)$	$-\frac{1}{2\mu^2}$	$-\sqrt{-2\theta}$	σ^2	μ	μ^3
$\mathrm{NB}(\mu,\kappa)$	$\ln \frac{\kappa\mu}{1+\kappa\mu}$	$-\frac{1}{\kappa}\ln(1-\kappa\mathrm{e}^{\theta})$	1	μ	$\mu(1+\kappa\mu)$

Zdroj: P.de Jong, G.Z. Heller – Generalized Linear Models for Insurance Data

ZOBECNĚNÉ LINEÁRNÍ MODELY – od LM ke GLM

- **LM (lineární model):**
  - $Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2 I)$ tj. $Y|X \sim N(X\beta, \sigma^2 I)$
- **GLM zobecňuje (Normální) Lineární model:**
  - Rozdělení Y…odezvy / vysvětlované proměnné… nemusí být normální (ani se mu blížit)
  - Pomocí X regresorů / vysvětlujících proměnných (resp. jejich lineární kombinací $\eta = X\beta$) nemodelujeme $\mathbb{E}Y$, ale její transformaci $\boldsymbol{g}(\mathbb{E}\boldsymbol{Y})$
  - Důsledek: odezvy Y mohou být (a obvykle jsou) heteroskedastické

ZOBECNĚNÉ LINEÁRNÍ MODELY – 3 pilíře GLM

Definice modelu GLM

- **Rozdělení exponenciálního typu**…Y … $f \in$ EDF
  - $f(y;\theta,\varphi) = exp\left\{\frac{y\theta - b(\theta)}{\varphi} + c(y,\varphi)\right\}, \quad y \in \mathbb{R}$
- **Lineární prediktor**… η *(čti éta)*… lineární kombinace regresorů
  - $\eta = X\beta$
- **Linková (spojovací) funkce**…g…striktně monotónní, dvakrát spojitě diferencovatelná
  - $g(\mu) = \eta = g(X\beta) = g(\mathbb{E}Y)$ $\qquad\qquad$ $\mu = \mathbb{E}Y = g^{-1}(\eta)$
- Další předpoklady: 1. rozdělení Y závisí na X. 2a) (Y,X) nezávislé N.Ve. Nebo 2b) Y jsou nezávislé N.V. a X měřené konstanty.

GLM – KANONICKÝ LINK

- **Linková (spojovací) funkce**…g…striktně monotónní, dvakrát spojitě diferencovatelná
  - $g(\mu) = \eta = X\beta = g(\mathbb{E}Y)$ $\qquad\qquad$ $\mu = \mathbb{E}Y = g^{-1}(\eta)$
- **Kanonický link**
  - $g \equiv (b')^{-1} \implies g(\mu) = \eta = \theta$,
  - tj. lineární prediktor je roven kanonickému parametru
  - Platí : $g'(\mu) = {}^{1}/_{V(\mu)}$ $\qquad\qquad$ $V(\mu) \equiv b''[(b')^{-1}(\mu)]$
  - Kanonický link zjednodušuje vtahy při odhadu parametrů

Linková funkce

- **Linková funkce:**
  - $\boldsymbol{g(\mu) = \eta = X\beta = g(\mathbb{E}Y)}$;
  - $\boldsymbol{\mu = \mathbb{E}Y = g^{-1}(\eta)}$
- **Kanonický link:**
  - $\boldsymbol{g \equiv (b')^{-1} \Rightarrow}$
  - $\boldsymbol{g(\mu) = \eta = \theta}$,

Příklady

Linková funkce	$g(\mu)$	Rozdělení y
identita	$\boldsymbol{\mu}$	Normální
log	$\ln(\mu)$	Poisson
inverze	$\frac{1}{\mu}$	Gama
mocnina	$\mu^p = \mu^{-1}$	Gama p= -1
mocnina	$\mu^p = \mu^{-2}$	Inverzní Gauss p= -2
odmocnina	$\sqrt{\mu}$	
logit	$ln\frac{\mu}{1-\mu}$	Binomické

GLM – VZTAHY MEZI PARAMETRY

$$\boldsymbol{X\beta} = \boldsymbol{\eta} \qquad \boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\mu}) \qquad \boldsymbol{\mu} = \boldsymbol{b}'(\boldsymbol{\theta})$$

$$\boldsymbol{\beta} \longrightarrow \boldsymbol{\eta} \longleftrightarrow \boldsymbol{\mu} \longleftrightarrow \boldsymbol{\theta}$$

$$\boldsymbol{g}^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu} \qquad (\boldsymbol{b}')^{-1}(\boldsymbol{\mu}) = \boldsymbol{\theta}$$

η… lineární prediktor

β… regresní parametr

g… linková funkce

μ… střední hodnota EY,

θ… kanonický parametr

b…kumulantová funkce

7 KROKŮ K MODELU GLM

- Volba rozdělení odezvy / vysvětlované proměnné… $\boldsymbol{f(y)}$
- Volba linkové funkce… $\boldsymbol{g(\mu)}$
- Volba regresorů / vysvětlujících proměnných…$\boldsymbol{x}$
- Výběr (nezávislých) dat: pozorování $y_1, \dots, y_n$ a odpovídajících hodnot regresorů $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$
- Odhad regresních parametrů β a případně disperzního parametru φ
- Model fit: kvalita modelu, výběr podmodelů
- Kalibrace modelu, odhad predikční chyby modelu

GLM – 1) volba rozdělení odezvy … $f(y)$

- Normální rozdělení
- Gama rozdělení
- Inverzní Gaussovo rozdělení
- Poissonovo rozdělení
- Binomické rozdělení
- Negativně binomické rozdělení
- Alternativní rozdělení
- Jiné $\in$ EDF

GLM – 2) volba linkové funkce… $g(\mu)$

- kanonický link
- jiná linková funkce

Linková funkce	$g(\mu)$	Rozdělení y
identita	μ	Normální
log	$\ln(\mu)$	Poisson
inverze	$\frac{1}{\mu}$	Gama
mocnina	$\mu^p = \mu^{-1}$	Gama p= -1
mocnina	$\mu^p = \mu^{-2}$	Inverzní Gauss p= -2
odmocnina	$\sqrt{\mu}$	
logit	$ln\frac{\mu}{1-\mu}$	Binomické

GLM – 3) volba regresorů …x

- **REGRESORY – Vysvětlující proměnné**
  - **Metrické proměnné**: věk, výše škody
  - **Faktory**: pohlaví, typ vozidla, počet dětí
  - **Interakce:** MxM, FxF, MxF
- **Metrické proměnné:**
  - model odhaduje vliv jednotlivých proměnných na odezvu
  - β_j ... vyjadřuje změnu při jednotkové změně regresoru *x*
  - Lze i polynomy
- **Faktory:**
  - Diskrétní proměnné nebo katerorie
  - Dummy proměnné (o 1 katerorii méně, referenční kategorie - typická)
  - β_j ... vyjadřuje změnu oproti referenční kategorii
- **Interakce:** $\boldsymbol{\beta_{1,2} x_1 x_2}$

GLM – 3) volba regresorů - OFFSET

- Člen v lineárním prediktoru s pevně daným koeficientem
- Regresní koeficient roven 1, neodhaduje se
- Obvykle použit jako korekce modelu s ohledem na expozici v riziku
  - velikost skupiny,
  - různá doba pozorování
- Nejčastěji: pro logaritmický link; zde příklad pro expozici n_i. řádku:

$$\eta_i = \ln n_i + x_i^{\mathrm{T}}\beta \qquad \mu_i = e^{\eta_i} = n_i \cdot e^{x_i^{\mathrm{T}}\beta}$$

- PRAXE: délka platnosti smlouvy, počet rizik

GLM – 3) volba regresorů – váhy pozorování w

- Do modelu je možné zahrnou apriorní váhy pro jednotlivá pozorování … w
- Parametrizace v EDF: $a(\phi) = {}^{\varphi}/_{w}$

$$\mathrm{E}[\boldsymbol{Y}] = b'(\theta), \qquad \mathrm{var}[\boldsymbol{Y}] = b''(\theta) \cdot {}^{\varphi}/_{w} = V(\theta) \cdot {}^{\varphi}/_{w}$$

