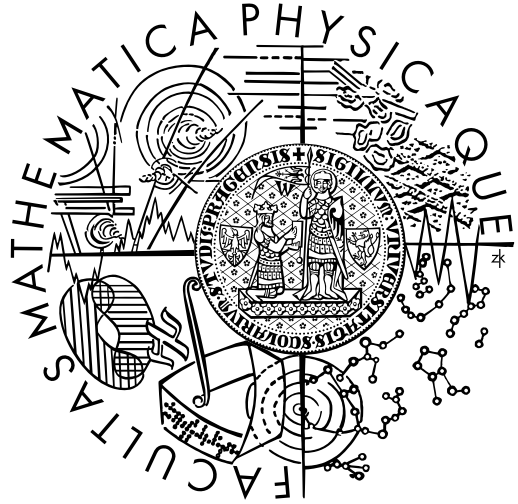
Analýza stratifikovaných dvoufázových studií s kalibrovanými a odhadnutými vahami

Michal Kulich

Katedra pravděpodobnosti a matematické statistiky
Matematicko-fysikální fakulta University Karlovy



Robust 2010 Králíky

Osnova

Semiparametrické modely

Pozorování

Máme nezávislá a stejně rozdělená pozorování
$X_1, \ldots, X_n \in (\mathcal{X}, \mathcal{A})$.
Každé X se skládá z

- Y, které nás zajímá a chceme jej modelovat
- Z, které nás nezajímá, ale může na něm záviset Y

Semiparametrické modely

Model

Nechť $X = (Y, Z)$ má rozdělení $P_0 = P_{\theta_0,\eta_0}$, kde

$$P_0 \in \mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta \subseteq \mathbb{R}^k, \eta \in \mathcal{H}\}$$

kde $P_{\theta,\eta}$ jsou absolutně spojité vzhledem k míře μ a $\mathcal{H}$ je podmnožina nějakého Banachova prostoru $\mathcal{B}$.

- θ je k-rozměrný vektor parametrů, které nás zajímají
- η je neznámá funkce, na níž závisí rozdělení dat

Semiparametrické modely

Příklady

- *Regrese s obecným rozdělením chyb*
  $Y = \mu_\theta(Z) + \sigma_\theta(Z)\varepsilon$, kde ε má nějaké rozdělení s nulovou střední hodnotou a jednotkovým rozptylem a μ_θ a σ_θ jsou známé funkce.

  Parametr θ: θ
  Parametr η: hustota ε
- *Regrese s chybami v prediktorech*
  $Y = \alpha + \beta U + \varepsilon$, $Z = U + \zeta$, kde $(\varepsilon, \zeta) \sim \mathrm{N}(0, \Sigma)$ nezávislé na U a rozdělení U není známo.

  Parametr θ: (α, β, Σ)
  Parametr η: hustota U

Semiparametrické modely

Příklady

- *Transformační regresní modely*
  $\eta(Y) = \beta^{\mathsf{T}} Z + \varepsilon$, kde ε má rozdělení $\mathrm{N}(0, \sigma^2)$ a η je nějaká rostoucí funkce.

  Parametr θ: (β, σ^2)
  Parametr η: transformace η
- *Zobecněné aditivní modely*
  $g(\mathrm{E}\, Y) = \alpha + \eta(Z_1) + \beta^{\mathsf{T}} Z_2$, kde rozdělení Y má známý tvar s dispersním parametrem ϕ, g je známá funkce a η je neznámá transformace Z_1.

  Parametr θ: (α, β, ϕ)
  Parametr η: transformace η

Semiparametrické modely

Příklady

- *Coxův model* $\lambda(t \mid Z) = \lambda(t) \exp\{\beta^{\mathsf{T}} Z\}$,
  kde $\lambda(t \mid Z)$ je podmíněná riziková funkce náhodné veličiny T^* a pozorovaná odezva je censorovaná zprava, tj. $Y = (T, \delta)$, $T = \min(T^*, C)$ a $\delta = I(T^* \leq C)$.

