

M7988 Modely ztrát v neživotním pojištění

Parametrický model a úlohy matematické statistiky

Model: X_1, \dots, X_n je náhodný výběr z rozdělení s distribuční funkcí $F(x, \theta)$. Tuto distribuční funkci známe až na neznámý parametr $\theta \in \Theta \subset \mathbb{R}^p$.

Úlohy matematické statistiky:

- Bodový odhad parametru θ .
- Intervalový odhad parametru θ .
- Testy hypotéz o parametru θ .

Někdy nás místo samotného parametru θ zajímá nějaká jeho funkce, tzv. *parametrická funkce* $\gamma(\theta)$, kde $\gamma : \Theta \rightarrow \Theta^* \subset \mathbb{R}$ je reálná funkce.

Dále budeme uvažovat jednorozměrný parametr $\theta \in \Theta \subset \mathbb{R}$.

Bodové odhady

Řekneme, že $T : \mathbb{R}^n \rightarrow \mathbb{R}$ je bodový odhad parametru $\theta \in \Theta \subset \mathbb{R}$, jestliže T je měřitelnou funkcí náhodného výběru X_1, \dots, X_n . Tedy $T = T(X_1, \dots, X_n)$ je náhodná veličina.

Vlastnosti bodových odhadů:

- T je nestranný odhad parametru θ , jestliže $\mathbb{E}T = \theta$ pro všechna $\theta \in \Theta$.
- T je asymptoticky nestranný odhad parametru θ , jestliže $\lim_{n \rightarrow \infty} \mathbb{E}T = \theta$ pro všechna $\theta \in \Theta$.
- T je konzistentní odhad parametru θ , jestliže $T = T(X_1, \dots, X_n) \rightarrow \theta$ v pravděpodobnosti pro $n \rightarrow \infty$ pro všechna $\theta \in \Theta$.

Který odhad je nejlepší?

- Necht' T_1, T_2 jsou dva nestranné odhady parametru θ . Řekneme, že T_1 je více eficientní (*lepší*) než T_2 , jestliže $DT_1 \leq DT_2$ pro všechna $\theta \in \Theta$.
- Necht' T je nestranný odhad parametru θ . Řekneme, že T je nejlepší nestranný odhad parametru θ , jestliže $DT \leq DT^*$ pro všechna $\theta \in \Theta$ a pro všechny nestranné odhady T^* .
- Necht' T je odhad parametru θ . Střední čtvercovou (kvadratickou) chybu odhadu definujeme jako $\text{MSE}(T) = \mathbb{E}(T - \theta)^2$.
- Je-li T je nestranný odhad parametru θ , pak $\text{MSE}(T) = DT$.
- Necht' T_1, T_2 jsou dva odhady parametru θ . Řekneme, že T_1 je více eficientní (*lepší*) než T_2 , jestliže $\text{MSE}(T_1) \leq \text{MSE}(T_2)$ pro všechna $\theta \in \Theta$.
- (Stejněměně) nejlepší odhad parametru θ neexistuje.

Metoda momentů

- Dále předpokládejme, že neznámý parametr θ je p -rozměrný ($\Theta \subset \mathbb{R}^p$).
- Necht' existují obecné momenty $\mu'_k = \mu'_k(\theta) = \mathbb{E}X_1^k$ pro $k = 1, \dots, p$.
- Označme jejich výběrové protějšky $M'_k = \frac{1}{n} \sum_{i=1}^n X_i^k$ pro $k = 1, 2, \dots$
- Řekneme, že $\tilde{\theta}$ je odhad parametru θ metodou momentů, jestliže $\mu'_k(\tilde{\theta}) = M'_k$ pro $k = 1, \dots, p$.
- Je-li řešení předchozí soustavy nejednoznačné (rovnice jsou lineárně závislé), přidáme další rovnici pro $k = p + 1$, pokud ovšem existuje příslušný moment.

Metoda maximální věrohodnosti

- Označme sdruženou hustotu náhodného vektoru $(X_1, \dots, X_n)'$ jako

$$L(\theta) = \prod_{i=1}^n f(x_i, \theta).$$

- $f(x, \theta)$ je hustota (pravděpodobnostní hustota, pravděpodobnostní funkce) náhodné veličiny X_i .
- $L(\theta) = L(\theta, x_1, \dots, x_n)$ se nazývá věrohodnostní funkce.
- $\hat{\theta}$ se nazývá maximálně věrohodným odhadem parametru θ , jestliže $L(\hat{\theta}) \geq L(\theta), \forall \theta \in \Theta$.
- $\hat{\theta} = \arg \max\{L(\theta); \theta \in \Theta\}$.
- $l(\theta) = \log L(\theta) = \sum_{i=1}^n \log f(x_i, \theta)$ se nazývá logaritmická věrohodnostní funkce.
- $\hat{\theta} = \arg \max\{l(\theta); \theta \in \Theta\}$.

Regulární systém hustot

Řekneme, že systém hustot $\{f(x, \theta), \theta \in \Theta\}$ je regulární, jestliže

- 1 $\Theta \subset \mathbb{R}^p$ je otevřená borelovská množina.
- 2 Množina $M = \{x \in \mathbb{R} : f(x, \theta) > 0\}$ nezávisí na hodnotě parametru θ .
- 3 Pro všechna $x \in M$ existuje konečná parciální derivace

$$f'_i(x, \theta) = \frac{\partial f(x, \theta)}{\partial \theta}, \quad i = 1, \dots, p.$$

- 4 Pro všechna $\theta \in \Theta$ a všechna $i = 1, \dots, p$ platí

$$\int_M f'_i(x, \theta) dx = 0.$$

- 5 Pro všechna $\theta \in \Theta$ a pro každou dvojici (i, j) existuje konečný integrál

$$J_{i,j}(\theta) = \int_M \frac{f'_i(x, \theta) f'_j(x, \theta)}{f_i^2(x, \theta)} f_i(x, \theta) dx$$

- 6 Matice $J(\theta) = (J_{i,j}(\theta))_{i,j=1}^p$ je pozitivně definitní pro všechna $\theta \in \Theta$.

Vlastnosti maximálně věrohodných odhadů

- Necht' systém hustot $\{f(x, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ je regulární, pak maximálně věrohodný odhad parametru $\boldsymbol{\theta}$ je asymptoticky nestranný, konzistentní a má asymptoticky normální rozdělení.
- $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ má asymptoticky rozdělení $\mathcal{N}_p(\mathbf{0}, \mathbf{J}^{-1}(\boldsymbol{\theta}))$.

$$\mathbf{J}(\boldsymbol{\theta}) = \left(\mathbb{E} \frac{\partial \log f(X_1, \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(X_1, \boldsymbol{\theta})}{\partial \theta_j} \right)_{i,j=1}^p$$

je Fisherova informační matice o parametru $\boldsymbol{\theta}$ příslušná X_1 .

- $$\mathbf{J}(\boldsymbol{\theta}) = -\mathbb{E} \left(\frac{\partial^2 \log f(X_1, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right)_{i,j=1}^p .$$
- $\hat{\boldsymbol{\theta}} \approx \mathcal{N}_p(\boldsymbol{\theta}, \frac{1}{n} \mathbf{J}^{-1}(\boldsymbol{\theta}))$.

Maximálně věrohodné odhady pro parametrickou funkci γ

- Necht' $\gamma : \Theta \rightarrow \Theta^*$ je parametrická funkce.
- Funkci $\tilde{L}(\theta^*) = \sup\{L(\theta); \theta \in \Theta : \gamma(\theta) = \theta^*\}$ pro $\theta^* \in \Theta^*$ nazveme věrohodnostní funkcí indukovanou parametrickou funkcí γ .
- $\hat{\theta}^*$ je maximálně věrohodný odhad parametrické funkce $\gamma(\theta) = \theta^*$, jestliže $\tilde{L}(\hat{\theta}^*) \geq \tilde{L}(\theta^*)$ pro všechna $\theta^* \in \Theta^*$.
- Zehnaova věta (princip invariance MLE): Je-li $\hat{\theta}$ maximálně věrohodný odhad parametru θ , pak $\gamma(\hat{\theta})$ je maximálně věrohodný odhad parametrické funkce $\gamma(\theta)$.

Delta metoda

Theorem

Nechť $\{\mathbf{X}_n\}_{n=1}^{\infty}$ je posloupnost p -rozměrných náhodných vektorů takových, že $\sqrt{n}(\mathbf{X}_n - \boldsymbol{\theta})$ má asymptoticky normální rozdělení $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Dále buď $\gamma : \mathbb{R}^p \rightarrow \mathbb{R}$ měřitelná funkce, která má totální diferenciál v bodě $\boldsymbol{\theta}$. Pak platí: $\sqrt{n}(\gamma(\mathbf{X}_n) - \gamma(\boldsymbol{\theta}))$ má asymptoticky normální rozdělení $\mathcal{N}(0, \sigma^2)$, kde $\sigma^2 = \nabla \gamma'(\boldsymbol{\theta}) \boldsymbol{\Sigma} \nabla \gamma(\boldsymbol{\theta})$, kde

$$\nabla \gamma(\boldsymbol{\theta}) = \left(\frac{\partial \gamma(\boldsymbol{\theta})}{\partial \theta_1}, \dots, \frac{\partial \gamma(\boldsymbol{\theta})}{\partial \theta_p} \right)'$$

je gradient funkce γ v bodě $\boldsymbol{\theta}$.

Aplikace na MLE

- Již víme, že je-li systém hustot $\{f(x, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ regulární, pak: $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ má asymptoticky rozdělení $\mathcal{N}_p(\mathbf{0}, \mathbf{J}^{-1}(\boldsymbol{\theta}))$.
- Aplikací delta metody dostaneme, že za splnění podmínek regularity platí:

$\sqrt{n}(\gamma(\hat{\boldsymbol{\theta}}) - \gamma(\boldsymbol{\theta}))$ má asymptoticky $\mathcal{N}(0, \nabla\gamma'(\boldsymbol{\theta})\mathbf{J}^{-1}(\boldsymbol{\theta})\nabla\gamma(\boldsymbol{\theta}))$.