- PRAXE: průměrné výše škody: w … počet škod na dané smlouvě
- PRAXE: škodní frekvence: w … délka expozice / délka platnosti smlouvy

GLM – 4) Výběr dat: $\ldots y_1, \ldots, y_n$

- nezávislá data
- náhodný výběr

GLM – 5a) odhad regresních parametrů β

- **ODHADY REGRESNÍCH PARAMETRŮ V GLM modelu – *MODEL FIT***
  - Odhady a INFERENCE v GLM jsou založeny na teorii Maximální věrohodnosti
  - Maximalizace vyžaduje iterativní řešení
    - Metoda iterativních vážených čtverců? IWLS X IRLS
    - Newton –Raphsonův iterační algoritmus
    - Fisherova metoda skórů
  - MLE – Maximálně věrohodné odhady – MAXIMUM LIKELIHOOD
    - Vlastnosti MLE odhadů:
      - Asymptoticky nezkreslené, konzistentní, asymptoticky vydatné, invariantní vůči monotónní funkci, asymptotický normální
  - IRLS algoritmus –Iteratively Re-Weighted Least Squares

GLM – 5b) odhad disperzního parametru φ

- Disperzní (škálový) parametr φ
- Obvykle není znám
- Pro MLE odhady regresních parametrů β_0 není nutné znát odhad skutečné hodnoty disperzního parametru φ_0. MLE odhady jsou stejné v obou případech. Asymptotické vlastnosti platí.
- Odhad disperzního parametru φ_0 je však potřeba pro odhad asymptotického rozptylu.
- MLE odhad φ_0 není vždy možné vypočítat. Pro odhady se proto obvykle používá modifikovaná Momentová metoda.
- Odhad založen na Pearsonově χ^2 statistice

$$\widehat{\varphi_n} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i-\widehat{\mu_i})^2}{V(\widehat{\mu_i})} = \frac{1}{n-p}X^2$$

GLM – 6) Výběr modelu

- Kvalita modelu a testy
- Saturovaný model:
  - Počet parametrů = počet pozorování
  - Prakticky se nepoužívá = perfect fit
  - Teoretická aplikace = je v něm dosaženo maximální možné věrohodnosti
- Deviance
  - $D = 2 * \left[l(Y) - l(\hat{\beta}) \right]$ (škálová deviance)
  - Analogie indexu determinace v lineárním modelu
- Waldovy testy
- Testy poměrem věrohodnosti

GLM – 6) Výběr modelu – INFORMAČNÍ KRITÉRIA

- Čím více vysvětlujících proměnných v modelu, tím lepší „fit“ modelu, ale: tím více parametrů a horší přesnost jejich odhadů (zvyšuje se rozptyl)
- => hledáme kompromis (vysoká věrohodnost × malý počet parametrů)

- **Akaikeho informační kritérium:**
- $AIC = -2\left(l\left(\boldsymbol{Y};\hat{\beta},\hat{\varphi}\right) - m\right)$
- m … počet parametrů β (penalizace)

- **Bayesovské informační kritérium BIC**
- $BIC = -2\, l\left(\boldsymbol{Y};\hat{\beta},\hat{\varphi}\right) + \ln(n)\cdot \dim(\beta)$

- Vybíráme modely s nízkým AIC a BIC.
- BIC více penalizuje za počet parametrů, v porovnání s AIC vybírá modely s menším počtem vysvětlujících proměnných (někdy až příliš chudé???)
- Porovnávané modely musí být založeny na stejných pozorováních

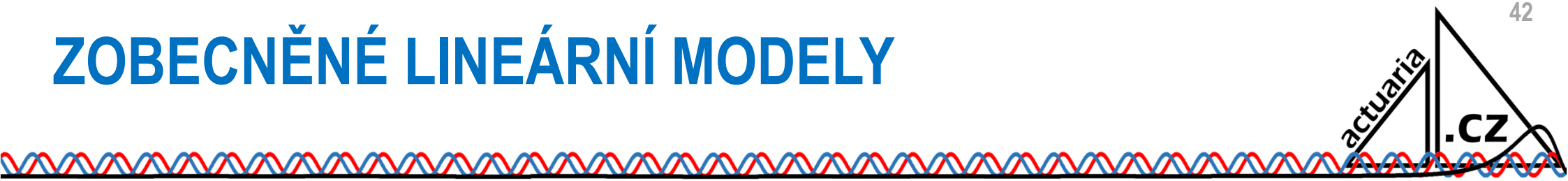
GLM – 6) Výběr modelu – metody výběru (pod)modelů

- **Best Subset Selection** – regresní model sestaven pro daný počet regresorů a všechny možné jejich kombinace, vybrán nejlepší model pro zvolené kritérium (AIC, BIC, CD), výpočetně náročné až nereálné
- **Kroková regrese** (Stepwise Regression) – postupné přidávání/ubírání regresorů, postup ve směru největšího poklesu hodnoty kritéria, výpočetně jednodušší
- **Regularizovaná regrese** – ve věrohodnostní funkci se penalizuje nárůst regresních parametrů
- **Sekvenční výběr proměnných** – manuální verze krokové regrese

GLM – 7) Kalibrace modelu, odhad predikční chyby

- **Validace** : rozdělení dat na Trénovací část (kalibrace - odhad modelu) a Testovací část (stanovení predikční chyby)
- **Cross-validace**: data rozdělena na několik částí: na všech kromě jedné odhadujeme model, na poslední testujeme; opakujeme pro všechny kombinace
- **Nová data** – pro stanovení predikční chyby

ZOBECNĚNÉ LINEÁRNÍ MODELY



▲ CO SE DO «« GLM »» NEVEŠLO

CO DÁL? - ROZŠÍŘENÍ / ZOBECNĚNÍ GLM

- V praxi však nebývají všechny předpoklady splněny…
- Je možné rozšíření teorie GLM na případy, kdy jsou předpoklady porušeny?
  - Rozdělení závislých proměnných / odezvy není exponenciálního typu
  - Rozdělení odezvy není blíže specifikováno
    - Známe první 2 momenty
    - Známe vztah mezi $\mathbb{E}Y_i$ a $var\ Y_i$
  - Data nejsou nekorelovaná / nezávislá
  - Regresní parametry / efekty β nejsou pevné, ale náhodné
  - Data vykazují nadměrnou disperzitu oproti teoretickým hodnotám

CO DÁL? - ROZŠÍŘENÍ / ZOBECNĚNÍ GLM

OVERDISPERSION / NADMĚRNÁ DISPERZITA

- Pozorovaná data mají vyšší variabilitu/disperzitu, než by se očekávalo při platnosti zvoleného modelu
- Průměr je většinou možné parametricky upravit, aby odpovídal teoretické hodnotě.
- Vyšší momenty se však (obzvláště u malých výběrů) upravují těžko
- S nadměrnou disperzí se často setkáváme u modelů četností (Poissonovo rozdělení, či binomické), obzvláště u malých výběrů a heterogenních populací
  - Př: nadměrná disperze v Binomických datech -> Beta-binomické rozdělení
  - Př: nadměrná disperze v Poissonovských datech -> Poisson-Gama rozdělení

OVERDISPERSION / NADMĚRNÁ DISPERZITA

- **Příklad: Poissonovo rozdělení**
  - $Y_1, \dots, Y_n \sim \boldsymbol{Po}(\lambda_0)$ nezávislá pozorování
  - $var[Y_i] = \mathbb{E}[Y_i] = \lambda_0, \quad Y_i \in EDF$
  - Parametr$[\lambda_i]$považujme za náhodnou veličinu (nikoli parametr), $\mathbb{E}[\lambda_i] = \lambda_0$
  - $Y_i|\lambda_0 \sim \boldsymbol{Po}(\lambda_0), \quad \mathbb{E}[\lambda_i] = \lambda_0 \quad var[\lambda_{ii}] = \sigma 2_\lambda$
  - $\mathbb{E}[Y_i] = \lambda_0 \quad var[Y_i] = \lambda_0 + \sigma^2{}_\lambda$
  - Předpokládejme: $\lambda_i \sim \Gamma(a, a\lambda_0)$
    - -> Poisson-Gama rozdělení - není EDP, GLM neumí řešit
    - Speciální případ: negativně-binomické rozd., geometrické

