Bayesiánská analýza

IX. Modely kvalitativních a omezených vysvětlujících proměnných

Obsah tématu

Úvod

- Normální lineární regresní model – omezující (předpoklad normality).
- Kvalitativní vysvětlovaná proměnná.
- Omezená vysvětlovaná proměnná.
- Zavedení latentních dat (mají normální rozdělení).
- Příklady: ekonomie dopravy, ekonomie práce, analýza investiční aktivity firem.

Obsah tématu

Značení

- Vysvětlovaná proměnná $y^* = (y_1^*, \ldots, y_N^*)'$.

$$y_i^* = x_i'\beta + \epsilon_i.$$

- $x_i = (1, x_{i2}, \ldots, x_{ik})'$.
- Maticově:

$$y^* = X\beta + \epsilon.$$

Náhodná složka

1. ϵ z vícerozměrného normálního rozdělení se střední hodnotou 0_N a kovarianční maticí $h^{-1}I_N$,
2. všechny prvky matice X jsou pevná čísla (tj. nenáhodné veličiny). Pro náhodné veličiny jsou prvky X nezávislé na všech prvcích vektoru ϵ; $p(X|\lambda)$, kde λ neobsahuje β ani h.

Princip odhadu

- Pokud y^* pozorovatelné – standardní analýza.
- y^* obsahuje latentní data nějak propojena s y.
- Pro „funkčnost“ metod: $p(\beta, h|y^*, y) = p(\beta, h|y^*)$ (v případě přirozeně konjugované apriorní hustoty) resp. $p(\beta|y^*, y, h) = p(\beta|y^*, h)$ a $p(h|y^*, y, \beta) = p(h|y^*, \beta)$ (nezávislá apriorní hustota).
- Pokud pozorujeme y^*, nepřinese dodatečné pozorování y žádnou novou informaci.
- Standardní posteriorní simulace (Gibbsův vzorkovač): výběry z $p(\beta, h|y^*)$ a $p(y^*|y, \beta, h)$ resp. $p(\beta|y^*, h)$, $p(h|y^*, \beta)$ a $p(y^*|y, \beta, h)$.
- Vše kromě $p(y^*|y, \beta, h)$ umíme generovat.

Obsah tématu

Vztah nepozorovaných a pozorovaných dat

- Příklad požadovaných investic.

$$\begin{array}{lll} y_i = y_i^* & \text{pokud} & y_i^* > 0 \\ y_i = 0 & \text{pokud} & y_i^* \leq 0 \end{array}$$

- Pokud známe y^*, známe $y \Rightarrow p(\beta, h|y^*) = p(\beta, h|y, y^*)$.

Posteriorní hustota

- Nezávislost latentních proměnných (stejně jako pozorovaná):

$$\begin{array}{rll} y_i = y_i^* & \text{pokud} & y_i^* > 0 \\ y_i = 0 & \text{pokud} & y_i^* \leq 0 \end{array}$$

- Využíváme omezené normální rozdělení (vycházíme z předpokladu nepodmíněné normality y_i^*).

$$\begin{array}{rll} y_i^* = y_i & \text{pokud} & y_i > 0 \\ y_i^* | y_i, \beta, h \sim N(x_i'\beta, h^{-1}) 1(y_i^* < 0) & \text{pokud} & y_i = 0 \end{array}$$

- Standardní analýza + možnost zobecnění pro omezující bod c (rozšíření i pro neznámý parametr).

Empirická ilustrace

- BUDE ČASEM DOPLNĚNO!

Obsah tématu

Úvod

- Předpoklad rozhodování mezi dvěma alternativami.
- U_{ij} užitek jednotlivce i (pro $i = 1, \ldots, N$) z volby j (pro $j = 0, 1$).
- Pravidlo: volba 1 pokud $U_{1i} \geq U_{0i}$ a volba 0 jinak.
- Výběr závisí na rozdílu v užitcích:

$$y_i^* = U_{1i} - U_{0i}.$$

Probit model

- Diference v užitcí odpovídá normálnímu lineárnímu regresnímu modelu.
- Závislost na pozorovaných charakteristikách x_i.
- Random utility model.

$$\begin{aligned} y_i = 1 \quad &\text{pokud} \quad y_i^* \geq 0 \\ y_i = 0 \quad &\text{pokud} \quad y_i^* < 0 \end{aligned}$$

- Pokud známe y^*, známe $y \Rightarrow p(\beta, h|y^*) = p(\beta, h|y, y^*)$.