- $$\gamma(\hat{\boldsymbol{\theta}}) \approx \mathcal{N}(\gamma(\boldsymbol{\theta}), \frac{1}{n}\nabla\gamma'(\boldsymbol{\theta})\mathbf{J}^{-1}(\boldsymbol{\theta})\nabla\gamma(\boldsymbol{\theta})).$$
- Je-li θ jednorozměrný parametr, pak $\gamma(\hat{\theta}) \approx \mathcal{N}\left(\gamma(\theta), \frac{[\gamma'(\theta)]^2}{nJ(\theta)}\right)$.

Intervalová data

- Nepozorujeme přímo hodnoty náhodného výběru X_1, \dots, X_n .
- O každém pozorování víme jen to, do kterého intervalu patří.
- Obor hodnot náhodného výběru je rozdělen na intervaly $(c_0, c_1], (c_1, c_2], \dots, (c_{k-1}, c_k]$.
- c_0 může být i $-\infty$ a c_k může být i ∞ .
- Označme n_j počet pozorování, která leží v intervalu $(c_{j-1}, c_j]$.
- $n = n_1 + \dots + n_k$.

Empirická distribuční funkce pro intervalová data

- Standardně pro hodnoty náhodného výběru X_1, \dots, X_n definujeme empirickou distribuční funkci

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{X_i \leq x\}}.$$

- Podle předchozí definice můžeme určit její přesné hodnoty v bodech c_0, c_1, \dots, c_k :

$$\hat{F}_n(c_0) = 0, \quad \hat{F}_n(c_j) = \frac{n_1 + \dots + n_j}{n}, \quad \hat{F}_n(c_k) = 1.$$

- Mezi těmito body empirickou distribuční funkci spojitě dodefinujeme, například lomennou čarou (ogive).
-

$$\hat{F}_n(x) = \begin{cases} 0, & x \leq c_0, \\ \frac{c_j - x}{c_j - c_{j-1}} \hat{F}_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} \hat{F}_n(c_j), & c_{j-1} < x < c_j, \\ 1, & x \geq c_k. \end{cases}$$

Histogram pro intervalová data

- Histogram je po částech konstantní odhad hustoty původních veličin X_1, \dots, X_n .
- $f_n(x) = \widehat{F}'_n(x)$ pro všechna $x \neq c_0, \dots, c_k$.

-

$$f_n(x) = \frac{n_j}{n} \cdot \frac{1}{c_j - c_{j-1}}, \quad c_{j-1} < x < c_j.$$

- V bodech c_0, \dots, c_k můžeme definovat libovolně.
- Hodnota histogramu v intervalu $(c_{j-1}, c_j]$ je rovna relativní četnosti pozorování v daném intervalu dělená délkou tohoto intervalu.

Empirická kvantilová funkce pro intervalová data

- Můžeme ji určit jako inverzní funkci k empirické distribuční funkci (ogive), tj. $\hat{Q}_n(\alpha) = \hat{F}_n^{-1}(\alpha)$ pro $0 < \alpha < 1$.
- Nebo opět určit přesné hodnoty v bodech $0, \frac{n_1}{n}, \frac{n_1+n_2}{n}, \dots, 1$:

$$\hat{Q}_n(0) = c_0, \quad \hat{Q}_n(\alpha_j) = c_j, \quad \hat{Q}_n(1) = c_k,$$

kde $\alpha_j = \frac{n_1 + \dots + n_j}{n}$.

- Mezi těmito body empirickou kvantilovou funkci opět spojitě dodefinujeme, například lomennou čarou.
-

$$\hat{Q}_n(\alpha) = \frac{\alpha_j - \alpha}{\alpha_j - \alpha_{j-1}} \hat{Q}_n(\alpha_{j-1}) + \frac{\alpha - \alpha_{j-1}}{\alpha_j - \alpha_{j-1}} \hat{Q}_n(\alpha_j), \quad \alpha_{j-1} < \alpha < \alpha_j.$$

Metoda maximální věrohodnosti pro intervalová data

- Necht' každá z nepozorovaných náhodných veličin X_1, \dots, X_n má distribuční funkci $F(x, \theta)$, kde $\theta \in \Theta$ je neznámý parametr.
- Pravděpodobnost, že dané pozorování X_i leží v intervalu $(c_{j-1}, c_j]$ je

$$P(X_i \in (c_{j-1}, c_j]) = F(c_j, \theta) - F(c_{j-1}, \theta).$$

- Věrohodnostní funkce pro naše data je:

$$L(\theta) = \prod_{i=1}^n P(X_i \in (c_{j-1}, c_j]) = [F(c_1, \theta) - F(c_0, \theta)]^{n_1} \cdot [F(c_2, \theta) - F(c_1, \theta)]^{n_2} \cdot \dots \cdot [F(c_k, \theta) - F(c_{k-1}, \theta)]^{n_k} = \prod_{j=1}^k [F(c_j, \theta) - F(c_{j-1}, \theta)]^{n_j}.$$

- $l(\theta) = \log L(\theta) = \sum_{j=1}^k n_j \log[F(c_j, \theta) - F(c_{j-1}, \theta)]$ je logaritmická věrohodnostní funkce pro naše data.
- $\hat{\theta} = \arg \max\{l(\theta); \theta \in \Theta\}$.

Metoda minimálního χ^2

- Označme $p_j(\theta)$ teoretickou pravděpodobnost, že náhodná veličina X_i nabude hodnoty z intervalu $(c_{j-1}, c_j]$, tedy

$$p_j(\theta) = P(X_i \in (c_{j-1}, c_j]) = F(c_j, \theta) - F(c_{j-1}, \theta).$$

- Dohromady máme celkem n pozorování. V intervalu $(c_{j-1}, c_j]$ by tedy mělo být $np_j(\theta)$ pozorování.
- Porovnejme očekávaný a skutečný (pozorovaný) počet pozorování v jednotlivých intervalech $(c_{j-1}, c_j]$ pomocí Pearsonovy χ^2 statistiky testu dobré shody:

$$\chi^2(\theta) = \sum_{j=1}^k \left(\frac{n_j - np_j(\theta)}{\sqrt{np_j(\theta)}} \right)^2.$$

•

$$\chi^2(\theta) = \sum_{j=1}^k \frac{n_j^2}{np_j(\theta)} - n.$$

- Odhad parametru θ metodou minimálního χ^2 minimalizuje $\chi^2(\theta)$ přes všechny hodnoty $\theta \in \Theta$, tj. $\tilde{\theta} = \arg \min \{ \chi^2(\theta), \theta \in \Theta \}$.

Metoda minimálního χ^2 pro klasická data

- Obor hodnot náhodné veličiny X_i musíme na k po dvou disjunktních intervalů rozdělit sami, stejně tak spočítat n_j počet pozorování v intervalu $(c_{j-1}, c_j]$.
- Na tato umělá intervalová data aplikujeme předchozí postup.
- Intervaly $(c_{j-1}, c_j]$ by se měly volit stejně pravděpodobné, tj. $p_j(\tilde{\theta}) = \frac{1}{k}$ pro $j = 1, \dots, k$.
- Volba počtu tříd k - heuristická pravidla, např. $k \doteq 15 \left(\frac{n}{100}\right)^{2/5}$, nebo $k \doteq 2n^{2/5}$.

Bayesovské odhady (Bayesovská statistika)

- Kombinuje informaci obsaženou v datech (parametrický model) s apriorní informací o neznámém parametru θ (zkušenosti, domněnky, dřívější pozorování).
- Závěry (odhady) vyvozuje až z aposteriorního rozdělení.
- Idea: Naše informace o hodnotě neznámého parametru může být vyjádřena pomocí pravděpodobnostního rozdělení, tj. neznámý parametr θ považujeme za náhodný vektor.

Matematický model

- X_1, \dots, X_n je náhodný výběr z rozdělení s hustotou $f(x, \theta)$, kde $\theta \in \Theta$.
- θ je nyní náhodný vektor s hustotou $q(\theta)$.
- Označme podmíněnou hustotu náhodného vektoru $(X_1, \dots, X_n)'$ při dané hodnotě parametru θ jako $r(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i, \theta)$, kde $\mathbf{x} = (x_1, \dots, x_n)'$.

Theorem (Bayesova věta)

Pro podmíněnou hustotu náhodného vektoru θ při daných hodnotách $\mathbf{X} = \mathbf{x}$ platí:

$$\pi(\theta|\mathbf{x}) = \begin{cases} \frac{r(\mathbf{x}|\theta)q(\theta)}{\int_{\Theta} r(\mathbf{x}|\theta)q(\theta)d\theta}, & \text{pokud } \int_{\Theta} r(\mathbf{x}|\theta)q(\theta)d\theta \neq 0, \\ 0, & \text{jinak.} \end{cases}$$

- $q(\theta)$ se nazývá apriorní hustota - vyjadřuje informaci o parametru θ ještě před realizací náhodného výběru \mathbf{X} .
- $\pi(\theta|\mathbf{x})$ se nazývá aposteriorní hustota - vyjadřuje informaci o parametru θ až po realizaci náhodného výběru \mathbf{X} .
- Při Bayesovském přístupu používáme kromě dat (realizace náhodného výběru) ještě informaci o parametru θ nezávisle na našich datech.
- Tato informace může mít objektivní i subjektivní charakter.