CO DÁL? - ROZŠÍŘENÍ / ZOBECNĚNÍ GLM

KVAZIVĚROHODNOST / QUASI-LIKELIHOOD

- Neznáme rozdělení, ale známe rozptylovou funkci $V(.)$, která určuje vztah mezi prvními dvěma momenty,
  - $V(\mu) = b''[(b')^{-1}(\mu)] \quad \longrightarrow \quad var\, Y_i = \varphi\, V(\mu) = \varphi\, V(\mathbb{E}Y_i)$
- MODEL:
  - Mějme n náhodných vektorů $(Y_i, \boldsymbol{X}_i)$, $i = 1, \dots, n$, kde $\boldsymbol{X_i} = (X_{i1}, X_{i2}, \dots, X_{ip})^{\mathrm{T}}$ jsou regresory a Y_i jsou odezvy.
  - Předpokládejme:
    - 1. $Y_1, \dots, Y_n$ jsou nezávislé
    - 2. $\mu_i = \mathbb{E}[Y_i]$, $g(\mu_i) = \eta_i = \boldsymbol{X_i}^{\mathrm{T}}\boldsymbol{\beta}_0$, η_i… lineární prediktor, ¨g… linková funkce
    - 3. $var\, Y_i = \varphi\, V(\mu_i)$, φ… disperzní parametr, V… varianční funkce
- GLM je parametrický model, toto semi-parametrický model: rozdělení není dáno, pouze vztah mezi momenty
- Metodu maximální věrohodnosti (MLE) lze použít pro odhady pouze pro parametrické modely.
- Odhady regresních parametrů – „Metoda maximální kvazivěrohosnosti“ (?MQLE)
- $\frac{1}{\sqrt{n}}\, U_n(\boldsymbol{\beta}_0) \xrightarrow{d} N_p(\mathbf{0}, \mathbb{I}(\boldsymbol{\beta}_0)) \qquad \sqrt{n}(\widehat{\boldsymbol{\beta}_n} - \boldsymbol{\beta}_0) \xrightarrow{d} N_p(\mathbf{0}, \mathbb{I}^{-1}(\boldsymbol{\beta}_0))$

- Je možné aplikovat Waldovy a skórové testy; nikoli však devianci či AIC.

REGRESNÍ MODELY

Zobecněné odhadovací rovnice

Generalised Estimating Equations

GEE

GEE - Zobecněné odhadovací rovnice

- **GEE – Generalised Estimating Equations**
  – Zobecněné odhadovací rovnice
  - *rozšíření GLM – korelace, skupinově závislá data*
  - GEE model navržen v článku z r. 1986 – K.Y. Liang a S.L. Zeger: *Longitudial data analysis using generalized linear models*
  - Umožňuje použít postupy z GLM i v případě, kdy nejsou splněny předpoklady pro GLM model
  - **Data nejsou nekorelovaná**

GEE - DEFINICE MODELU pro skupinově závislá data

- **Data**: $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K$ nezávislé náhodné vektory $\boldsymbol{Y}_i = \left(Y_{i1}, \ldots, Y_{in_i}\right)^{\mathrm{T}}, i = 1 \ldots K,\ \sum_{\mathrm{i}=1}^{\mathrm{K}} n_i = n$
- Data rozdělena do K nezávislých skupin (shluků, subjektů, jedinců)
- V každé skupině je různý počet (n_i) vzájemně korelovaných pozorování
- Tj. data jsou: **závislá v rámci jedné skupiny a nezávislá mezi jednotlivými skupinami** (př. zuby pacienta, mláďata z jednoho vrhu, škody na jedné pojistné smlouvě, budovy v jedné obci)
- **Shluková data** (bez uspořádání), **opakovaná měření** (uspořádání v čase), **longitudinální data** (se záznamem o čase), **panelová data** (ekonomická data)
- Ke každému pozorování (odezvě, závisle/vysvětlované proměnná) Y_{ij}

  přísluší vektor regresorů (nezávislých/vysvětlujících proměnných) $\mathbf{X}_{ij} = \left(\mathrm{X}_{\mathrm{ij1}}, \ldots, \mathrm{X}_{\mathrm{ijp}}\right)^{\mathrm{T}}, \mathrm{i} = 1 \ldots \mathrm{K},\ \mathrm{j} = 1, \ldots \mathrm{n_i},$ $\sum_{\mathrm{i}=1}^{\mathrm{K}} n_i = n$.

GEE - DEFINICE MODELU pro skupinově závislá data

- **Data**: $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_K$ nezávislé náhodné vektory $\boldsymbol{Y}_i = \left(Y_{i1}, \dots, Y_{in_i}\right)^{\mathrm{T}}, i = 1 \dots K,\ \sum_{\mathrm{i=1}}^{\mathrm{K}} n_i = n$
- Ke každému pozorování (odezvě) Y_{ij} přísluší vektor regresorů $\mathbf{X}_{\mathrm{ij}} = \left(X_{ij1}, \dots, X_{ijp}\right)^{T}$,
  $\mathrm{i} = 1, \dots \mathrm{K},\ \mathrm{j} = 1, \dots \mathrm{n_i},\ \sum_{\mathrm{i=1}}^{\mathrm{K}} n_i = n$.
- **Cíl**: popsat závislost střední hodnoty pozorování $\mu_{\mathrm{ij}} = \mathbb{E}\mathrm{Y}_{\mathrm{ij}}$ na regresorech $\boldsymbol{X}_{ij}$ pomocí regresního modelu.
- **Předpoklad**: vztah je definován pomocí linkové funkce $g\left(\mu_{ij}\right) = \mathbb{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0$, stejně jako v GLM
- $\boldsymbol{g}$ … linková funkce: striktně monotónní, dvakrát spojitě diferencovatelná
- $\boldsymbol{\beta}_0 = \left(\beta_{01}, \dots, \beta_{0p}\right)^{\mathrm{T}}$ neznámý vektor regresních parametrů (skutečná hodnota)
- Předpokládejme :
  - $\mathbb{E}\boldsymbol{Y}_i = \boldsymbol{\mu}_i = \left(\mu_{i1}, \mu_{i2}, \dots, \mu_{in_i}\right)^{\mathrm{T}} = \left(g^{-1}\left(\mathbb{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_0\right), \dots, g^{-1}\left(\mathbb{X}_{in_i}^{\mathrm{T}}\boldsymbol{\beta}_0\right)\right)^{\mathrm{T}}$
  - $var\ \boldsymbol{Y_i}$ nespecifikováno, bez předpokladů o rozptylu či kovarianci

GEE - ODHADY REGRESNÍCH PARAMETRŮ v modelu pro skupinově závislá data

- $\mathbb{X}_i = \left(\boldsymbol{X}_{i1}^{\mathrm{T}}, \ldots, \boldsymbol{X}_{ip}^{\mathrm{T}}\right)^{\mathrm{T}}$ … regresní matice $(n_i \times p)$

$$\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)_{(p \times n_i)} = \mathbb{X}_i^{\mathrm{T}} \begin{pmatrix} g'(\mu_{i1}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & g'(\mu_{in_i}) \end{pmatrix}^{-1}$$