Posteriorní hustota

- Z nezávislosti:

$$p(y^*|y, \beta, h) = \prod_{i=1}^{N} p(y_i^*|y_i, \beta, h)$$

- Předpoklad normální lineární regrese $\rightarrow p(y_i^*|\beta, h)$ normální.
- Kombinace s informací o $y_i \rightarrow p(y_i^*|y_i, \beta, h)$:

$$\begin{array}{llll} y_i^*|y_i, \beta, h \sim N(x_i'\beta, h^{-1})1(y_i^* \geq 0) & \text{pokud} & y_i = 1 \\ y_i^*|y_i, \beta, h \sim N(x_i'\beta, h^{-1})1(y_i^* < 0) & \text{pokud} & y_i = 0 \end{array}$$

Pravděpodobnosti volby

- Pro dané parametry:

$$\begin{aligned}\Pr(y_i = 1|\beta, h) &= \Pr(y_i^* \geq 0|\beta, h) \\ &= \Pr(x_i'\beta + \epsilon_i \geq 0|\beta, h) = \Pr(\sqrt{h}\epsilon_i \geq -\sqrt{h}x_i'\beta|\beta, h)\end{aligned}$$

- Díky normalitě – poslední člen jedna mínus kumulativní distribuční funkce standardního normálního rozdělení (tj. $\sqrt{h}\epsilon_i$ odpovídá $N(0,1)$).
- Značení $\Phi(a)$ pro CDF $\rightarrow 1 - \Phi(-\sqrt{h}x_i'\beta)$.
- Standardní analýza (funkce parametrů).

Identifikační problém

- Více kombinací hodnot parametrů modelu vede ke stejné hodnotě věrohodnostní funkce.
- Probit: nekonečný počet hodnot parametrů β a h vede k témuž modelu.
- $\Pr(x_i{}'\beta + \epsilon_i \geq 0|\beta, h) = \Pr(x_i' c\beta + c\epsilon_i \geq 0|\beta, h)$ pro jakoukoli kladnou konstantu c.
- Transformovaná náhodná veličina $c\epsilon_i$ má rozdělení $N(0, c^2 h^{-1}) \rightarrow$ totožné probit modely s jinými koeficienty a přesností chyb.
- Alternativně: hodnoty věrohodnostní funkce stejné pro $(\beta = \beta_0, h = h_0)$ a $(\beta = c\beta_0, h = \frac{h_0}{c^2})$.
- Nelze rozlišit odděleně β a h (jenidentifikace $\beta\sqrt{h}$).
- Řešení: nastavení $h = 1$ (preferováno) nebo některý z β na 1 (apriorní kladný vliv této proměnné na pravděpodobnost!).

Empirická ilustrace

- BUDE ČASEM DOPLNĚNO!

Obsah tématu

Úvod

- Vysvětlované proměnné kvalitativní (dobrý, průměrný, slabý), ale uspořádatelné.
- Klíčový vztah mezi (vektory) y^* a y (y_i má hodnoty $j = 1, \ldots, J$; J je počet uspořádaných alternativ):

$$y_i = j \quad \text{pokud} \quad \gamma_{j-1} < y_i^* \le \gamma_j,$$

- $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_J)'$ je vektor parametrů, kde $\gamma_0 \le \ldots \le \gamma_J$.
- Normalita regresního modelu pro latentní data:

$$\begin{aligned}
\Pr(y_i = j|\beta, \gamma) &= \Pr(\gamma_{j-1} < y_i^* \le \gamma_j|\beta, \gamma) \\
&= \Pr(\gamma_{j-1} < x_i'\beta + \epsilon_i \le \gamma_j|\beta, \gamma) \\
&= \Pr(\gamma_{j-1} - x_i'\beta < \epsilon_i \le \gamma_j - x_i'\beta|\beta, \gamma).
\end{aligned}$$

- ϵ_i z $N(0, 1)$ (z důvodu identifikace $h = 1$):

$$\Pr(y_i = j|\beta, \gamma) = \Phi(\gamma_j - x_i'\beta) - \Phi(\gamma_{j-1} - x_i'\beta)$$

Problém identifikace

- Uspořádaný probit: pravděpodobnosti volby na základě normálního rozdělení a volba $\gamma_0, \ldots, \gamma_J$ pro rozdělení pravděpodobností mezi všechny možnosti volby.
- Potřeba více omezení: např. pro $J = 3$ máme normální rozdělení s volbou střední hodnoty $(x_i{}'\beta)$ a čtyři body (tj. γ_0, γ_1, γ_2 a γ_3).
- x_i jen úrovňová konstanta a chceme $\Pr(y_i = 1|\beta, \gamma) = 0.025$, $\Pr(y_i = 2|\beta, \gamma) = 0.95$ a $\Pr(y_i = 3|\beta, \gamma) = 0.025$.
- Řešení: $\beta = 0$, $\gamma_0 = -\infty$, $\gamma_1 = -1.96$, $\gamma_2 = 1.96$ a $\gamma_3 = \infty$ nebo $\beta = 1$, $\gamma_0 = -\infty$, $\gamma_1 = -0.96$, $\gamma_2 = 2.96$ a $\gamma_3 = \infty$ atd.
- Obvyklé řešení problému identifikace: $\gamma_0 = -\infty$, $\gamma_1 = 0$ a $\gamma_J = \infty$.