Volba apriorního rozdělení

- Pokud máme informace (výsledky) z minulosti
 - přesná znalost rozdělení
 - jádrové odhady hustoty
 - parametrický model
- Pokud nemáme informace (výsledky) z minulosti
 - neinformativní (rovnoměrné) rozdělení $q(\boldsymbol{\theta}) \propto 1$
 - Jeffreysovo apriorní rozdělení $q(\boldsymbol{\theta}) \propto \sqrt{|\mathbf{J}(\boldsymbol{\theta})|}$
 - konjugované apriorní rozdělení

Bodové odhady

- Definujme ztrátovou funkci $L(\theta, \hat{\theta})$ - ztráta, kterou utrpíme, když odhadneme parametr θ pomocí odhadu $\hat{\theta}$.
- Dále definujme bayesovské riziko (průměrná aposteriorní ztráta):

$$r(\theta) = \int_{\Theta} L(\theta, \hat{\theta}) \pi(\theta | \mathbf{x}) d\theta.$$

- Hledáme odhad, který minimalizuje bayesovské riziko. Necht' dále θ je jednorozměrný parametr.
- Pro kvadratickou ztrátovou funkci $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ je bayesovským odhadem aposteriorní střední hodnota, tj. $\hat{\theta} = \mathbb{E}(\theta | X_1, \dots, X_n)$.
- Pro absolutní ztrátovou funkci $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ je bayesovským odhadem aposteriorní medián, tj. $\hat{\theta} = \text{med}(\theta | X_1, \dots, X_n)$.
- Pro 0-1 ztrátovou funkci $L(\theta, \hat{\theta}) = \mathbb{I}\{\theta \neq \hat{\theta}\}$ je bayesovským odhadem aposteriorní modus, tj. $\hat{\theta} = \arg \max \pi(\theta | X_1, \dots, X_n)$.

Intervalové odhady

Definition

100(1 - α)% věrohodnostní interval pro parametr θ je takový interval $[a, b] = [a(X_1, \dots, X_n), b(X_1, \dots, X_n)]$, pro který $P(a \leq \theta \leq b | \mathbf{X}) = 1 - \alpha$.

- Equal tail - je takový interval $[a, b]$, kde a je $\alpha/2$ -kvantil aposteriorního rozdělení $\theta | \mathbf{X}$ a b je $1 - \alpha/2$ -kvantil aposteriorního rozdělení $\theta | \mathbf{X}$.
- HPD (interval o největší aposteriorní hustotě) je takový interval $[a, b]$ takový, že $\pi(\theta | \mathbf{x}) \geq c$ pro všechna $\theta \in [a, b]$ a $c > 0$ je nejmenší číslo takové, že $P(\pi(\theta | \mathbf{x}) \geq c) = 1 - \alpha$.

Theorem

Je-li $\pi(\theta | \mathbf{x})$ spojitá a unimodální, pak HPD interval je nejkratší mezi všemi věrohodnostními intervaly.

Predikce budoucího pozorování

- Necht' se budoucí pozorování X_{n+1} řídí stejným modelem jako X_1, \dots, X_n - tedy má hustotu $f(x, \theta)$.
- Chceme predikovat (předpovídat) jeho budoucí hodnotu.
- S pomocí Bayesovy věty můžeme odvodit aposteriorní prediktivní hustotu

$$f(x_{n+1}|\mathbf{x}) = \int_{\Theta} f(x_{n+1}, \theta) \pi(\theta|\mathbf{x}) d\theta.$$

Model selection (výběr modelu)

- 1 Je náš model vhodný? Popisuje dobře naše data?
- 2 Máme-li více modelů, který z nich je nejlepší? Který máme použít?

Grafické metody pro posouzení vhodnosti modelu

- Jsou založené na porovnání teoretického (parametrický model) a empirického rozdělení (neparametrický odhad).
- Porovnání teoretické distribuční funkce $F(x, \hat{\theta})$ a empirické distribuční funkce $\hat{F}_n(x)$ v jednom grafu.
- Porovnání teoretické a empirické distribuční funkce pomocí funkce $D(x) = \hat{F}_n(x) - F(x, \hat{\theta})$.
- Porovnání teoretické hustoty $f(x, \hat{\theta})$ a empirické hustoty (histogram, jádrový odhad) v jednom grafu.
- Q-Q plot
- P-P plot

Q-Q plot

- Porovnává teoretické a empirické kvantily.
- Uspořádané hodnoty náhodného výběru označme

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

- Podle upravené definice $x_{(i)}$ je $p_i = \frac{i-\beta}{n+1-2\beta}$ -tý výběrový kvantil, kde $0 \leq \beta < 1$ je korekční faktor.
- Q-Q plot je graf $\left[F^{-1}(p_i, \hat{\theta}), x_{(i)} \right]$ pro $i = 1, \dots, n$.
- Je-li náš model správný, pak by se body Q-Q plotu měly náhodně vyskytovat kolem osy prvního kvadrantu.

P-P plot

- porovnává hodnoty teoretické a empirické distribuční funkce.
- P-P plot je graf $\left[F(x_{(i)}, \hat{\theta}), \frac{i}{n+1} \right]$ pro $i = 1, \dots, n$.
- Je-li náš model správný, pak by se body P-P plotu měly náhodně vyskytovat kolem osy prvního kvadrantu.

Statistické testy

- Potřebujeme obecnější model: X_1, \dots, X_n je náhodný výběr z rozdělení s distribuční funkcí F (libovolná).
- Formálně testujeme nulovou hypotézu $H_0: F = F(x, \theta)$ pro nějaké $\theta \in \Theta$ proti alternativě, že H_0 neplatí.
- Testujeme tedy, že námi specifikovaný model je vhodný pro naše data.
- Testy založené na porovnání distribučních funkcí (Kolmogorovův - Smirnovův test, Andersonův - von Darlingův test, Cramérův - von Misesův test), testy dobré shody (Pearsonův χ^2 test) a další.
- Pro odvození testů budeme ještě potřebovat pomocnou nulovou hypotézu $H_0^*: F = F(x, \theta^*)$, kde θ^* je známá hodnota.

Kolmogorovův - Smirnovův test

- Nulová hypotéza H_0^* : $F = F(x, \theta^*)$, kde θ^* je známá hodnota.
- Testová statistika
$$D_n = \max_{x \in \mathbb{R}} \{|\widehat{F}_n(x) - F(x, \theta^*)|\} = \max_{i=1, \dots, n} \left\{ \left| \frac{i}{n} - F(x_{(i)}, \theta^*) \right| \right\}.$$
- Za platnosti H_0^* má $\sqrt{n}D_n$ asymptotické rozdělení stejné jako $\sup_{t \in [0,1]} |B(t)|$, kde $B(t)$ je Brownův most v $\mathcal{C}(0, 1)$.
- To má distribuční funkci $1 - 2 \sum_{j=0}^{\infty} (-1)^{j+1} e^{-2j^2 y^2}$, pro $y > 0$.
- A aproximativní kvantilovou funkci $\sqrt{\frac{1}{2} \log \frac{2}{1-\alpha}}$, pro $0 < \alpha < 1$.
- Test se dá použít jen, když hodnota θ^* je známá. Pokud jsou k jejímu odhadu použita data, test nefunguje - je příliš konzervativní.

Kolmogorovův - Smirnovův test (modifikace)

- Uvažujme původní hypotézu $H_0: F = F(x, \theta)$ pro nějaké $\theta \in \Theta$.
- Neznámý parametr θ nejprve odhadneme z dat.

- Testová statistika

$$D_n = \max_{x \in \mathbb{R}} \{|\widehat{F}_n(x) - F(x, \widehat{\theta})|\} = \max_{i=1, \dots, n} \{|\frac{i}{n} - F(x_{(i)}, \widehat{\theta})|\}.$$

- Rozdělení testové statistiky D_n za platnosti H_0 závisí na daném rozdělení, ze kterého data pocházejí (a v některých situacích i na jeho parametrech).
- Pro testování normality bylo toto rozdělení odvozeno - Lillieforsův test.
- Pro ostatní rozdělení lze použít simulace - odhadnout příslušnou p -hodnotu testu pomocí parametrického bootstrapu.

Kolmogorovův - Smirnovův test (parametrický bootstrap)

- 1 Spočítáme hodnotu testové statistiky D_n pro naše data s odhadnutým parametrem $\hat{\theta}$, označme ji t .
- 2 Nageneryjeme si nový náhodný výběr o rozsahu n z rozdělení s distribuční funkcí $F(x, \hat{\theta})$, realizaci označme x_1^*, \dots, x_n^* .
- 3 Odhadneme neznámý parametr θ , označme jej $\tilde{\theta}$.
- 4 Pro tuto realizaci spočítáme hodnotu testové statistiky
$$D_n = \max_{x \in \mathbb{R}} \{|\hat{F}_n^*(x) - F(x, \tilde{\theta})|\} = \max_{i=1, \dots, n} \left\{ \left| \frac{i}{n} - F(x_{(i)}^*, \tilde{\theta}) \right| \right\}.$$
- 5 Body (2) – (4) několikrát opakujeme.
- 6 p -hodnotu testu poté odhadneme jako relativní četnost případů, kdy $D_n \geq t$.

Cramerův - von Misesův test

- Patří do třídy testů s testovou statistikou

$$n \int_{-\infty}^{\infty} \left(\widehat{F}_n(x) - F(x, \widehat{\theta}) \right)^2 w(F(x, \widehat{\theta})) f(x, \widehat{\theta}) dx$$

pro nějakou váhovou funkci w .

- Cramerův - von Misesův test používá váhovou funkci $w(y) = 1$ pro $0 < y < 1$.
- Testová statistika se dá zjednodušit do tvaru

$$T = \frac{1}{12n} + \sum_{i=1}^n \left[\frac{2i-1}{2n} - F(x_{(i)}, \widehat{\theta}) \right]^2.$$

- Rozdělení testové statistiky T i pro θ známé závisí na testovaném rozdělení.
- Příslušnou p -hodnotu testu musíme získat pomocí simulací, například pomocí parametrického bootstrapu.