… matice parciálních derivací

- Připomenutí:
  skórová funkce GLM modelu: $U_i(\beta) = \frac{\partial \mu_i}{\partial \beta} \frac{1}{\varphi\, V(\mu_i)} (Y_i - \mu_i)$, kde $V(\mu_i)$ je rozptylová funkce.
- Zobecnění pro vícerozměrné $\boldsymbol{Y}_i$:
  pseudo-skórová funkce příslušející i. skupině pozorování: $U_i(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}}\ {}_i\mathbb{Q}^{-1}(\boldsymbol{\mu_i})\ (Y_i - \mu_i)$.
- $\mathbb{Q}_i(\boldsymbol{\mu}_i) = \varphi\ \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)\ \mathbb{R}_i\ \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)$ … „**pracovní kovarianční matice**“, která reprezentuje naši představu (guess) o rozptylu $var\ \boldsymbol{Y_i}$.
- $\mathbb{V}_i$ …diagonální matice, na diagonále členy $V(\mu_{i1}), \ldots, V(\mu_{in_i})$, které představují „pracovní rozptyl“ pro pozorování $\boldsymbol{Y}_i = \left(Y_{i1}, \ldots, Y_{in_i}\right)^{\mathrm{T}}$.
- $\mathbb{R}_i$ … „**pracovní korelační matice**“ $(n_i \times n_i)$.
- Nepředpokládáme, že $\mathbb{R}_i$ a $V(\mu_{ij})$, jsou známé či správně odhadnuté. (V opačném případě bychom rovnou použili $cor\boldsymbol{Y}_i$ a $var\ Y_{ij}$ místo $\mathbb{R}_i$ a $\varphi V(\mu_{ij})$.)
- **GEE odhad regresních parametrů $\boldsymbol{\beta}$:** $\widehat{\boldsymbol{\beta}_K}$ je definován jako řešení soustavy „**zobecněných odhadovacích rovnic**“:

$$U_K\left(\widehat{\boldsymbol{\beta}_K}\right) = \sum_{i=1}^{K} U_i\left(\widehat{\boldsymbol{\beta}_K}\right) = \mathbf{0} \qquad \text{(GEE)}$$

GEE - asymptotické vlastnosti odhadů regresních parametrů

- Označme: $\mathbb{D} = -\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}^{\mathrm{T}}} U_i(\boldsymbol{\beta}_0)\right] = \mathbb{E}\left[\left(\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}}\right) \mathbb{Q}^{-1}(\boldsymbol{\mu_i}) \left(\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}}\right]$
- Pro $K \to \infty$ platí:
- (i) $\widehat{\boldsymbol{\beta}_K} \xrightarrow{p} \boldsymbol{\beta}_0$ (konzistence)
- (ii) $\frac{1}{\sqrt{K}} U_K(\boldsymbol{\beta}_0) \xrightarrow{d} N_p(\mathbf{0}, \boldsymbol{\Sigma})$
- (iii) $\sqrt{K}(\widehat{\boldsymbol{\beta}_K} - \boldsymbol{\beta}_0) \xrightarrow{d} N_p(\mathbf{0}, \mathbb{D}^{-1}\boldsymbol{\Sigma}\mathbb{D}^{-1})$. (asymptotická normalita)
- Asymptotika funguje pro velká $K \to \infty$. **Potřebujeme velký počet nezávislých skupin.**
- Počet (korelovaných) pozorování v jednotlivých skupinách není podstatný.

GEE - odhad asymptotického rozptylu

- Asymptotický rozptyl $\mathbb{D}^{-1}\boldsymbol{\Sigma}\mathbb{D}^{-1}$ odhadujeme pomocí sendvičového odhadu (sandwich estimator / White estimator):
- $\widehat{\mathbb{D}^{-1}\boldsymbol{\Sigma}\mathbb{D}^{-1}} = \widehat{\mathbb{D}}^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbb{D}}^{-1}$
- kde: $\widehat{\mathbb{D}} = \frac{1}{K}\sum_{i=0}^{n}\left(\frac{\partial\widehat{\boldsymbol{\mu}}_{\boldsymbol{i}}}{\partial\boldsymbol{\beta}}\right)\mathbb{Q}^{-1}(\widehat{\boldsymbol{\mu}}_{\boldsymbol{i}})\left(\frac{\partial\widehat{\boldsymbol{\mu}}_{\boldsymbol{i}}}{\partial\boldsymbol{\beta}}\right)^{\mathrm{T}}$
- a $\widehat{\boldsymbol{\Sigma}} = \frac{1}{K}\sum_{i=1}^{K} U_i\left(\widehat{\boldsymbol{\beta}}_K\right)^{\otimes 2}$.

- $\mathbb{Q}(\boldsymbol{\mu}_i) = \varphi\ \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)\ \mathbb{R}_i\ \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)$ … „**pracovní kovarianční matice**“, která reprezentuje naši představu (guess) o rozptylu $var\ \boldsymbol{Y}_{\boldsymbol{i}}$.
- $\mathbb{V}_i$ …diagonální matice, na diagonále členy $V(\mu_{i1}), \dots, V(\mu_{in_i})$, které představují „pracovní rozptyl“ pro pozorování $\boldsymbol{Y}_i = \left(Y_{i1}, \dots, Y_{in_i}\right)^{\mathrm{T}}$.
- $\mathbb{R}_i$ … „**pracovní korelační matice**“ $(n_i \times n_i)$.
- Nepředpokládáme, že $\mathbb{R}_i$ a $V(\mu_{ij})$, jsou známé či správně odhadnuté.

GEE - řešení soustavy zobecněných odhadovacích rovnic

$$U_K(\widehat{\boldsymbol{\beta}_K}) = \sum_{i=1}^{K} U_i(\widehat{\boldsymbol{\beta}_K}) = \mathbf{0} \qquad \text{(GEE)}$$

- Soustava (GEE) se řeší iteračně.¨
- GEE odhad $\widehat{\boldsymbol{\beta}}_K$ hledáme pomocí modifikované metody IWLS (Irerrative Re-Weighted Least Squares).
- Iterujeme: $\widehat{\boldsymbol{\beta}} = \left[\sum_{i=0}^{n} \left(\frac{\partial \widehat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}}\right) \mathbb{Q}^{-1}(\widehat{\boldsymbol{\mu}}_i) \left(\frac{\partial \widehat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}}\right]^{-1} \left[\sum_{i=0}^{n} \left(\frac{\partial \widehat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}}\right) \mathbb{Q}^{-1}(\widehat{\boldsymbol{\mu}}_i)\ \widehat{\boldsymbol{Z}}_i\right]$,
- kde: $\widehat{\boldsymbol{Z}}_i = (\hat{Z}_{i1}, \ldots, \hat{Z}_{in_i})^{\mathrm{T}}$ a $\hat{Z}_{ij} = \frac{\hat{\eta}_{ij} + (Y_{ij} - \hat{\mu}_{ij}) g'(\hat{\mu}_{ij})}{g'(\hat{\mu}_{ij})}$

GEE – volba korelační struktury

- Rozptylová funkce $V(\mu)$ v GLM vyjadřuje vztah mezi rozptylem a střední hodnotou.

- **Pracovní kovarianční matice** $\mathbb{Q}(\boldsymbol{\mu}_i) = \varphi\ \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)\ \mathbb{R}_i\ \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)$,
  resp. **pracovní korelační matice** $\mathbb{R}_i$,
  reprezentuje v GEE rovněž naši představu o rozptylu.

- Jak zvolit vhodnou pracovní korelační matici?

GEE – volba korelační struktury

- **pracovní nezávislost / working independence**
- **metoda parametrizované korelace / parametrized correlation**
  - pásová korelace 1. řádu
  - pásová korelace m. řádu:
  - exchangeable correlations
  - AR(1) korelace a další korelace na bázi časových řad

GEE – volba korelační struktury
pracovní nezávislost

- Nejjednodušší volba, analyzujeme **data**, jako by byla **nezávislá**.
- Volíme pracovní korelační matici : $\mathbb{R}_i \equiv \mathbb{I}_{n_i}$ (čtvercová jednotková matice $(n_i \times n_i)$).
- Pracovní kovarianční matice je diagonální: $\mathbb{Q}\,(\boldsymbol{\mu}_i) = \varphi\ \mathbb{V}_i\,(\boldsymbol{\mu}_i) = \begin{pmatrix} \varphi\, V(\mu_{i1}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varphi\, V(\mu_{in_i}) \end{pmatrix}$.
- Odhady regresních parametrů $\boldsymbol{\beta}_0$ jako při nezávislých datech, pomocí standardního IWLS algoritmu.
- Rozptyl odhadů je poté upraven pomocí sendvičového odhadu, zohlednění možných korelací a případné nevhodné volby rozptylové funkce $V(\cdot)$.
- Odhady jsou konzistentní. V případě velké korelace mezi daty nejsou eficientní.