Problém identifikace a další intuice

- Alternativně: probit model pro $J = 2 \Rightarrow \gamma_0 = -\infty$, $\gamma_1 = 0$ a $\gamma_2 = \infty$.
- y^* jako užitek $\rightarrow$ pravděpodobnosti volby jako integrály na sekvenčních oblastech normálního rozdělení.
- Při mírném zvýšení užitku možnost přechodu jen do sousední kategorie (předpoklad uspořádání alternativ) $\times$ jinak multinomiální probit.

Bayesovská analýza I

- Gibbsův vzorkovač s obohacenými daty: $p(\beta|y^*, \gamma)$, $p(\gamma|y^*, y, \beta)$ a $p(y^*|y, \beta, \gamma)$.
- Standardní posteriorní hustoty pro β ($h = 1$), $p(y_i^*|y_i, \beta, \gamma)$:

$$y_i^*|y_i = j, \beta, \gamma \sim N(x_i'\beta, 1)1(\gamma_{j-1} < y_i^* \leq \gamma_j).$$

- Podmíněná hustota pro γ, $p(\gamma|y_i^*, y_i, \beta)$.
- Nepravá apriorní hustota (možnost i jiných priorů s mírnými modifikacemi výsledku): $p(\gamma_j) \propto c$ (zjednodušuje výběr γ v jednom běhu).
- Z volby $\gamma_0 = -\infty$, $\gamma_1 = 0$ a $\gamma_J = \infty$: $p(\gamma_j|y^*, y, \beta, \gamma_{(-j)})$ pro $j = 2, \ldots, J-1$.
- Označení $\gamma_{(-j)}$: vektor γ bez prvku γ_j.

$$\gamma_{(-j)} = (\gamma_0, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_J)'$$

Bayesovská analýza II

- $p(\gamma_j|y^*, y, \beta, \gamma_{(-j)})$ snadno odvoditelná $\rightarrow$.
  1. Hustota podmíněna vektorem $\gamma_{(-j)} \Rightarrow \gamma_j$ musí ležet v $[\gamma_{j-1}, \gamma_{j+1}]$.
  2. Hustota podmíněna vektorem y a $y^* \Rightarrow$ lze vyvodit jaké hodnoty latentních dat odpovídají příslušným hodnotám skutečných dat.
  3. V argumetnech podmíněné hustoty není přítomna žádná další informace o γ_j.
- Rovnoměrné rozdělení:

$$\gamma_j|y^*, y, \beta, \gamma_{(-j)} \sim U(\overline{\gamma}_{j-1}, \overline{\gamma}_{j+1})$$

- Pro $j = 2, \ldots, J-1$, kde

$$\overline{\gamma}_{j-1} = \max\left\{\max\left\{y_i^* : y_i = j\right\}, \gamma_{j-1}\right\}$$
$$\overline{\gamma}_{j+1} = \min\left\{\min\left\{y_i^* : y_i = j+1\right\}, \gamma_{j+1}\right\}$$

- $\max\left\{y_i^* : y_i = j\right\}$ označuje maximální hodnotu latentních dat mezi všemi jednotlivci, kteří si zvolili alternativu j (analogicky $\min\left\{y_i^* : y_i = j+1\right\}$).

Obsah tématu

Úvod

- Více alternativ volby.
- y_i pro $\{j = 0, \ldots, J\} \rightarrow J + 1$ alternativ, kdy $J > 1$.
- Motivace: U_{ji} je užitek i-tého jednotlivce volícího alternativu j (pro $i = 1, \ldots, N$ a $j = 0, \ldots, J$).
- Alternativa 0 jako základní volba a definujeme latentní proměnnou:

$$y_{ji}^* = U_{ji} - U_{0i}$$

  pro $j = 1, \ldots, J$.
- Multinomiální probit model předpokládá:

$$y_{ji}^* = x_{ji}' \beta_j + \epsilon_{ji}$$

- x_{ji} je k_j-rozměrný vektor obsahující vysvětlující proměnné, které ovlivňují užitek spojený s volbou j (relativně vzhledem k volbě 0), β_j je odpovídající vektor regresních koeficientů a ϵ_{ji} je chybový člen regrese.