Andersonův - Darlingův test

- Patří do třídy testů s testovou statistikou

$$n \int_{-\infty}^{\infty} \left(\widehat{F}_n(x) - F(x, \widehat{\theta}) \right)^2 w(F(x, \widehat{\theta})) f(x, \widehat{\theta}) dx$$

pro nějakou váhovou funkci w .

- Andersonův - Darlingův test používá váhovou funkci $w(y) = \frac{1}{y(1-y)}$ pro $0 < y < 1$.
- Testová statistika se dá zjednodušit do tvaru

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) \left[\log(F(x_{(i)}, \widehat{\theta})) + \log(1 - F(x_{(n+1-i)}, \widehat{\theta})) \right].$$

- Rozdělení testové statistiky A^2 i pro θ známé závisí na testovaném rozdělení.
- Příslušnou p -hodnotu testu musíme získat pomocí simulací, například pomocí parametrického bootstrapu.

Pearsonův χ^2 test dobré shody

- Začneme opět nejprve s nulovou hypotézou H_0^* : $F = F(x, \theta^*)$, kde θ^* je známá hodnota.
- Definujme si intervaly $(c_{j-1}, c_j]$, $j = 1, \dots, k$.
- Označme n_j počet pozorování, které padnou do intervalu $(c_{j-1}, c_j]$ pro $j = 1, \dots, k$.
- Určíme očekávaný počet pozorování (za platnosti H_0^*), které by měly padnout do intervalu $(c_{j-1}, c_j]$:

$$e_j = np_j(\theta^*) = nP(X_1 \in (c_{j-1}, c_j]) = n(F(c_j, \theta^*) - F(c_{j-1}, \theta^*)).$$

- Testová statistika

$$\chi^2 = \sum_{j=1}^k \frac{(n_j - e_j)^2}{e_j}.$$

- χ^2 má za platnosti nulové hypotézy H_0^* asymptoticky χ^2 rozdělení s $k - 1$ stupni volnosti.

Modifikace Pearsonova χ^2 testu dobré shody

- Vraťme se k původní hypotéze $H_0: F = F(x, \theta)$ pro nějaké $\theta \in \Theta$.
- Nejprve z dat odhadneme neznámý p -rozměrný parametr θ , označme jej $\hat{\theta}$, a postupujeme stejně jako v předchozím.
- Testová statistika

$$\chi^2 = \sum_{j=1}^k \frac{(n_j - e_j)^2}{e_j},$$

kde $e_j = n(F(c_j, \hat{\theta}) - F(c_{j-1}, \hat{\theta}))$.

- Upravená testová statistika χ^2 má za platnosti nulové hypotézy H_0 asymptoticky χ^2 rozdělení s $k - 1 - p$ stupni volnosti.

Volba tříd:

- $e_j = \frac{n}{k}$ pro $j = 1, \dots, k$.
- Heuristická pravidla pro počet tříd – $k \doteq 2n^{2/5}$, nebo $k \doteq 15(n/100)^{2/5}$.

Výběr modelu z několika kandidátů

Princip Occamovy břitvy - vybíráme co nejjednodušší vhodný model.

- 1 Judgement-based přístup - založený na subjektivním úsudku analytika
 - rozhodnutí založené na různých grafech či tabulkách (tail vs. mod fit)
 - rozhodnutí založené na předchozí zkušenosti (Paretovo rozdělení pro výši příjmů, Benfordovo pro četnost prvních číslic)
 - model je plně určen situací, kterou má popisovat (házení mincí - alternativní rozdělení)
- 2 Score-based přístup - založený na číselných charakteristikách
 - nejnižší hodnota statistiky nějakého statistického testu
 - nejvyšší p -hodnota nějakého statistického testu
 - nejvyšší hodnota věrohodnosti
 - nejnižší hodnota nějaké penalizované funkce, např. AIC, BIC:
 - $AIC = -2l(\hat{\theta}) + 2p.$
 - $BIC = -2l(\hat{\theta}) + p \log n.$

Cíl:

- odhadnout $P(X > x)$ pro x velké.
- odhadnout $F^{-1}(\alpha)$ pro α blízké 1.
- určit výši plnění, kterou nárokuje jen malé procento klientů s nevyšším plněním.
- určit, jak často budou klienti nárokovat vysoké pojistné plnění.

Metody:

- Metoda blokových maxim.
- Metoda založená na překročení meze (peaks-over-threshold; POT).

Chování maxima náhodného výběru

- Necht' X_1, \dots, X_n je náhodný výběr z rozdělení s distribuční funkcí $F(x)$. Označme $M_n = \max\{X_1, \dots, X_n\}$ maximum X_1, \dots, X_n . Počítejme jeho distribuční funkci:

$$\begin{aligned} G_n(x) &= P(M_n \leq x) = P(\max\{X_1, \dots, X_n\} \leq x) \\ &= P(X_1 \leq x, \dots, X_n \leq x) = P(X_1 \leq x) \cdots P(X_n \leq x) = F(x)^n. \end{aligned}$$

- Hledejme jeho asymptotické rozdělení

$$\lim_{n \rightarrow \infty} G_n(x) = \begin{cases} 0, & \text{pokud } x < x_F \\ 1, & \text{pokud } x \geq x_F, \end{cases}$$

kde $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$ je pravý koncový bod nosiče F .

- Limitní rozdělení M_n je degenerované v bodě x_F , nebo "utíká" do nekonečna.
- Budeme hledat posloupnosti konstant $\{a_n\}$ a $\{b_n\}$ tak, aby $\frac{M_n - b_n}{a_n}$ (normovaná maxima) konvergovala k nějakému nedegenerovanému rozdělení.

Rozdělení extrémních hodnot

1 Gumbelovo rozdělení



$$G_0(x) = e^{-e^{-x}}, \quad x \in \mathbb{R}.$$



$$g_0(x) = e^{-(e^{-x}+x)}, \quad x \in \mathbb{R}.$$

2 Fréchetovo rozdělení s parametrem tvaru $\alpha > 0$



$$G_1(x) = e^{-x^{-\alpha}}, \quad x > 0.$$



$$g_1(x) = \frac{\alpha}{x^{\alpha+1}} e^{-x^{-\alpha}}, \quad x > 0.$$

3 Weibullovo (extremální) rozdělení s parametrem tvaru $\alpha < 0$



$$G_2(x) = e^{-(-x)^{-\alpha}}, \quad x < 0.$$



$$g_2(x) = -\alpha(-x)^{-\alpha-1} e^{-(-x)^{-\alpha}}, \quad x < 0.$$

Rozdělení extrémních hodnot s parametry polohy a měřítka

- Předchozí tři rozdělení jsou standardizované.
- Přidáme parametr polohy μ a parametr měřítka $\sigma > 0$.
- $G_{i,\mu,\sigma}(x) = G_i\left(\frac{x-\mu}{\sigma}\right)$.

1 Gumbelovo rozdělení

$$G_{0,\mu,\sigma}(x) = e^{-e^{-\frac{x-\mu}{\sigma}}}, \quad x \in \mathbb{R}.$$

2 Fréchetovo rozdělení s parametrem tvaru $\alpha > 0$

$$G_{1,\mu,\sigma}(x) = e^{-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}}, \quad x > \mu.$$

3 Weibullovo (extremální) rozdělení s parametrem tvaru $\alpha < 0$

$$G_{2,\mu,\sigma}(x) = e^{-\left(-\frac{x-\mu}{\sigma}\right)^{-\alpha}}, \quad x < \mu.$$

Předchozí tři distribuce můžeme zapsat jedním vzorcem.

Zobecněné rozdělení extrémních hodnot (GEV rozdělení)

- GEV rozdělení s parametrem $\gamma \in \mathbb{R}$ ve standardizovaném tvaru

$$G_\gamma(x) = e^{-(1+\gamma x)^{-\frac{1}{\gamma}}}, \quad 1 + \gamma x > 0.$$

- Pro $\gamma = 0$ dostaneme Gumbelovo rozdělení.
- Pro $\gamma > 0$ dostaneme Fréchetovo rozdělení.
- Pro $\gamma < 0$ dostaneme Weibullovo rozdělení.
- Opět budeme potřebovat přidat parametr polohy μ a parametr měřítka $\sigma > 0$.
- GEV rozdělení s parametrem polohy $\mu \in \mathbb{R}$, parametrem měřítka $\sigma > 0$ a parametrem tvaru $\gamma \in \mathbb{R}$

$$G_{\gamma,\mu,\sigma}(x) = e^{-(1+\frac{\gamma(x-\mu)}{\sigma})^{-\frac{1}{\gamma}}}, \quad 1 + \frac{\gamma(x-\mu)}{\sigma} > 0.$$

Theorem (Fisherova - Tippettova věta)

*Nechť X_1, \dots, X_n je náhodný výběr z rozdělení s distribuční funkcí $F(x)$.
Nechť existují posloupnosti konstant $\{a_n\}$ a $\{b_n\}$ tak, že
 $P\left(\frac{M_n - b_n}{a_n} \leq x\right) \rightarrow H(x)$ pro $n \rightarrow \infty$ pro nějakou nedegenerovanou
distribuční funkci $H(x)$. Pak $H(x)$ je distribuční funkce GEV rozdělení.*

- Předchozí věta říká, že GEV je jediné možné limitní rozdělení maxim.
- Maxima náhodného výběru budeme modelovat pomocí GEV rozdělení s parametry μ, σ a γ .