GEE – volba korelační struktury
metoda parametrizované korelace

- Volíme **kovarianční strukturu**, která **není nezávislá**.
- Zavedeme nový vektorový parametr $\boldsymbol{\alpha} \in \mathbb{R}^m$.
- Položíme:
  - pracovní korelační matice: $\mathbb{R}_i \equiv \mathbb{R}_i(\boldsymbol{\alpha})$
  - pracovní korelační matice: $\mathbb{Q}_i(\boldsymbol{\mu}_i) \equiv \mathbb{Q}_i(\boldsymbol{\mu}_i, \boldsymbol{\alpha}) = \varphi\ \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)\ \mathbb{R}_i(\boldsymbol{\alpha})\ \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)$
  - odhad pracovní korelační matice: $\widehat{\mathbb{Q}}_i(\boldsymbol{\mu}_i) \equiv \mathbb{Q}_i(\boldsymbol{\mu}_i, \widehat{\boldsymbol{\alpha}}) = \hat{\varphi}\mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)\ \mathbb{R}_i(\widehat{\boldsymbol{\alpha}})\ \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)$
  - kde $\widehat{\boldsymbol{\alpha}}$ je $\sqrt{K}$-konzistentní odhad parametru $\boldsymbol{\alpha}$, např. momentový odhad.
- Dostáváme modifikované skóre: $\widehat{U}_i(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\ \widehat{\mathbb{Q}_i}^{-1}(\boldsymbol{\mu}_i)\ (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$.

  (! Ztráta nezávislosti vektorů)
- Odhady regresních parametrů $\widehat{\boldsymbol{\beta}}$ jsou řešením soustavy $U_K(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{K} \widehat{U}_i(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$.

GEE – volba korelační struktury
obecný postup odhadu parametru α

- Obecný postup při odhadování parametru $\boldsymbol{\alpha}$: (založen na reziduích):
- Odhad $\boldsymbol{\beta}$ za předpokladu pracovní nezávislosti (Volíme: $\mathbb{R}_i = \mathbb{I}_{n_i}$)
- Spočteme Personova rezidua: $r_{ij}^P = \frac{Y_{ij} - \hat{\mu}_{ij}^{(I)}}{\sqrt{V\left(\hat{\mu}_{ij}^{(I)}\right)}}$
- Pokud $\hat{\mu}_{ij}^{(I)}$ správně odhadují střední hodnotu $\mathbb{E}Y_{ij}$, potom:
- $\mathbb{E}\left[r_{ij}^P\right] \approx 0, \quad var\left[r_{ij}^P\right] \approx \varphi, \quad \mathbb{E}\left[r_{ij}^P, r_{ik}^P\right] \approx \varphi\, cor\left(Y_{ij}, Y_{ik}\right) = \varphi\left(R_{ijk}(\boldsymbol{\alpha})\right)$
- Hledáme momentové odhady parametru $\boldsymbol{\alpha}$ spočtené ze součinů Personových reziduí $r_{ij}^P \cdot r_{ik}^P$ y dat ze stejné skupiny.

GEE – volba korelační struktury metoda parametrizované korelace
Příklady volby korelační struktury

- **Pásová korelace 1. řádu:**

- $R_i(\boldsymbol{\alpha}) = cor(Y_{ij}, Y_{ik}) = \begin{pmatrix} 1 & \alpha & 0 & \cdots & \cdots & 0 \\ \alpha & 1 & \alpha & \ddots & & \vdots \\ 0 & \alpha & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \alpha & 0 \\ \vdots & & \ddots & \alpha & 1 & \alpha \\ 0 & \cdots & 0 & 0 & \alpha & 1 \end{pmatrix}$

- Konzistentní odhad parametru $\boldsymbol{\alpha}$: $\widehat{\boldsymbol{\alpha}} = \frac{1}{\widehat{\varphi}} \frac{1}{n-K-p} \sum_{i=1}^{K} \sum_{j=1}^{n_i-1} r_{ij}^{P} \cdot r_{i,j+1}^{P}$

GEE – volba korelační struktury metoda parametrizované korelace
Příklady volby korelační struktury

- **Pásová korelace m. řádu: (m=2)**

$$R_i(\boldsymbol{\alpha}) = cor(Y_{ij}, Y_{ik}) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & 0 & \cdots & 0 \\ \alpha_1 & 1 & \alpha_1 & \ddots & & \vdots \\ \alpha_2 & \alpha & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \alpha_1 & \alpha_2 \\ \vdots & & \ddots & \alpha_1 & 1 & \alpha_1 \\ 0 & \cdots & 0 & \alpha_2 & \alpha_1 & 1 \end{pmatrix}$$

- matici rozšíříme na více pásem, m nesmí být moc vysoké
- Konzistentní odhad parametru $\boldsymbol{\alpha}$: $\hat{\boldsymbol{\alpha}} = \frac{1}{\hat{\varphi}} \frac{1}{n-K*l-p} \sum_{i=1}^{K} \sum_{j=1}^{n_i-1} r_{ij}^{P} \cdot r_{i,j+1}^{P}$

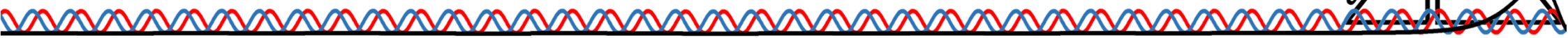
GEE – volba korelační struktury metoda parametrizované korelace
Příklady volby korelační struktury

- **Exchangable correlation**

- $R_i(\boldsymbol{\alpha}) = cor(Y_{ij}, Y_{ik}) = \begin{pmatrix} 1 & \alpha & \alpha & \cdots & \cdots & \alpha \\ \alpha & 1 & \alpha & \ddots & & \vdots \\ \alpha & \alpha & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \alpha & \alpha \\ \vdots & & \ddots & \alpha & 1 & \alpha \\ \alpha & \cdots & \alpha & \alpha & \alpha & 1 \end{pmatrix}$

- Konzistentní odhad parametru $\boldsymbol{\alpha}$: $\hat{\boldsymbol{\alpha}} = \frac{1}{\hat{\varphi}} \frac{1}{....} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \Sigma\, r_{ij}^{P} \cdot r_{i,}^{P}$

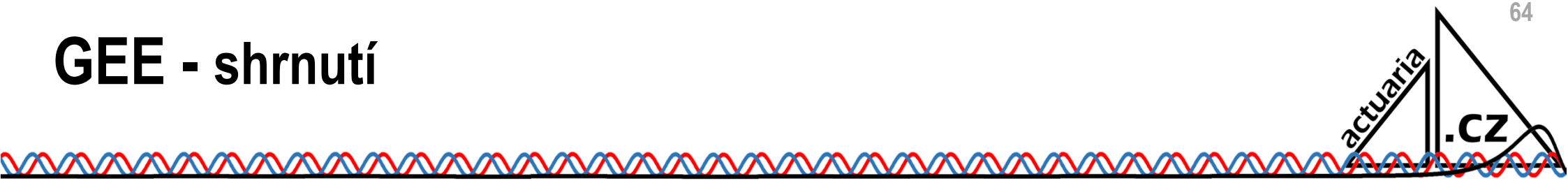
GEE – volba korelační struktury metoda parametrizované korelace
Příklady volby korelační struktury



- **AR(1) correlation**

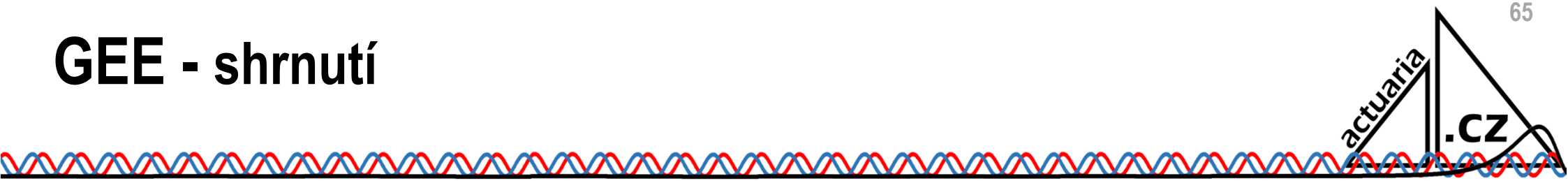
- $Y_{i1}, \dots, Y_{in_i}$ jsou z AR(1) časové řady
- $cor(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}$