Značení I

- J rovnic $\Rightarrow$ simulátor pro SUR model v kombinaci s metodami poskytujícími výběry pro latentní rozdíly užitků.
- Přepis do SUR modelu: $y_i^* = (y_{1i}^*, \dots, y_{Ji}^*)'$, $\epsilon_i = (\epsilon_{1i}, \dots, \epsilon_{Ji})'$,

$$\beta = \begin{pmatrix} \beta_1 \\ \cdot \\ \cdot \\ \beta_J \end{pmatrix} \quad X_i = \begin{pmatrix} x'_{1i} & 0 & \cdot & \cdot & 0 \\ 0 & x'_{2i} & 0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & 0 & x'_{Ji} \end{pmatrix}$$

- Definujeme $k = \sum_{j=1}^{J} k_j$ a

$$y_i^* = X_i \beta + \epsilon_i$$

Značení II

- Dále:

$$y^* = \begin{pmatrix} y_1^* \\ \cdot \\ \cdot \\ y_N^* \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \cdot \\ \cdot \\ \epsilon_N \end{pmatrix} \quad X = \begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ X_N \end{pmatrix}$$

- Model:

$$y^* = X\beta + \epsilon$$

Další předpoklady

- ϵ_i nezávisle a stejnoměrně rozděleny, $N(0, H^{-1})$ pro $i = 1, \ldots, N$, kdy H je matice přesností chyb rozměrů $J \times J$.
- Alternativně: ϵ odpovídá $N(0, \Omega)$, kde Ω je blokově diagonální matice rozměru $NJ \times NJ$:

$$\Omega = \begin{pmatrix} H^{-1} & 0 & \cdot & \cdot & 0 \\ 0 & H^{-1} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & 0 & H^{-1} \end{pmatrix}$$

- Vztah latentních a pozorovaných proměnných:

$$\begin{array}{lll} y_i = 0 & \text{pokud} & \max\left(y_i^*\right) < 0 \\ y_i = j & \text{pokud} & \max\left(y_i^*\right) = y_{ji}^* \geq 0 \end{array}$$

- $\max\left(y_i^*\right)$ je maximum J-rozměrného vektoru y_i^*.

Posteriorní hustota

- Gibbsův vzorkovač: $p(\beta|y^*, H)$ a $p(H|y^*, \beta)$, a jistou podobu vícerozměrného ohraničeného normálního rozdělení pro podmíněnou hustotu $p(y^*|y, \beta, H)$.
- Nezávislost chování mezi jendotlivci:

$$p(y^*|y, \beta, H) = \prod_{i=1}^{N} p(y_i^*|y_i, \beta, H)$$

- $p(y_i^*|\beta, H)$ odpovídá normální hustotě pravděpodobnosti + informace i y_i:

$$\begin{array}{rll} y_i^*|y_i, \beta, H \sim N(X_i'\beta, H^{-1})1(\max{(y_i^*)} < 0) & \text{pokud} & y_i = 0 \\ y_i^*|y_i, \beta, H \sim N(X_i'\beta, H^{-1})1(\max{(y_i^*)} = y_{ji}^* \geq 0) & \text{pokud} & y_i = j \end{array}$$

- Ekonometrická analýza mnoho let mimo oblast hlavního zájmu (jak z hlediska bayesovského, tak i klasického přístupu) $\Leftrightarrow$ výpočetní obtíže vztahující se k ohraničenému normálnímu rozdělení.

Bayesiánská analýza I

- β a H: nezávislá normální-Wishartova apriorní hustota (využití výsledků pro SUR model).
- Problém identifikace: jednorozměrný probit model nastavoval $h = 1$.
- Multinomiální probit model: složitější.
- Kovarianční matice chyb $\Sigma = H^{-1}$ a σ_{ij} jako ij-tý prvek matice $\Sigma \rightarrow$ standardní způsob řešení identifikovatelnosti volbou $\sigma_{11} = 1$.
- Za těchto podmínek $p(H|y^*, \beta)$ nebude odpovídat Wishartovu rozdělení a nelze tak využít výsledky analýzy SUR modelu.
- Možnost řešení (viz literatura): ignorovat problém a prezentovat výsledky pro $\frac{\beta}{\sigma_{11}}$.
- Práce s neidentifikovanými modely záludná $\rightarrow$ nebezpečná práce s neinformativními apriorními hustotami (výpočetní problémy).
- Obvyklá bayesovská analýza multinomiálního probit modelu s využitím informativní apriorní hustoty ovšem s ignorováním identifikačních omezení.