Aplikace metody blokových maxim

- Necht' X_1, \dots, X_N je náhodný výběr z rozdělení s distribuční funkcí $F(x)$. Pozor, nyní počet pozorování značíme N .
- Data rozdělíme do m bloků o velikosti n ($N = m \cdot n$).
- V každém bloku najdeme maximum, maximum v i -tém bloku označme $M_i = M_i^{(n)}$.
- Dále budeme modelovat veličiny M_1, \dots, M_m (jsou nezávislé a stejně rozdělené).
- Ty budeme modelovat pomocí GEV rozdělení s parametry μ, σ a γ .
- Délka bloku n musí být dostatečně velká, aby "fungovala" aproximace pomocí GEV rozdělení.
- Počet bloků m musí být taky dostatečně velký, aby odhady parametrů byly "přesné".

Metoda maximální věrohodnosti pro GEV rozdělení

- Model: M_1, \dots, M_m je náhodný výběr z GEV rozdělení s parametry μ, σ a γ .
- Logaritmická věrohodnostní funkce je

$$l(\gamma, \mu, \sigma) = -m \log \sigma - \left(1 + \frac{1}{\gamma}\right) \sum_{i=1}^m \log \left(1 + \frac{\gamma(M_i - \mu)}{\sigma}\right) - \sum_{i=1}^m \left(1 + \frac{\gamma(M_i - \mu)}{\sigma}\right)^{-\frac{1}{\gamma}},$$

pro $1 + \frac{\gamma(M_1 - \mu)}{\sigma} > 0, \dots, 1 + \frac{\gamma(M_m - \mu)}{\sigma} > 0$.

- Tu maximalizujeme přes všechny hodnoty γ, μ a $\sigma > 0$ takové, že $1 + \frac{\gamma(M_1 - \mu)}{\sigma} > 0, \dots, 1 + \frac{\gamma(M_m - \mu)}{\sigma} > 0$.
- Funkce $l(\gamma, \mu, \sigma)$ není diferencovatelná, proto ji musíme maximalizovat numericky.
- Pro $\gamma > -\frac{1}{2}$, maximálně věrohodný odhad je asymptoticky nestranný, konzistentní a asymptoticky normální.

Metoda pravděpodobnostně vážených momentů pro GEV rozdělení

Definition

Nechť X je náhodná veličina s distribuční funkcí $F(x)$, pak čísla $M_{p,r,s} = \mathbb{E}[X^p F(X)^r (1 - F(X))^s]$ pro $p, r, s \in \mathbb{R}$ nazveme pravděpodobnostně vážené momenty.

- Speciálně položme $p = 1$ a $s = 0$ a označme $\beta_r = M_{1,r,0} = \mathbb{E}[X \cdot F(X)^r]$ pro $r = 0, 1, 2$.
- Pro GEV rozdělení je $\beta_r = \frac{1}{r+1} \left\{ \mu - \frac{\sigma}{\gamma} [1 - (r+1)^\gamma \Gamma(1-\gamma)] \right\}$ pro $\gamma < 1, \gamma \neq 0$.
- Jeho odhad je $\hat{\beta}_0 = \frac{1}{m} \sum_{i=1}^m M_i$ a $\hat{\beta}_r = \frac{1}{m} \sum_{i=1}^m \left(\prod_{j=1}^r \frac{i-j}{m-j} M_{(i)} \right)$ pro $r = 1, 2$.
- Odhady parametrů metodou pravděpodobnostně vážených momentů získáme jako řešení soustavy $\beta_r = \hat{\beta}_r$, pro $r = 0, 1, 2$.

Odhad pravděpodobnosti překročení vysoké hranice



$$P(M_1 \leq x) = P(X_i \leq x)^n = [1 - P(X_i > x)]^n.$$



$$P(X_i > x) = 1 - P(M_1 \leq x)^{\frac{1}{n}} = 1 - [G_{\gamma, \mu, \sigma}(x)]^{\frac{1}{n}}.$$

- Pro x velké můžeme odhadnout

$$P(\widehat{X_i > x}) = 1 - [G_{\widehat{\gamma}, \widehat{\mu}, \widehat{\sigma}}(x)]^{\frac{1}{n}} = 1 - e^{-\frac{1}{n} \left(1 + \frac{\widehat{\gamma}(x - \widehat{\mu})}{\widehat{\sigma}}\right)^{-\frac{1}{\widehat{\gamma}}}}.$$

Odhad vysokého kvantilu

- q_α je α -kvantil veličiny X_i , jestliže $P(X_i \leq q_\alpha) = \alpha$.



$$1 - \alpha = P(X_i > q_\alpha) = 1 - P(M_1 \leq q_\alpha)^{\frac{1}{n}} = 1 - [G_{\gamma, \mu, \sigma}(q_\alpha)]^{\frac{1}{n}}.$$



$$q_\alpha = G_{\gamma, \mu, \sigma}^{-1}(\alpha^n) = \mu + \frac{\sigma}{\gamma} \left((-\log(\alpha^n))^{-\gamma} - 1 \right).$$

- Tedy α -kvantil náhodné veličiny X_i je roven α^n -kvantilu GEV rozdělení.
- Pro α blízké 1 můžeme odhadnout

$$\hat{q}_\alpha = G_{\hat{\gamma}, \hat{\mu}, \hat{\sigma}}^{-1}(\alpha^n) = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\gamma}} \left((-\log(\alpha^n))^{-\hat{\gamma}} - 1 \right).$$

Odhad doby návratu

- Cílem je stanovit průměrnou frekvenci výskytu extrémního jevu, tj. jak často je překračována nějaká vysoká hranice.
- Frekventistická definice pravděpodobnosti: je-li $P(X_i > x) = p$, pak X_i překročí hranici x v průměru jednou za $\frac{1}{p}$ časových okamžiků.
- Předpokládejme, že máme danou hranici x .
- Označme k průměrnou frekvenci, tj. $k = \frac{1}{p}$ a hledejme jej tak, že platí

$$P(X_i > x) = \frac{1}{k}.$$

- Tedy

$$k = \frac{1}{P(X_i > x)} = \frac{1}{1 - [G_{\gamma, \mu, \sigma}(x)]^{\frac{1}{n}}}.$$

- Tedy odhad doby návratu je

$$\hat{k} = \frac{1}{1 - e^{-\frac{1}{n} \left(1 + \frac{\hat{\gamma}(x - \hat{\mu})}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\gamma}}}}.$$

Odhad úrovně návratu

- Cílem je stanovit hranici x , která je překračována v průměru jednou za k časových okamžiků.
- Opět začneme s frekventistickou definicí pravděpodobnosti:

$$P(X_i > x) = \frac{1}{k}.$$

- Potom x je $(1 - \frac{1}{k})$ -kvantil X_i , tj.

$$x = G_{\gamma, \mu, \sigma}^{-1} \left(\left(1 - \frac{1}{k}\right)^n \right).$$

- Tedy odhad úrovně návratu je

$$\hat{x} = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\gamma}} \left(\left(-\log \left(\left(1 - \frac{1}{k}\right)^n \right) \right)^{-\hat{\gamma}} - 1 \right).$$

Odhad doby a úrovně návratu II

- V praxi často doba a úroveň návratu chápe časové okamžiky jako počty bloků.
- Označme k^* průměrnou frekvenci (v počtech bloků), pak

$$P(M_i > x) = \frac{1}{k^*} = 1 - G_{\gamma, \mu, \sigma}(x).$$

- Tedy odhad doby návratu (v počtech bloků) je

$$\widehat{k^*} = \frac{1}{1 - G_{\widehat{\gamma}, \widehat{\mu}, \widehat{\sigma}}(x)} = \frac{1}{1 - e^{-(1 + \frac{\widehat{\gamma}(x - \widehat{\mu})}{\widehat{\sigma}})^{-\frac{1}{\widehat{\gamma}}}}}.$$

- A odhad úrovně návratu

$$\widehat{x} = G_{\widehat{\gamma}, \widehat{\mu}, \widehat{\sigma}}^{-1}\left(1 - \frac{1}{\widehat{k^*}}\right) = \widehat{\mu} + \frac{\widehat{\sigma}}{\widehat{\gamma}} \left(\left(-\log\left(1 - \frac{1}{\widehat{k^*}}\right) \right)^{-\widehat{\gamma}} - 1 \right).$$

Chování excesů náhodného výběru

- Necht' X_1, \dots, X_n je náhodný výběr z rozdělení s distribuční funkcí $F(x)$. Zvolme nějakou hranici (práh, threshold) u a definujme $Y_i^{(u)} = X_i - u$ pro $X_i > u$ výši jeho překročení (exces) pro pozorování, která tuto hranici překročila.
- Označme N_u počet pozorování, která překročila hranici u .
- Hledejme distribuční funkci $Y_i^{(u)}$, označme ji $F_u(x)$:

$$\begin{aligned} F_u(x) &= P(Y_i^{(u)} \leq x) = P(X_i - u \leq x | X_i > u) \\ &= \frac{P(u < X_i \leq u + x)}{P(X_i > u)} = \frac{F(u + x) - F(u)}{1 - F(u)}, \quad x \geq 0. \end{aligned}$$

- Hledejme jeho asymptotické rozdělení pro $u \nearrow x_F$.
- To bude degenerované v bodě 0.
- Budeme hledat posloupnosti konstant $\{a_n\}$ a $\{b_n\}$ tak, aby $\frac{Y_i^{(u)} - b_n}{a_n}$ (normované excesy) konvergovaly k nějakému nedegenerovanému rozdělení.