GEE - shrnutí



- GEE metoda vhodná pro regresní analýzu dat rozdělených do K nezávislých skupin. Uvnitř skupin jsou data vzájemně korelovaná.
- Není nutně přesně určit rozdělení dat Y, ani jejich rozptyl či korelační strukturu uvnitř skupin.
- Regresní parametry β mají „population-average“ interpretaci

GEE - shrnutí



- GEE metoda není založena na věrohodnosti, pouze na kvazi-věrohodnosti. Pro analýzu modelu proto nelze použít statistiky na bázi věrohodnosti (tj. Deviance, AIC, BIC).
- Existují analogické statistiky na bázi kvazi-věrohodnosti (QIC (Quasilikelihood under the Independence model Information Criterion), QIC_{HH}, CIC(Correlation Information Criterion), CIC_{HH}), více viz Hudecová&Pešta, , podrobněji Hardin&Hilbe. Je rovněž možné použít skorové testy. (dj-h)
- GEE – ilustrace
  - Příklad: Vehicle Insurance Claims – DeJong, Heller: Logistická regrese s korelovanými pozorováními, longitudinální data, navrhovaná korelační struktura AR(1).
- Využití GEE metod pro rezervování škod. Možnost modelování závislosti vývoje škodních trojúhelníků v jednotlivých letech – viz. Hudecová&Pešta, Gerthofer

REGRESNÍ MODELY

Lineární smíšené modely

Linear Mixed (Effect) Models

LMM

LLM

- Odezvy $Y_1,\ldots, Y_K$ splňují **jednoúrovňový model LMM**, pokud platí:
  - $\boldsymbol{Y_i} = \mathbb{X}_i\boldsymbol{\beta_i} + \mathbb{Z}_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}, i = 1, \ldots, K$
  - $\boldsymbol{b_i}$ (náhodné efekty)…nezávislé vektory t.ž. $\boldsymbol{b_i} \sim \mathrm{N}_q(\mathbf{0}, \mathbb{D})$,
  - $\boldsymbol{\varepsilon_i}$ … náhodné vektory, t.ž.: $\boldsymbol{\varepsilon_i} \sim \mathrm{N}_{n_i}(\mathbf{0}, \sigma_e^2\, \mathbb{I}_{n_i})$,
  - $\mathbb{X}_i$… regresní matice pro pevné efekty $\boldsymbol{\beta_i}$
  - $\mathbb{Z}_i$… regresní matice pro náhodné efekty $\boldsymbol{b_i}$
- pevná složka: $\mathbb{X}_i\boldsymbol{\beta_i}$
- náhodná složka: $\mathbb{Z}_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}$, $\mathbb{E}[\mathbb{Z}_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}] = \mathbf{0}$, $\mathrm{var}[\mathbb{Z}_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}] = \mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^{\boldsymbol{T}} + \sigma_e^2\, \mathbb{I}_{n_i}$
- (neznámé) parametry modelu MLE: $\boldsymbol{\beta}$, σ_e^2 a $\mathbb{D}$ (kovariantní matice, symetrická a pozitivně definitní)
- Matice $\mathbb{D}$ se nahrazuje maticí Δ, t.ž. $\Delta^{\boldsymbol{T}}\Delta = \sigma_e^2\mathbb{D}^{-1}$

LLM

- Odezvy $Y_1,\ldots, Y_K$ splňují **jednoúrovňový model LMM**, pokud platí:
  - $\boldsymbol{Y_i} = \mathbb{X}_i\boldsymbol{\beta_i} + \mathbb{Z}_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}, i = 1, \ldots, K$
  - $\boldsymbol{b_i}$ (náhodné efekty)…nezávislé vektory t.ž. $\boldsymbol{b_i} \sim \mathrm{N}_q(\mathbf{0}, \mathbb{D})$,
  - $\boldsymbol{\varepsilon_i}$ … náhodné vektory, t.ž.: $\boldsymbol{\varepsilon_i} \sim \mathrm{N}_{n_i}(\mathbf{0}, \sigma_e^2\, \mathbb{I}_{n_i})$,

- **Marginální tvar** - alternativní zápis modelu LMM: Marginální tvar:
  - $\boldsymbol{Y_i} \sim \mathrm{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta}, \sigma_e^2\, \Sigma_i)$, kde $\Sigma_i = \mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^{\boldsymbol{T}}/\sigma_e^2 + \mathbb{I}_{n_i}$
    - nebo
  - $\boldsymbol{Y} \sim \mathrm{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma_e^2\, \Sigma)$

LLM – odhady parametrů

- Marginální tvar modelu LMM:
- $\boldsymbol{Y_i} \sim \mathrm{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta}, \sigma_e^2\,\Sigma_i)$,
  - kde $\Sigma_i = \mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^{\boldsymbol{T}}/\sigma_e^2 + \mathbb{I}_{n_i}$ modelu je funkcí parametrů β a θ.
- 1. Marginální věrohodnost:
- Logaritmická věrohodností funkce: $l_n(\beta, \theta, \sigma_e^2)$
  - Minimalizujeme $l_n(\beta, \theta, \sigma_e^2)$ vzhledem k β pro dané θ:
  - řešení metodou vážených nejmenších čtverců
    - řešení: $\hat{\beta}(\theta) = \left(\sum_{i=1}^{K} \mathbb{X}_i{}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbb{X}_i\right)^{-1} \sum_{i=1}^{K} \mathbb{X}_i{}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}_i$
  - Minimalizujeme $l_n(\hat{\beta}(\theta), \theta, \sigma_e^2)$ vzhledem k σ_e^2 pro dané θ:
    - řešení: $\hat{\sigma}_e^2(\theta) = \frac{1}{n}\sum_{i=1}^{K}\left(\boldsymbol{Y_i} - \mathbb{X}_i\hat{\beta}(\theta)\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y_i} - \mathbb{X}_i\hat{\beta}(\theta)\right)$
  - Maximalizujeme profilovou věrohodnost: $l_n{}^*(\theta) = l_n(\hat{\beta}(\theta), \theta, \hat{\sigma}_e^2(\theta))$ přes θ:
    - řešení: $\hat{\theta}$
  - Toto je však těžko řešitelné, aplikujeme jiný přístup využívající strukturu náhodných efektů a dekompozici věrohodnosti

LLM – odhady parametrů
Hendersonovy rovnice pro smíšený model

- Obecný tvar pro smíšený model:
- $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \mathbb{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}$
  - kde $\boldsymbol{b} \sim \mathrm{N}_{q^*}(\mathbf{0}, \mathbb{D}_*)$, $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Lambda}_*)$, $cov(b, \varepsilon) = 0$, $\Sigma = var(Y) = \mathbb{Z}\,\mathbb{D}_*\mathbb{Z}^{\mathrm{T}} + \boldsymbol{\Lambda}_*$.
- jednoúrovňový model LMM je speciálním případem. $q^* = Kq$, $n = \sum_{i=1}^{K} n_i$, …, $\boldsymbol{\Lambda}_* = \sigma_e^2 \mathbb{I}_n$
- Sdruženou hustotu $(Y, b) = f(y, b; \beta) = f(y|b; \beta) f(b)$ považujme za věrohodností funkci neznámých parametrů (β, b), maximalizujeme současně přes β i b, odhady $(\hat{\beta}, \hat{b})$, jako řešení Hendersonových rovnic pro smíšený model
- $\hat{\beta} = (\mathbb{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbb{X})^{-1}(\mathbb{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}Y)$
- $\hat{b} = \mathbb{D}_*\mathbb{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbb{X}\hat{\beta})$

LLM – odhady parametrů
Hendersonovy rovnice pro smíšený model

- Odhady regresních parametrů pomocí Hendersonových rovnic pro smíšený model:
  - $\hat{\beta} = (\mathbb{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbb{X})^{-1}(\mathbb{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}Y)$
  - $\hat{b} = \mathbb{D}_{*}\mathbb{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbb{X}\hat{\beta})$
- Pro jednoúrovňový LMM model dostáváme
  - $\hat{b} = \mathbb{D}\mathbb{Z}_i^{\boldsymbol{T}}(\mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^{\boldsymbol{T}} + \sigma_e^2\mathbb{I}_{n_i})^{-\mathbf{1}}(\boldsymbol{Y_i} - \mathbb{X}_i\hat{\beta})$
- $\hat{\beta}$ je BLUE (nejlepší lineární nestranný odhad) a konzistentní odhad parametru β bez ohledu na rozdělení Y
- $\hat{b}$ je BLUP (nejlepší lineární nestranný prediktor) pro b.