  Parametr θ: β
  Parametr η: $\int_0^t \lambda(s)\, ds$
- *Coxův model s intervalovým censorováním*
  $\lambda(t \mid Z) = \lambda(t) \exp\{\beta^{\mathsf{T}} Z\}$; pozorovaná odezva je intervalově censorovaná, tj. $Y = (T_L, T_U)$ a víme $T^* \in (T_L, T_U\rangle$.

  Parametr θ: β
  Parametr η: $\int_0^t \lambda(s)\, ds$

Odhady parametrů: úplná data

Předpoklady

Nechť hustotu $f_{\theta,\eta}$ rozdělení $P \in \mathcal{P}$ vzhledem k μ lze psát ve tvaru

$$f_{\theta,\eta}(x) = p_{\theta,\eta}(y,z)g_Z(z)$$

kde $g_Z(z)$ nezávisí ani na θ ani na η.

- Rozdělení Z je tedy možné zcela eliminovat ze všech úvah o odhadování θ a η.
- Soustředíme se vlastně na modelování podmíněného rozdělení Y, je-li dáno Z

Odhady parametrů: úplná data

Definice

Definice (Regulární parametrický model)

Model $\mathcal{P}$ indexovaný parametrem $\theta \in \Theta \subseteq \mathbb{R}^k$ se nazývá regulární právě když $\forall \theta \in \Theta$ platí:

1. θ je vnitřní bod Θ
2. Existuje vektor $\dot{s}(\theta)$ funkcí z $L_2(\mu)$ takový, že

$$\|\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \dot{s}(\theta)^\mathsf{T} h\| = o(|h|)$$

   pro $|h| \to 0$. (Fréchetova derivace funkcionálu $\theta \to \sqrt{p_\theta}$)
3. Matice $\int \dot{s}(\theta)\dot{s}(\theta)^\mathsf{T}\, d\mu$ je regulární

a zobrazení $\theta \mapsto \dot{s}_j(\theta)$ z Θ do $L_2(\mu)$ je spojité.

Odhady parametrů: úplná data

Definice

Definice

Nechť $\mathcal{P}$ je regulární parametrický model.

1. $\dot{\ell}(\theta) = 2\dfrac{\dot{s}(\theta)}{s(\theta)} I(s(\theta) > 0)$ se nazývá skórová funkce
2. $I(\theta) = 4 \int \dot{s}(\theta)\dot{s}(\theta)^{\mathsf{T}} \, d\mu$ se nazývá informační matice

Odhady parametrů: úplná data

Definice

Definice (Regulární parametrický odhad)

Odhad T_n parametru $\theta_0 = \nu(P_0) \in \mathbb{R}^k$ se nazývá [lokálně] regulární právě když pro každou posloupnost θ_n takovou, že $\sqrt{n}\,|\theta_n - \theta_0| < M < \infty \; \forall n$, platí

$$\mathcal{L}_{P_{\theta_n}}(\sqrt{n}(T_n - \nu(P_{\theta_n}))) \xrightarrow{d} \mathcal{L}_0$$

a limitní rozdělení nezávisí na θ_n.

Odhady parametrů: úplná data

Definice

Definice (Regulární parametrický submodel)

Každá podmnožina $\mathcal{Q}$ modelu $\mathcal{P}$, která má regulární euklidovskou parametrisaci, se nazývá regulární parametrický submodel modelu $\mathcal{P}$.

Definice (Regulární odhad)

Odhad T_n se nazývá [lokálně] regulární [v modelu $\mathcal{P}$], jestliže je [lokálně] regulární v každém regulárním parametrickém submodelu $\mathcal{Q} \subset \mathcal{P}$.

Odhady parametrů: úplná data

Zavedení odhadů

Označme

$$Pf = \mathrm{E}_P f(X) = \int f(x)\, dP, \quad \mathbb{P}_n f = \int f(x)\, d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

Maximálně věrohodné odhady $(\hat{\theta}, \hat{\eta})$ řeší soustavu rovnic

$$\hat{\Psi}_{n1}(\theta, \eta) \equiv \mathbb{P}_n \dot{\ell}_{\theta,\eta} = 0$$
$$\hat{\Psi}_{n2}(\theta, \eta)h \equiv \mathbb{P}_n B_{\theta,\eta} h = 0 \quad \forall h \in \mathcal{H}_0$$

kde $\mathcal{H}_0$ je nějaký vhodně zvolený systém funkcí vyjadřující směry, z nichž se prvky jednodimensionálních submodelů pro η přibližují k η_0.