Rozdělení pro modelování excesů

1 Exponenciální rozdělení



$$W_0(x) = 1 - e^{-x}, \quad x \geq 0.$$



$$w_0(x) = e^{-x}, \quad x \geq 0.$$

2 Paretovo rozdělení s parametrem tvaru $\alpha > 0$



$$W_1(x) = 1 - x^{-\alpha}, \quad x \geq 1.$$



$$w_1(x) = \frac{\alpha}{x^{\alpha+1}}, \quad x \geq 1.$$

3 Beta rozdělení s parametrem tvaru $\alpha < 0$



$$W_2(x) = 1 - (-x)^{-\alpha}, \quad -1 \leq x \leq 0.$$



$$w_2(x) = -\alpha(-x)^{-\alpha-1}, \quad -1 \leq x \leq 0.$$

Rozdělení pro modelování excesů s parametry polohy a měřítka

- Předchozí tři rozdělení jsou standardizované.
- Přidáme parametr polohy μ a parametr měřítka $\sigma > 0$.
- $W_{i,\mu,\sigma}(x) = W_i\left(\frac{x-\mu}{\sigma}\right)$.
- ① Exponenciální rozdělení

$$W_{0,\mu,\sigma}(x) = 1 - e^{-\frac{x-\mu}{\sigma}}, \quad x \geq \mu.$$

- ② Paretovo rozdělení s parametrem tvaru $\alpha > 0$

$$W_{1,\mu,\sigma}(x) = 1 - \left(\frac{x-\mu}{\sigma}\right)^{-\alpha}, \quad x \geq \mu + \sigma.$$

- ③ Beta rozdělení s parametrem tvaru $\alpha < 0$

$$W_{2,\mu,\sigma}(x) = 1 - \left(-\frac{x-\mu}{\sigma}\right)^{-\alpha}, \quad \mu - \sigma \leq x \leq \mu.$$

Předchozí tři distribuce můžeme zapsat jedním vzorcem.

Zobecněné Paretovo rozdělení (GPD rozdělení)

- GPD rozdělení s parametrem $\gamma \in \mathbb{R}$ ve standardizovaném tvaru

$$W_\gamma(x) = 1 - (1 + \gamma x)^{-\frac{1}{\gamma}}.$$

- Pro $\gamma = 0$ dostaneme exponenciální rozdělení ($x \geq 0$).
- Pro $\gamma > 0$ dostaneme Paretovo rozdělení ($x \geq 0$).
- Pro $\gamma < 0$ dostaneme beta rozdělení ($0 \leq x \leq -\frac{1}{\gamma}$).
- Nyní budeme potřebovat přidat pouze parametr měřítka $\sigma > 0$.
- GPD rozdělení s parametrem měřítka $\sigma > 0$ a parametrem tvaru $\gamma \in \mathbb{R}$

$$W_{\gamma,\sigma}(x) = 1 - \left(1 + \frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}}.$$

Modelování excesů

Theorem (Balkemova - de Haanova - Pickandsova věta)

Nechť X_1, \dots, X_n je náhodný výběr z rozdělení s distribuční funkcí $F(x)$.

Nechť existují posloupnosti konstant $\{a_n\}$ a $\{b_n\}$ tak, že

$P\left(\frac{Y_i^{(u_n)} - b_n}{a_n} \leq x\right) \rightarrow H(x)$ pro $u_n \nearrow x_F$ pro nějakou nedegenerovanou

spojitou distribuční funkci $H(x)$. Pak $H(x)$ je distribuční funkce GPD rozdělení.

- Předchozí věta říká, že GPD rozdělení je jediné možné limitní rozdělení excesů.
- Excesy náhodného výběru budeme modelovat pomocí GPD rozdělení s parametry σ a γ .

Aplikace POT metody

- Necht' X_1, \dots, X_N je náhodný výběr z rozdělení s distribuční funkcí $F(x)$. Pozor, počet pozorování opět značíme N .
- Zvolíme dostatečně vysokou hranici u .
- A definujeme excesy $Y_i = X_i - u$, pokud $X_i > u$, pro $i = 1, \dots, N_u$ (N_u je počet excesů).
- Dále budeme modelovat veličiny Y_1, \dots, Y_{N_u} (jsou nezávislé a stejně rozdělené).
- Ty budeme modelovat pomocí GPD rozdělení s parametry σ a γ .
- Velikost prahu u musí být dostatečně velká, aby "fungovala" aproximace pomocí GPD rozdělení.
- Počet excesů N_u musí být taky dostatečně velký, aby odhady parametrů byly "přesné".

Metoda maximální věrohodnosti pro GPD rozdělení

- Model: Y_1, \dots, Y_{N_u} je náhodný výběr z GPD rozdělení s parametry σ a γ .
- Logaritmická věrohodnostní funkce je

$$l(\gamma, \sigma) = -N_u \log \sigma - \left(1 + \frac{1}{\gamma}\right) \sum_{i=1}^{N_u} \log \left(1 + \frac{\gamma}{\sigma} Y_i\right)$$

pro $1 + \frac{\gamma}{\sigma} Y_1 > 0, \dots, 1 + \frac{\gamma}{\sigma} Y_{N_u} > 0$.

- Tu maximalizujeme přes všechny hodnoty γ a $\sigma > 0$ takové, že $1 + \frac{\gamma}{\sigma} Y_1 > 0, \dots, 1 + \frac{\gamma}{\sigma} Y_{N_u} > 0$.
- Funkce $l(\gamma, \sigma)$ není diferencovatelná, proto ji musíme maximalizovat numericky.
- Pro $\gamma > -\frac{1}{2}$, maximálně věrohodný odhad je asymptoticky nestranný, konzistentní a asymptoticky normální.

Metoda pravděpodobnostně vážených momentů pro GPD rozdělení

Připomeňme definici pravděpodobnostně vážených momentů:

Definition

Nechť X je náhodná veličina s distribuční funkcí $F(x)$, pak čísla $M_{p,r,s} = \mathbb{E}[X^p F(X)^r (1 - F(X))^s]$ pro $p, r, s \in \mathbb{R}$ nazveme pravděpodobnostně vážené momenty.

- Speciálně položme $p = 1$ a $r = 0$ a označme $\alpha_s = M_{1,0,s} = \mathbb{E}[X \cdot (1 - F(X))^s]$ pro $s = 0, 1$.
- Pro GPD rozdělení je $\alpha_s = \frac{\sigma}{(s-\gamma+1)(s+1)}$ pro $\gamma < 1$.
- Jeho odhad je $\hat{\alpha}_0 = \frac{1}{N_u} \sum_{i=1}^{N_u} Y_i$ a $\hat{\alpha}_1 = \frac{1}{N_u} \sum_{i=1}^{N_u} \frac{N_u - i}{N_u - 1} Y_{(i)}$.
- Odhady parametrů metodou pravděpodobnostně vážených momentů získáme jako řešení soustavy $\alpha_s = \hat{\alpha}_s$, pro $s = 0, 1$.

Odhad pravděpodobnosti překročení vysoké hranice



$$P(X_i > x | X_i > u) = \frac{P(X_i > x)}{P(X_i > u)}, \text{ pro } x \geq u.$$



$$\begin{aligned} P(X_i > x) &= P(X_i > u)P(X_i > x | X_i > u) \\ &= P(X_i > u) [1 - P(Y_i \leq x - u | X_i > u)] \\ &= P(X_i > u) [1 - F_u(x - u)] \\ &= P(X_i > u) (1 - W_{\gamma, \sigma}(x - u)) \\ &= P(X_i > u) \left(1 + \frac{\gamma(x - u)}{\sigma}\right)^{-\frac{1}{\gamma}}. \end{aligned}$$

- Pro x velké můžeme odhadnout

$$P(\widehat{X}_i > x) = \frac{N_u}{N} \left(1 + \frac{\widehat{\gamma}(x - u)}{\widehat{\sigma}}\right)^{-\frac{1}{\widehat{\gamma}}}.$$

Odhad vysokého kvantilu

- q_α je α -kvantil veličiny X_i , jestliže $P(X_i \leq q_\alpha) = \alpha$.



$$1 - \alpha = P(X_i > q_\alpha) = P(X_i > u) (1 - W_{\gamma, \sigma}(q_\alpha - u)).$$



$$q_\alpha = u + W_{\gamma, \sigma}^{-1} \left(1 - \frac{1 - \alpha}{P(X_i > u)} \right) = u + \frac{\sigma}{\gamma} \left[\left(\frac{P(X_i > u)}{1 - \alpha} \right)^\gamma - 1 \right].$$

- Tedy α -kvantil náhodné veličiny X_i je roven $\left(1 - \frac{1 - \alpha}{P(X_i > u)}\right)$ -kvantilu GPD rozdělení plus u .
- Pro α blízke 1 můžeme odhadnout

$$\hat{q}_\alpha = u + W_{\hat{\gamma}, \hat{\sigma}}^{-1} \left(1 - \frac{1 - \alpha}{\widehat{P(X_i > u)}} \right) = u + \frac{\hat{\sigma}}{\hat{\gamma}} \left[\left(\frac{N_u}{N(1 - \alpha)} \right)^{\hat{\gamma}} - 1 \right].$$

Odhad doby návratu

- Cílem je stanovit průměrnou frekvenci výskytu extrémního jevu, tj. jak často je překračována nějaká vysoká hranice.
- Frekventistická definice pravděpodobnosti: je-li $P(X_i > x) = p$, pak X_i překročí hranici x v průměru jednou za $\frac{1}{p}$ časových okamžiků.
- Předpokládejme, že máme danou hranici x .
- Označme k průměrnou frekvenci, tj. $k = \frac{1}{p}$ a hledejme jej tak, že platí

$$P(X_i > x) = \frac{1}{k}.$$

- Tedy

$$k = \frac{1}{P(X_i > x)} = \frac{1}{P(X_i > u) (1 - W_{\gamma, \sigma}(x - u))} = \frac{\left(1 + \frac{\gamma(x-u)}{\sigma}\right)^{\frac{1}{\gamma}}}{P(X_i > u)}.$$

- Tedy odhad doby návratu je

$$\hat{k} = \frac{N}{N_u (1 - W_{\hat{\gamma}, \hat{\sigma}}(x - u))} = \frac{N \left(1 + \frac{\hat{\gamma}(x-u)}{\hat{\sigma}}\right)^{\frac{1}{\hat{\gamma}}}}{N_u}.$$

Odhad úrovně návratu

- Cílem je stanovit hranici x , která je překračována v průměru jednou za k časových okamžiků.
- Opět začneme s frekventistickou definicí pravděpodobnosti:

$$P(X_i > x) = \frac{1}{k}.$$

- Potom x je $(1 - \frac{1}{k})$ -kvantil X_i , tj.

$$x = u + W_{\gamma, \sigma}^{-1} \left(1 - \frac{1}{kP(X_i > u)} \right).$$

- Tedy odhad úrovně návratu je

$$\hat{x} = u + W_{\hat{\gamma}, \hat{\sigma}}^{-1} \left(1 - \frac{N}{kN_u} \right) = u + \frac{\hat{\sigma}}{\hat{\gamma}} \left[\left(\frac{kN_u}{N} \right)^{\hat{\gamma}} - 1 \right].$$

Definition

Střední hodnotu překročení prahu u za podmínky, že k překročení došlo (mean excess) definujeme jako $e(u) = \mathbb{E}(X_i - u | X_i > u)$.