LLM – odhady parametrů
Hendersonovy rovnice pro smíšený model

- Odhady rozptylů regresních parametrů pomocí Hendersonových rovnic pro jednoúrovňový smíšený model:
  - $var\,(\hat{\beta}) = \left(\sum_{i=1}^{K} \mathbb{X}_i^{\mathrm{T}}\, \boldsymbol{\Sigma_i}^{-1} \mathbb{X}_i\ \right)^{-1}$
  - $var\,(\widehat{b_i}) = \mathbb{D}\mathbb{Z}_i^{\boldsymbol{T}} \left[\, \boldsymbol{\Sigma_i}^{-1} - \boldsymbol{\Sigma_i}^{-1} \mathbb{X}_i \quad \left(\sum_{i=1}^{K} \mathbb{X}_i^{\mathrm{T}}\, \boldsymbol{\Sigma_i}^{-1} \mathbb{X}_i\ \right)^{-1} \mathbb{X}_i^{\mathrm{T}}\, \boldsymbol{\Sigma_i}^{-1} \right] \mathbb{Z}_i\, \mathbb{D}$
  - $var\,(\widehat{b_i} - \hat{b}) = \mathbb{D} - var\,(\widehat{b_i})$

- Odhady rozptylů regresních parametrů lze řešit i pomocí REML metody (Restricted Maximum Likelihood Estimators), výsledky se mírně liší od věrohodnostních odhadů.
- Pro velké počty nezávislých skupin K je rozdíl zanedbatelný. Asymptoticky jsou výsledky shodné.

GEE x LMM

- **LMM:**
  - poskytuje detailní model pro *var Y*
  - Předpokládá Normalitu b_i a ε_i
  - Poskytuje informaci o struktuře rozptylu
    - Dekompozice
    - Odhady složek rozptylu
    - Náhodné efekty
    - Testy hypotéz o struktuře rozptylu
  - Pokud struktura rozptylu NENÍ dobře specifikovaná, závěry o pevných efektech β (testy, intervaly spolehlivosti) jsou NEPLATNÉ
  - Vyžaduje velké K (počet subjektů, nezávislých skupin)
- **GEE:**
  - používá „pracovní model“ pro *var Y,* o kterém se však nepředpokládá, že je správný
  - Neklade žádný předpoklad o rozdělení *Y*
  - Neposkytuje dostatek informací o struktuře rozptylu
  - Pokud je K (počet subjektů, nezávislých skupin) dost velké, závěry o βjsou platné i když pracovní struktura rozptylu není správná
  - Vyžaduje velké K (počet subjektů, nezávislých skupin), pro malé selhává

REGRESNÍ MODELY

Zobecněné lineární smíšené modely

Generalised Linear Mixed Models

GLMM

GLMM

- GLMM - další možnost jak analyzovat data, u nichž je **porušen předpoklad nezávislosti / nekorelovanosti**.
- GLMM lze pohlížet jako na **zobecnění Lineárních smíšených modelů (LMM**), kdy rozšíříme skupinu rozdělení, z nichž pochází odezva/vysvětlovaná proměnná Y.
- GLMM lze pohlížet jako na **zobecnění Zobecněných lineárních modelů GLM**, kdy do modelu kromě pevných regresních parametrů β (fixed effects, pevné efekty), **zavedeme další prvek náhodnosti pomocí náhodných efektů $\boldsymbol{b}$** (random effects).

GLMM - DEFINICE MODELU pro skupinově závislá data

- **1. Data**: $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_K$ nezávislé náhodné vektory $\boldsymbol{Y}_i = \left(Y_{i1}, \dots, Y_{in_i}\right)^{\mathrm{T}}, i = 1 \dots K,\ \sum_{\mathrm{i=1}}^{\mathrm{K}} n_i = n$
- **2.** náhodné efekty: $\boldsymbol{b_i}$ … nezávislé q-rozměrné vektory s hustotou $h(b;\psi)$
- Složky $Y_{i1}, \dots, Y_{in_i}$ vektoru $\boldsymbol{Y}_i$ jsou při daném $\boldsymbol{b_i}$ podmíněně nezávislé, podmíněná hustota $\in EDF$
- $f(y|\boldsymbol{b_i}) = exp\left\{\frac{y\theta_{ij} - b(\theta_{ij})}{\varphi} + c(y,\varphi)\right\}$
- 3. θ_{ij}…kanonický parametr, závisí na:
  - (p-rozměrných) regresorech pro pevné efekty… X_{ij}
  - pevných regresních parametrech $\beta \in \mathbb{R}$
  - (q-rozměrných) regresorech pro náhodné efekty… $Z_{ij}, q \leq p$
  - náhodných efektech b_i prostřednictvím
  - Lineárního prediktoru $\eta_{ij} = \boldsymbol{X_{ij}^{\mathrm{T}}\beta} + \boldsymbol{Z_{ij}^{\mathrm{T}}\, b_i}$
- 4. g …linková funkce, platí $g(\mu_{ij}) = \eta_{ij}$

GLMM - DEFINICE MODELU pro skupinově závislá data

- Data: $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K$ nezávislé náhodné vektory $\boldsymbol{Y}_i = \left(Y_{i1}, \ldots, Y_{in_i}\right)^{\mathrm{T}}, i = 1 \ldots K, \ \sum_{\mathrm{i=1}}^{\mathrm{K}} n_i = n$
- Náhodné efekty: $\boldsymbol{b_i}$ … nezávislé q-rozměrné vektory s hustotou $h(b;\psi)$
- Podmíněná hustota $f(y|\boldsymbol{b_i}) = exp\left\{\frac{y\theta_{ij} - b(\theta_{ij})}{\varphi} + c(y,\varphi)\right\}$
- Lineární prediktor $\eta_{ij} = \boldsymbol{X_{ij}^{\mathrm{T}}\beta} + \boldsymbol{Z_{ij}^{\mathrm{T}}\, b_i}$
- Linková funkce $g(\mu_{ij}) = \eta_{ij}$
- $\mathbb{E}\left[Y_{ij}|b_i\right] = \mu_{ij} = b'\left(\theta_{ij}\right) = g^{-1}(\eta_{ij}) = g^{-1}(\boldsymbol{X_{ij}^{\mathrm{T}}\beta} + \boldsymbol{Z_{ij}^{\mathrm{T}}\, b_i})$
- $var\left[Y_{ij}|b_i\right] = \varphi\, b''\left(\theta_{ij}\right) = \varphi\, V\left(\mu_{ij}\right)$

GLMM - DEFINICE MODELU pro skupinově závislá data

- Při daném $\boldsymbol{b_i}$, Y_{ij} podmíněně splňují **GLM** model,
  - podmíněná hustota $f(y|\boldsymbol{b_i}) \in EDF$
- stejně jako **LMM** modelu, má zavedení náhodných efektů $\boldsymbol{b_i}$ do všech lineárních prediktorů za následek korelaci mezi $Y_{i1}, \dots, Y_{in_i}$.