$B_{\theta,\eta} : \mathcal{H}_0 \to L_2(P_0)$ se nazývá *skórový operátor*

Odhady parametrů: úplná data

Asymptotická linearita

Za jistých podmínek platí, že

$$\dot{\psi}_0 \sqrt{n}(\hat{\theta}_n - \theta_0, \hat{\eta}_n - \eta_0) = -\sqrt{n}\Psi_n(\theta_0, \eta_0) + o_P(1),$$

kde $\dot{\psi}_0$ je Fréchetova derivace zobrazení Ψ, jež musí být spojitě invertovatelná.

Odhady parametrů: úplná data

Asymptotická normalita $\hat{\theta}$

Dále

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{\psi}_0(X_i) + o_P(1)$$

kde

- $\widetilde{\psi}_0 = \widetilde{I}_0^{-1}(I - B_0(B_0^* B_0)^{-1} B_0^*)\dot{\ell}_0$,
- $\widetilde{I}_0 = P_0[(I - B_0(B_0^* B_0)^{-1} B_0^*)\dot{\ell}_0 \dot{\ell}_0^{\mathrm{T}}]$,
- $B_0 = B_{\theta_0, \eta_0}$,
- B_0^* je adjungovaný operátor k B_0.

Tudíž $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathrm{N}(0, \widetilde{I}_0^{-1})$

Dvoufázová studie

Motivace

Nyní uvažujeme situaci, kdy n je velké a měření některých složek Z je drahé.

- *Dvoufázová studie* spočívá v provedení náhodného výběru o daleko menším rozsahu z původního počtu pozorování n.
- Výběr jedinců $1, \ldots, n$ nazýváme *první fází* studie. Některé veličiny z $X = (Y, Z)$ pozorujeme na všech jedincích vybraných do první fáze (Y), některé však měřeny nejsou (drahé složky Z).
- Pomocí předepsaného známého mechanismu vybereme $m < n$ jedinců do *druhé fáze*.
- Drahé složky Z pozorujeme pouze na podvýběru (jedincích z druhé fáze).
- Pokud selekce výběru vhodným způsobem závisí na naměřených hodnotách Y, můžeme dosáhnout toho, že asymptotický rozptyl odhadovaných parametrů se příliš nezvýší.

Dvoufázová studie

Provedení výběru

Označme jako V diskrétní veličinu s hodnotami $1, \ldots, K$, která je transformací dat pozorovaných v 1. fázi.

- Nechť $\xi_1, \ldots, \xi_n$ jsou binární veličiny popisující výběr do druhé fáze a nechť podmíněné rozdělení ξ_i, je-li dáno $V = k$ je $\mathrm{Alt}(p_k)$.
- Nechť ξ_i je podmíněně nezávislé na X_i, je-li dáno V_i.
- Označme $n_k = \sum I(V_i = k)$, $m_k = \sum \xi_i I(V_i = k)$, $q_k = \mathrm{P}(V_i = k)$.
- Označme $\pi_i = \sum_{k=1}^{K} p_k I(V_i = k)$.

Dvoufázová studie

Odhady

Analýza dvoufázové studie probíhá zavedením vah do obvyklé odhadovací procedury.
Označme

$$\mathbb{P}_n^\pi f = \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} f(X_i)$$

a definujme odhady $(\widetilde{\theta}, \widetilde{\eta})$ jako řešení soustavy

$$\begin{aligned} \mathbb{P}_n^\pi \dot{\ell}_{\theta,\eta} &= 0 \\ \mathbb{P}_n^\pi B_{\theta,\eta} h &= 0 \quad \forall h \in \mathcal{H}_0. \end{aligned}$$