- Má-li náhodná veličina X_i GPD rozdělení s parametry σ a γ , pak $\mathbb{E}X_i = \frac{\sigma}{1-\gamma}$.
- Má-li náhodná veličina X_i GPD rozdělení s parametry σ a γ , pak náhodná veličina $X_i - u | X_i > u$ má GPD s parametry $\sigma + \gamma u$ a σ .
- Má-li náhodná veličina X_i GPD rozdělení s parametry σ a γ , pak $e(u) = \frac{\sigma + \gamma u}{1-\gamma}$.
- Tedy $e(u)$ je lineární funkcí v u (charakteristická vlastnost GPD rozdělení).

Volba prahu u II

- Pro naše data můžeme spočítat odhad: $\hat{e}(u) = \frac{\sum_{i=1}^{N_u} Y_i}{N_u}$, kde $Y_i = X_i - u$ pro $X_i > u$, $i = 1, \dots, N_u$.
- Vykreslíme graf $[x_{(i)}, \hat{e}(x_{(i)})]$ pro $i = 1, \dots, N$.
- Ten se v praxi nazývá mean excess plot.
- Pokud data pochází z GPD rozdělení, pak by graf měl být lineární.
- Hranici u určíme jako bod z grafu, odkud křivka vykazuje lineární závislost.

Modelování dvourozměrných dat

- Sdružené rozdělení náhodného vektoru $(X, Y)'$ jednoznačně určuje marginální rozdělení náhodných veličin X a Y .
- Opačně to ale neplatí.
- Cíl: Popsat, jak vypadají všechna sdružená rozdělení s předem danými marginálními rozděleními.

Definition

Funkce $C : [0, 1] \times [0, 1] \rightarrow [0, 1]$ se nazývá kopula, jestliže

- 1 $C(u, 0) = C(0, u) = 0, \forall u \in [0, 1]$.
- 2 $C(u, 1) = C(1, u) = u, \forall u \in [0, 1]$.
- 3 $C(u_1, v_1) - C(u_1, v_2) - C(u_2, v_1) + C(u_2, v_2) \geq 0,$
 $\forall 0 \leq u_1 \leq u_2 \leq 1, 0 \leq v_1 \leq v_2 \leq 1.$

Modelování dvourozměrných dat

Definition (Alternativní definice kopuly)

Funkce $C : [0, 1] \times [0, 1] \rightarrow [0, 1]$ se nazývá kopula, jestliže existuje pravděpodobnostní prostor (Ω, \mathcal{A}, P) a na něm definovaný náhodný vektor $(U, V)'$ takový, že

- 1 $U \sim \mathcal{R}(0, 1)$.
- 2 $V \sim \mathcal{R}(0, 1)$.
- 3 $C(u, v) = P(U \leq u, V \leq v)$ je sdružená distribuční funkce $(U, V)'$.

Příklady:

- 1 $C(u, v) = \Pi(u, v) = u \cdot v$ (*součinná kopula*; U a V jsou nezávislé).
- 2 $C(u, v) = M(u, v) = \min\{u, v\}$ (*horní kopula*; $V = U$).
- 3 $C(u, v) = W(u, v) = \max\{u + v - 1, 0\}$ (*dolní kopula*; $V = 1 - U$).

Modelování dvourozměrných dat

Theorem (Fréchetovy - Hoeffdingovy meze)

Pro každou kopulu C platí:

$$W(u, v) \leq C(u, v) \leq H(u, v), \quad \forall u, v \in [0, 1].$$

Theorem (Sklarova věta)

Nechť náhodný vektor $(X, Y)'$ má distribuční funkci F a marginální distribuční funkce F_1 a F_2 . Pak existuje taková kopula C , že

$$F(x, y) = C(F_1(x), F_2(y)), \quad \forall x, y \in \mathbb{R}.$$

Je-li F spojitá, pak C je určená jednoznačně.

- Předchozí věta nám dává návod, jak modelovat dvourozměrná rozdělení.
- Předepíšeme si marginální rozdělení a zvolíme vhodnou kopulu, která popisuje závislost složek (nezávisle na marginálních rozděleních).

Vztah kopul a korelačních koeficientů

- Pearsonův korelační koeficient

$$\rho = \rho(X, Y) = \frac{C(X, Y)}{\sqrt{D(X)D(Y)}}.$$

- Jeho hodnota závisí na marginálních rozděleních X a Y .
- Spearmanův korelační koeficient

$$\rho_S = \rho_S(X, Y) = \rho(F_1(X), F_2(Y)) = 12 \int_0^1 \int_0^1 C(u, v) du dv - 3.$$

- Kendalllovo τ

$$\begin{aligned} \tau &= \tau(X, Y) = P((X_1 - X_2)(Y_1 - Y_2) > 0) \\ &\quad - P((X_1 - X_2)(Y_1 - Y_2) < 0) = 4 \int_0^1 \int_0^1 C(u, v) c(u, v) du dv - 1, \end{aligned}$$

kde $(X_1, Y_1)'$ a $(X_2, Y_2)'$ jsou dvě nezávislé kopie $(X, Y)'$ a

$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v}$ je hustota kopuly C .

Příklady kopul

- ① Archimédovské kopuly
 - Gumbelova kopula
 - Joeova kopula
 - Claytonova kopula
 - Frankova kopula
- ② Eliptické kopuly
 - Normální (gaussovská) kopula
 - Studentova t kopula
- ③ Kopuly extrémních hodnot
 - Gumbelova kopula
 - Galambosova kopula
 - Tawnova kopula

Archimédovské kopuly

- Dají se většinou vyjádřit v uzavřeném tvaru.
- Většinou obsahují jeden parametr.
-

$$C(u, v) = \phi^{-1}(\phi(u) + \phi(v)),$$

kde $\phi : [0, 1] \rightarrow [0, \infty]$ je spojitá, klesající a konvexní funkce s $\phi(1) = 0$.

- Funkce ϕ se nazývá generátor.
- Pro hodnotu Kendallova τ platí:

$$\tau = 1 + 4 \int_0^1 \frac{\phi(u)}{\phi'(u)} du.$$

Součinnová (nezávislá) kopula

- Je Archimédovská kopula s generátorem $\phi(u) = -\log u$.



$$C(u, v) = u \cdot v.$$

- Pro hodnotu Kendallova τ platí: $\tau = 0$.

Gumbelova (Gumbelova - Hougardova) kopula

- Je Archimédovská kopula s generátorem $\phi(u) = (-\log u)^\theta$.
- $\theta \geq 1$ je parametr.

-

$$C(u, v) = \exp \left\{ - \left[(-\log u)^\theta + (-\log v)^\theta \right]^{\frac{1}{\theta}} \right\}.$$

- Pro hodnotu Kendallova τ platí: $\tau = 1 - \frac{1}{\theta}$.
- Pro $\theta = 1$ je Gumbelova kopula rovna součinnové.
- Pro $\theta \rightarrow \infty$ se Gumbelova kopula blíží horní kopule.

Joeova kopula

- Je Archimédovská kopula s generátorem $\phi(u) = -\log[1 - (1 - u)^\theta]$.
- $\theta \geq 1$ je parametr.
-

$$C(u, v) = 1 - [(1 - u)^\theta + (1 - v)^\theta - (1 - u)^\theta \cdot (1 - v)^\theta]^{\frac{1}{\theta}}.$$

- Pro $\theta = 1$ je Joeova kopula rovna součinové.
- Pro $\theta \rightarrow \infty$ se Joeova kopula blíží horní kopule.

Claytonova kopula

- Je Archimédovská kopula s generátorem $\phi(u) = \frac{1}{\theta} (u^{-\theta} - 1)$.
- $\theta \in [-1, \infty) \setminus \{0\}$ je parametr.

-

$$C(u, v) = \max\{(u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}, 0\}.$$

- Pro hodnotu Kendallova τ platí: $\tau = \frac{\theta}{\theta+2}$.
- Pro $\theta = -1$ je Claytonova kopula rovna dolní kopule.
- Pro $\theta \rightarrow 0$ se Claytonova kopula blíží součinnové kopule.
- Pro $\theta \rightarrow \infty$ se Claytonova kopula blíží horní kopule.

Frankova kopula

- Je Archimédovská kopula s generátorem

$$\phi(u) = -\log\left(\frac{e^{-\theta u} - 1}{e^{-\theta} - 1}\right).$$

- $\theta \in \mathbb{R} \setminus \{0\}$ je parametr.
-

$$C(u, v) = -\frac{1}{\theta} \log\left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right).$$

- Pro $\theta \rightarrow -\infty$ se Frankova kopula blíží dolní kopule.
- Pro $\theta \rightarrow 0$ se Frankova kopula blíží součinnové kopule.
- Pro $\theta \rightarrow \infty$ se Frankova kopula blíží horní kopule.