- $\mathbb{E}[Y_{ij}|b_i] = g^{-1}(\boldsymbol{X_{ij}^{\mathrm{T}}\beta} + \boldsymbol{Z_{ij}^{\mathrm{T}}}\, \boldsymbol{b_i})$
- $\mathbb{E}[Y_{ij}] = \mathbb{E}\left[\mathbb{E}[Y_{ij}|b_i]\right] = \mathbb{E}[g^{-1}(\boldsymbol{X_{ij}^{\mathrm{T}}\beta} + \boldsymbol{Z_{ij}^{\mathrm{T}}}\, \boldsymbol{b_i})]$

GLMM – odhady parametrů

- Vzhledem k předpokladu podmíněného rozdělení odezvy/závislých proměnných, je možné pro odhady parametrů použít metodu maximální věrohodnosti
- Věrohodností funkce $l(\beta, \varphi \mid Y)$ však nemá explicitní řešení
- Je nutné řešit aproximačními metodami
- jedním z možných přístupů je použití Laplaceovy aproximace, která zjednoduší tvar věrohodnostní funkce a umožní odhady parametrů následujícím způsobem:
  - Odhad β: modifikovaný IWLS algoritmus
  - Odhad $\boldsymbol{b_i}$: Hendersonovy rovnice
  - Odhad φ a ϕ: momentová metoda

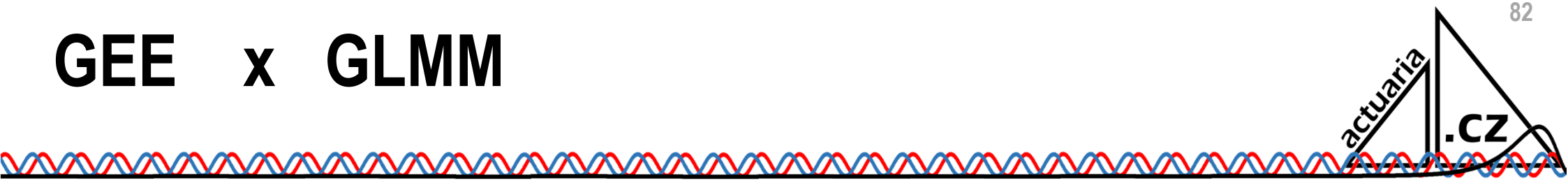
GLMM

- podmíněný model:
  - $\mathbb{E}[Y_{ij}|b_i] = g^{-1}(\boldsymbol{X}_{\boldsymbol{ij}}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{Z}_{\boldsymbol{ij}}^{\mathrm{T}}\,\boldsymbol{b}_{\boldsymbol{i}})$
  - $\boldsymbol{\beta}$ vyjadřuje podmíněný efekt $\boldsymbol{X}_{\boldsymbol{ij}}$na $\mathbb{E}[Y_{ij}]$ vzhledem k $\boldsymbol{b}_{\boldsymbol{i}}$
  - tzv: subject-specific efekty
  - Popisují vliv na $\mathbb{E}[Y_{ij}|b_i]$, když daný subjekt/jedinec mění hodnotu $\boldsymbol{X}_{\boldsymbol{ij}}$
  - Parametry $\boldsymbol{\beta}$ obecně nemohou porovnat dva různé subjekty, které se liší v hodnotě$\boldsymbol{X}_{\boldsymbol{ij}}$
- marginální model:
  - $\mathbb{E}[Y_{ij}] = \mathbb{E}\left[\mathbb{E}[Y_{ij}|b_i]\right] = \mathbb{E}[g^{-1}(\boldsymbol{X}_{\boldsymbol{ij}}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{Z}_{\boldsymbol{ij}}^{\mathrm{T}}\,\boldsymbol{b}_{\boldsymbol{i}})]$
  - Nepodmíněný efekt $\boldsymbol{X}_{\boldsymbol{ij}}$na $\mathbb{E}[Y_{ij}]$ vzhledem k $\boldsymbol{b}_{\boldsymbol{i}}$
  - tzv: population-average efekty
  - $\mathbb{E}[Y_{ij}]$ obecně nesplňuje předpoklady GLM , parametry $\boldsymbol{\beta}$ (v podmíněném modelu) obecně nemají population-average interpretaci.

REGRESNÍ MODELY

GEE x GLMM

GEE x GLMM



- GLM analyzuje nezávislá data
- V případě korelovaných dat GLM nedávají dobré výsledky
- Řešením může být použití modelů GLMM nebo GEE, která pracují s korelovanými daty
- Jak zvolit mezi GLMM a GEE?
- Dle požadované interpretace parametrů

GEE x GLMM

- GLMM: Conditional Model
- GEE: Marginal Model

- GLMM: Subject Specific interpretace:
- GEE: Population Average interpretace

- GLMM: regresní koeficienty se vztahují na každého jednotlivce (subjekt), nikoli však nutně na celou populaci,
- GEE: regresní koeficienty se vztahují na celou populaci, nemusí však platit pro každého jednotlivce

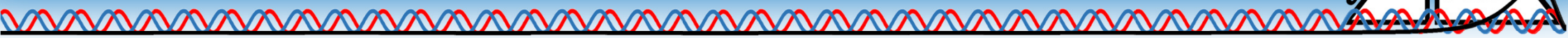
- GLMM odhaduje jiné parametr než GEE. Pokud oba modely mají stejnou linkovou funkci -> aspoň jeden model není správný (výjimky)
- Pokud je GLMM dobrý model, obvykle je i GEE dobrý, má však výrazně odlišné odhady parametrů

GEE x GLMM

- GLMM odhaduje jiné parametr než GEE. Pokud oba modely mají stejnou linkovou funkci -> aspoň jeden model není správný (výjimky)
- Pokud je GLMM dobrý model, obvykle je i GEE dobrý, má však výrazně odlišné odhady parametrů.
  - log-lineární model: liší se pouze v absolutním členu (intercept)
  - Logistický model: liší se i směrnice (slope)

- GLMM: závislost modeluje pomocí náhodných efektů přidaných do GLM modelu
- GEE: nepředpokládá žádné konkrétní rozdělení odezvy, stejně jako GLM vyžaduje specifikaci: linkové funkce a lineárního prediktoru. Místo konkrétního rozdělení proměnných však stačí specifikovat vztah mezi Střední hodnotou a kovariancí (analogie rozptylové funkce, v případě kvazivěrohodnosti v GLM), resp. Specifikovat „pracovní varianci“
- GEE: závislostní struktura je modelována pomocí „pracovní“ kovarianční matice, která nemusí odpovídat skutečné závislosti. Doporučuje se používat jednoduchou korelační strukturu

REGRESNÍ MODELY



A CO DÁL?

REFERENCE

- P. de Jong, G. Z. Heller: *Generalized Linear Models for Insurance Data.* Cambridge University Press, 2008.
- E. Ohlsson, B. Johansson: *Non-Life Insurance Pricing with Generalized Linear Models.* EAA Series, Springer-Verlag Berlin Heidelberg, 2010.
- W. N. Venables, B. D. Ripley: *Modern Applied Statistics with S.* 4th edition. Springer 2002
- S.Wood: Generalized Additive Models, An Introduction with R, Chapman & Hall/CRC Press, 2006
- C.E. McCulloch, S.R. Searle: *Generalized, Linear, and Mixed Models.* Wiley Series in Probability and Statistics, Wiley 2001
- J. W. Hardin, J. M. Hilbe: *Generalized Estimating Equations.* Chapman & Hall/CRC, 2003
- A. Agresti: *An Introduction to Categorical Data Analysis.* Wiley, 2007
- M. Branda: *Zobecněné lineární modely v pojišťovnictví.* MFF UK 2013
- Š. Hudecová, M. Pešta: *Modeling Dependencies in Claims Reserving with GEE.* MFF UK 2003
- M. Gerthofer: *Claims reserving within the panel data Framework.* MFF UK 2015, diplomová práce
- Poznámky k přednášce: *Pokročilé regresní modely / Advanced Regression Models* (NMST432), 2015, MFF UK, přednášející Doc. Mgr. Michal Kulich, Ph.D.
- Poznámky k přednášce: *Matematika neživotního pojištění* (NMFM402), 2014, MFF UK, přednášející RNDr. Lucie Mazurová, Ph.D.
- Poznámky k přednášce: *Vybraný software pro finance a pojišťovnictví / Selected Software Tools for Finance and Insurance* (NMFM404), 2014, MFF UK, přednášející RNDr. Michal Pešta, Ph.D.
- Poznámky ze semináře České společnosti aktuárů: *Zobecněné lineární modely (GLM) v pojišťovnictví*, 2012, přednášející Ing. Pavel Zimmernann, Ph.D.
- Poznámky ze semináře České společnosti aktuárů: *Aplikované modely storen*, 2015,
- https://onlinecourses.science.psu.edu/statprogram/stat504 - Analysis of Discrete Data