Dvoufázová studie

Vlastnosti odhadů

Platí

$$\begin{aligned}\sqrt{n}(\widetilde{\theta} - \theta_0) &= \sqrt{n}(\hat{\theta} - \theta_0) - \sqrt{n}(\widetilde{\theta} - \hat{\theta}) \\ &= \sqrt{n}\mathbb{P}_n\widetilde{\ell}_0 + \sqrt{n}(\mathbb{P}_n^{\pi} - \mathbb{P}_n)\widetilde{\ell}_0 + o_P(1).\end{aligned}$$

Lze ukázat, že

$$\sqrt{n}(\mathbb{P}_n - \mathbb{P}_0) \xrightarrow{w} \mathbb{G} \quad \text{na } \ell^{\infty}(\mathcal{F})$$

a

$$\sqrt{n}(\mathbb{P}_n^{\pi} - \mathbb{P}_0) \xrightarrow{w} \sum_{k=1}^{K} \sqrt{q_k}\sqrt{\frac{1-p_k}{p_k}}\mathbb{G}_k \quad \text{na } \ell^{\infty}(\mathcal{F}),$$

kde $(\mathbb{G}, \mathbb{G}_1, \ldots, \mathbb{G}_K)$ jsou vzájemně nezávislé Brownovy mosty.

Dvoufázová studie

Použití dat z první fáze

- Váhy $\frac{\xi_i}{\pi_i}$ eliminují veškerá data $W_i \subset (Y_i, Z_i)$ pozorovaná na subjektech, kteří vstoupili do první fáze, ale nebyli vybráni do fáze druhé (s výjimkou informace o stratech $V_i = v(W_i)$).
- Tato data W_i můžeme použít ke snížení rozptylu $(\widetilde{\theta}, \widetilde{\eta})$ tak, že s jejich pomocí provedeme malé perturbace ve vahách
- Existují dvě metody, jak to udělat:
  1. Metoda odhadnutých vah
  2. Metoda kalibrovaných vah

Dvoufázová studie

Metoda odhadnutých vah

Uvažujme logistický model pro pravděpodobnosti výběru

$$\mathrm{P}\left(\xi = 1 \mid W\right) = \pi(q(W); \alpha) = \frac{\exp\{\alpha^{\mathsf{T}} Q\}}{1 + \exp\{\alpha^{\mathsf{T}} Q\}},$$

kde Q je vektor prediktorů spočítaných z W.
Budiž $\pi(q(W_i); \alpha_0) = \pi_i \ \forall i$ [tj. stratum musí být obsaženo v Q jakožto kategoriální prediktor].

Na datech z 1. fáze najdeme maximálně věrohodné odhady $\hat{\alpha}_n$ parametru α_0 splňující

$$\sqrt{n}(\hat{\alpha}_n - \alpha_0) = \left\{P_0[\pi_0(1-\pi_0)]QQ^{\mathsf{T}}\right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\xi_i - \pi(Q_i; \alpha_0)] Z_i + o_P(1),$$

kde $\pi_0 = \pi(Q; \alpha_0)$.

Dvoufázová studie

Metoda odhadnutých vah

Zavedeme

$$\mathbb{P}_n^{\hat{\pi}} f = \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i}{\hat{\pi}_i} f(X_i)$$

kde

$$\hat{\pi}_i = \pi(Q_i; \hat{\alpha})$$

a předefinujme odhady $(\widetilde{\theta}, \widetilde{\eta})$ jako řešení soustavy

$$\mathbb{P}_n^{\hat{\pi}} \dot{\ell}_{\theta,\eta} = 0$$
$$\mathbb{P}_n^{\hat{\pi}} B_{\theta,\eta} h = 0 \quad \forall h \in \mathcal{H}_0.$$

Dvoufázová studie

Metoda odhadnutých vah

Dostaneme

$$\sqrt{n}(\widetilde{\theta} - \theta_0) \xrightarrow{d} \mathbb{G}\widetilde{\ell}_0 + \sum_{k=1}^{K} \sqrt{q_k}\sqrt{\frac{1-p_k}{p_k}}\mathbb{G}_k(\widetilde{\ell}_0 - p_k R Q),$$

kde

$$R = \left\{P_0[(1-\pi_0)]\widetilde{\ell}_0 Q^{\mathsf{T}}\right\}\left\{P_0[\pi_0(1-\pi_0)]QQ^{\mathsf{T}}\right\}^{-1}$$

Asymptotický rozptyl $\widetilde{\theta}$ je minimalizován, pokud Q obsahuje co nejlepší přiblížení k $\widetilde{\ell}_0$.