Normální (gaussovská) kopula

- Patří mezi eliptické kopuly.



$$C(u, v) = \Phi_{\rho}(\Phi^{-1}(u), \Phi^{-1}(v)).$$

- Φ^{-1} je kvantilová funkce standardizovaného normálního rozdělení $\mathcal{N}(0, 1)$.
- Φ_{ρ} je distribuční funkce dvourozměrného normálního rozdělení $\mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma})$, kde $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ je jeho varianční matice.
- $\rho \in [-1, 1]$ je parametr.
- Pro hodnotu Kendallova τ platí: $\tau = \frac{2}{\pi} \arcsin \rho$.
- Pro $\rho = -1$ je normální kopula rovna dolní kopule.
- Pro $\rho = 0$ je normální kopula rovna součinnové kopule.
- Pro $\rho = 1$ je normální kopula rovna horní kopule.

Studentova t kopula

- Patří mezi eliptické kopuly.



$$C(u, v) = t_{\nu, \rho}(t_{\nu}^{-1}(u), t_{\nu}^{-1}(v)).$$

- t_{ν}^{-1} je kvantilová funkce t rozdělení s ν stupni volnosti.
- $t_{\nu, \rho}$ je distribuční funkce dvourozměrného t rozdělení s ν stupni volnosti a varianční maticí $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.
- $\rho \in [-1, 1]$ a $\nu \in [1, \infty)$ jsou parametry.
- Pro hodnotu Kendallova τ platí: $\tau = \frac{2}{\pi} \arcsin \rho$.
- Pro $\rho = -1$ je Studentova kopula rovna dolní kopule.
- Pro $\rho = 0$ je Studentova kopula rovna součinné kopule.
- Pro $\rho = 1$ je Studentova kopula rovna horní kopule.

Kopuly extrémních hodnot

Jsou takové kopuly, pro které platí:

$$C(u^n, v^n) = C^n(u, v), \quad 0 \leq u, v \leq 1, \forall n = 1, 2, \dots$$

- 1 Gumbelova kopula

$$C(u, v) = \exp \left\{ - \left[(-\log u)^\theta + (-\log v)^\theta \right]^{\frac{1}{\theta}} \right\}, \quad \theta \geq 1 \text{ je parametr.}$$

- 2 Galambosova kopula

$$C(u, v) = u \cdot v \cdot \exp \left\{ \left[(-\log u)^{-\theta} + (-\log v)^{-\theta} \right]^{-\frac{1}{\theta}} \right\}, \quad \theta > 0 \text{ je parametr.}$$

- 3 Tawnova kopula

$$C(u, v) = u^{1-\alpha} \cdot v^{1-\beta} \cdot \exp \left\{ - \left[(-\alpha \log u)^\theta + (-\beta \log v)^\theta \right]^{\frac{1}{\theta}} \right\},$$

$\theta \geq 1, 0 \leq \alpha \leq 1, 0 \leq \beta \leq 1$ jsou parametry.

Generování dvourozměrných náhodných vektorů

- Naším úkolem je získat realizaci spojitého náhodného vektoru $(X, Y)'$ se sdruženou distribuční funkcí $F(x, y)$.
- Sklarova věta: $F(x, y) = C(F_1(x), F_2(y))$.
- Označme $c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v}$ hustotu kopuly $C(u, v)$.
- Hustota náhodné veličiny $U|V = v$ je rovna $c(u|v) = c(u, v)$.
- Distribuční funkce náhodné veličiny $U|V = v$ je rovna $\frac{\partial C(u, v)}{\partial v}$.

Generování dvourozměrných náhodných vektorů

- 1 Generujeme realizaci náhodné veličiny V z $\mathcal{R}(0, 1)$, označme ji v .
- 2 Generujeme realizaci náhodné veličiny U z podmíněného rozdělení $U|V = v$ s distribuční funkcí $\frac{\partial C(u,v)}{\partial v}$, označme ji u .
- 3 Tedy $(u, v)'$ je realizace kopuly C .
- 4 Hledanou realizaci $(x, y)'$ dostaneme jako $(x, y)' = (F_1^{-1}(u), F_2^{-1}(v))'$.

Modelování dvourozměrných dat

- **Model:** $(X_1, Y_1)', \dots, (X_n, Y_n)'$ je náhodný výběr z dvourozměrného rozdělení s distribuční funkcí $F(x, y)$.
- Naším cílem je odhadnout distribuční funkci $F(x, y)$.
- Sklarova věta: $F(x, y) = C(F_1(x), F_2(y))$, kde $F_1(x)$ je marginální distribuční funkce náhodných veličin X_i a $F_2(y)$ je marginální distribuční funkce náhodných veličin Y_i .
- **Parametrický přístup:** Zvolíme marginální rozdělení F_1 a F_2 a kopulu C a přidáme parametry.



$$F_{\theta}(x, y) = C_{\theta_3}(F_1(x, \theta_1), F_2(y, \theta_2)),$$

kde $\theta = (\theta'_1, \theta'_2, \theta'_3)'$ je vektor neznámých parametrů.

Modelování dvourozměrných dat - MLE



$$f_{\boldsymbol{\theta}}(x, y) = f_1(x, \boldsymbol{\theta}_1) f_2(y, \boldsymbol{\theta}_2) c_{\boldsymbol{\theta}_3}(F_1(x, \boldsymbol{\theta}_1), F_2(y, \boldsymbol{\theta}_2)),$$

kde $f_1(x, \boldsymbol{\theta}_1)$ je marginální hustota X_i , $f_2(y, \boldsymbol{\theta}_2)$ je marginální hustota Y_i a $c_{\boldsymbol{\theta}_3}(u, v)$ je hustota kopuly C .



$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^n \log f_{\boldsymbol{\theta}}(X_i, Y_i) = \sum_{i=1}^n \log f_1(X_i, \boldsymbol{\theta}_1) + \sum_{i=1}^n \log f_2(Y_i, \boldsymbol{\theta}_2) + \\ &+ \sum_{i=1}^n \log c_{\boldsymbol{\theta}_3}(F_1(X_i, \boldsymbol{\theta}_1), F_2(Y_i, \boldsymbol{\theta}_2)) = l_1(\boldsymbol{\theta}_1) + l_2(\boldsymbol{\theta}_2) + l_3(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3). \end{aligned}$$

- $\hat{\boldsymbol{\theta}} = \arg \max\{l(\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ je odhad parametru $\boldsymbol{\theta}$ metodou maximální věrohodnosti.

Modelování dvourozměrných dat - pseudoMLE

- Předchozí odhad je výpočetně náročný a nestabilní (dimenze θ je velká).
- Modifikace metody - pseudomaximálně věrohodný odhad (IFM; inference for marginal distributions).
- Nejprve odhadneme marginální rozdělení, tj. získáme maximálně věrohodné odhady parametrů θ_1 a θ_2 :

$$\hat{\theta}_1 = \arg \max l_1(\theta_1), \quad \hat{\theta}_2 = \arg \max l_2(\theta_2).$$

- Pro tyto pevné hodnoty $\hat{\theta}_1$ a $\hat{\theta}_2$ maximalizujeme $l_3(\theta_1, \theta_2, \theta_3)$:

$$\hat{\theta}_3 = \arg \max \{l_3(\hat{\theta}_1, \hat{\theta}_2, \theta_3); \theta_3\}.$$

- Potom $\hat{\theta} = (\hat{\theta}'_1, \hat{\theta}'_2, \hat{\theta}'_3)'$ je odhad parametru θ metodou pseudomaximální věrohodnosti.

Modelování dvourozměrných dat - CML

- Pro modelování můžeme využít i semiparametrický přístup (marginální rozdělení modelujeme neparametricky a kopulu parametricky).
- Marginální distribuční funkce odhadneme pomocí modifikovaných empirických distribučních funkcí $\hat{F}_1(x) = \frac{1}{n+1} \sum_{i=1}^n \mathbb{I}_{\{X_i \leq x\}}$ a $\hat{F}_2(y) = \frac{1}{n+1} \sum_{i=1}^n \mathbb{I}_{\{Y_i \leq y\}}$.
- Odhad parametru θ_3 příslušný kopule C metodou kanonické maximální věrohodnosti je

$$\begin{aligned}\hat{\theta}_3 &= \arg \max \sum_{i=1}^n \log c_{\theta_3}(\hat{F}_1(X_i), \hat{F}_2(Y_i)) \\ &= \arg \max \sum_{i=1}^n \log c_{\theta_3} \left(\frac{R_i}{n+1}, \frac{S_i}{n+1} \right),\end{aligned}$$

kde R_i je pořadí X_i mezi X_1, \dots, X_n a S_i je pořadí Y_i mezi Y_1, \dots, Y_n .

Modelování dvourozměrných dat - metoda inverze Kendallova τ

- Jedná se o semiparametrický přístup (marginální rozdělení modelujeme neparametricky a kopulu parametricky).
- Lze použít v případě, kdy kopula C závisí na jednorozměrném parametru θ a hodnotu Kendallova τ jsme schopni vyjádřit jako funkci θ , tj. $\tau = f(\theta)$.
- Marginální distribuční funkce odhadneme pomocí modifikovaných empirických distribučních funkcí $\hat{F}_1(x) = \frac{1}{n+1} \sum_{i=1}^n \mathbb{I}_{\{X_i \leq x\}}$ a $\hat{F}_2(y) = \frac{1}{n+1} \sum_{i=1}^n \mathbb{I}_{\{Y_i \leq y\}}$.
- Odhadneme hodnotu Kendallova τ z dat:

$$\hat{\tau} = \frac{\sum \sum_{i \neq j} \text{sign}(X_i - X_j) \text{sign}(Y_i - Y_j)}{n(n-1)} = \frac{2(a-b)}{n(n-1)},$$

kde a je počet konkordantních párů a b je počet diskordantních párů.

- Odhad parametru θ dostaneme jako řešení rovnice $f(\theta) = \hat{\tau}$.