Uvažujme opět vektor veličin Q, který lze spočítat z dat W, jež jsou k disposici v první fázi. Součet

$$Q_+ = \sum_{i=1}^{n} Q_i$$

je znám přesně.

Upravme váhy tak, aby se co nejméně lišily od $w_i = 1/\pi_i$, ale aby zajistily přesný odhad známého součtu Q_+.

Dvoufázová studie

Metoda kalibrovaných vah

Nechť $\nu(x, y)$ je nějaká metrika na $\mathbb{R}$
[např. $\nu(x, y) = (x - y)^2/(2y)$ nebo $\nu(x, y) = x \log(x/y) - x + y$].

Hledáme nezáporná čísla $c_1, \ldots, c_n$ tak, aby minimalisovala

$$\sum_{i=1}^{n} \xi_i \nu(w_i, c_i w_i)$$

za podmínek $\sum_{i=1}^{n} \xi_i c_i w_i Q_i = Q_+$.

Problém se řeší metodou Lagrangeových multiplikátorů.

Zavedeme

$$\mathbb{P}_n^c f = \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i c_i}{\pi_i} f(X_i)$$

a předefinujme odhady $(\widetilde{\theta}, \widetilde{\eta})$ jako řešení soustavy

$$\begin{aligned} \mathbb{P}_n^c \dot{\ell}_{\theta,\eta} &= 0 \\ \mathbb{P}_n^c B_{\theta,\eta} h &= 0 \quad \forall h \in \mathcal{H}_0. \end{aligned}$$

Dostaneme

$$\sqrt{n}(\widetilde{\theta} - \theta_0) \xrightarrow{d} \mathbb{G}\widetilde{\ell}_0 + \sum_{k=1}^{K} \sqrt{q_k} \sqrt{\frac{1-p_k}{p_k}} \mathbb{G}_k(\widetilde{\ell}_0 - R_C Q),$$

kde

$$R_C = \left\{P_0 \widetilde{\ell}_0 Q^{\mathsf{T}}\right\}\left\{P_0 Q Q^{\mathsf{T}}\right\}^{-1}$$

a $R_C Q$ je projekce $\widetilde{\ell}_0$ do prostoru lineárních kombinací komponent Q metodou nejmenších čtverců.
Asymptotický rozptyl $\widetilde{\theta}$ je minimalizován, pokud Q obsahuje co nejlepší přiblížení k $\widetilde{\ell}_0$.

Software

Metody pro analýzu dvoufázových studií

- v Coxově modelu
- v zobecněných lineárních modelech

pomocí

- odhadovaných vah
- kalibrovaných vah

jsou implementovány v

knihovně `survey` balíku R.

Shrnutí

- dvoufázové studie lze propojit s většinou moderních parametrických a semiparametrických metod pro odhadování parametrů
- odhadované váhy a kalibrace umožňují použití veškerých dat pozorovaných v první fázi
- existují-li dobré predikce neúplně pozorovaných veličin, je rozptyl dvoufázových odhadů téměř stejný jako rozptyl odhadů na úplných datech, ovšem za zlomek ceny
- řada metod je implementována v R

Literatura

- Bickel, Klaassen, Ritov, and Wellner (1993) *Efficient and Adaptive Estimation in Semiparametric Models*. Johns Hopkins University Press.
- van der Vaart (1998) *Asymptotic statistics*. Cambridge University Press.
- Breslow and Wellner (2007) Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand J Stat*, **34**, 86–102.
- Breslow, Lumley, Ballantyne, Chambless, and Kulich (2009) Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences*, **1**, 32–49.