Bayesiánská analýza II

- McCulloch, Polson, Rossi (2000): ϵ_i odpovídá $N(0, \Sigma)$.
- Rozdělení vektoru ϵ_i do podoby

$$\epsilon_i = \begin{bmatrix} \epsilon_{1i} \\ \upsilon_i \end{bmatrix}$$

kde $\upsilon_i = (\epsilon_{2i}, \dots, \epsilon_{Ji})'$.

- Rozdělení matice Σ:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \delta' \\ \delta & \Sigma_\upsilon \end{bmatrix}$$

Bayesiánská analýza III

- Zákony pravděpodobnosti: $p(\epsilon_i) = p(\epsilon_{1i})p(\upsilon_i|\epsilon_{1i})$.
- Z vlastností vícerozměrného rozdělení:

$$\epsilon_{1i} \sim N(0, \sigma_{11})$$
$$\upsilon_i|\epsilon_{1i} \sim N(\frac{\delta}{\sigma_{11}}\epsilon_{1i}, \Phi),$$

- $\Phi = \Sigma_\upsilon - \frac{\delta\delta'}{\sigma_{11}}$.
- Místo s maticí Σ rozměru $J \times J$ pracujeme s parametry σ_{11}, δ a $\Phi \rightarrow$ nastavení $\sigma_{11} = 1$ a volba apriorní hustoty pro δ a Φ.

Bayesiánská analýza III

- Obvykle normální apriorní hustota pro δ a Wishartova apriorní hustotu pro Φ^{-1}.

$$
\begin{aligned}
p(\delta, \Phi^{-1}) &= p(\delta)p(\Phi^{-1}) \\
p(\delta) &= f_N(\delta|\underline{\delta}, \underline{V}_\delta) \\
p(\Phi^{-1}) &= f_W(\Phi^{-1}|\underline{\nu}_\Phi, \underline{\Phi}^{-1})
\end{aligned}
$$

- Podmíněné posteriorní hustoty:

$$
\begin{aligned}
p(\delta|y^*, \Phi, \beta) &= f_N(\delta|\overline{\delta}, \overline{V}_\delta) \\
p(\Phi^{-1}|y^*, \delta, \beta) &= f_W(\Phi^{-1}|\overline{\nu}_\Phi, \overline{\Phi}^{-1})
\end{aligned}
$$

Bayesiánská analýza III

- Posteriorní parametry:

$$\overline{V}_\delta = \left(\underline{V}_\delta^{-1} + \Phi^{-1} \sum_{i=1}^{N} \epsilon_{1i}^2 \right)^{-1}$$

$$\overline{\delta} = \overline{V}_\delta \left(\underline{V}_\delta^{-1} \underline{\delta} + \Phi^{-1} \sum_{i=1}^{N} \upsilon_i \epsilon_{1i} \right)$$

$$\overline{\Phi}^{-1} = \left[\underline{\Phi} + \sum_{i=1}^{N} (\upsilon_i - \epsilon_{1i}\delta)(\upsilon_i - \epsilon_{1i}\delta)' \right]^{-1}$$

$$\overline{\upsilon}_\Phi = \underline{\upsilon}_\Phi + N$$

- Podmíněné hustoty $\rightarrow \epsilon_i = (\epsilon_{1i}, \upsilon_i')$ známý vektor.
- Kritika multinomiálního probitu kvůli přeparametrizaci v důsledku mnoha alternativ (Σ) $\rightarrow$ nepřesné odhady.
- Informativní priory pro dodatečnou strukturu: např. diagonální Σ (pokud rozumné, zjednodušení výpočtu a řešení přeparametrizace).

Empirická ilustrace

- BUDE ČASEM DOPLNĚNO!

Obsah tématu

Varianty probit a tobit

- Panelová data pro probit:

$$y_{it}^* = x_{it}'\beta_i + \epsilon_{it}$$

- Metody z části věnované panelovým datům.
- Panelový multiniomiální probit model s náhodnými koeficienty: odvození v rámci multinomiálního probit modelu, modelu náhodných koeficientů a SUR modelu.
- Multinomiální časový probit model (*multinomial multiperiod probit model*): řešení problému autokorelace.
- Modulární podstata nástrojů (kombinovatelnost): nelineárnost vztahů, heteroskedasticita (jak pro probit tak i pro tobit).

Další varianty

- Lineární regresní modly s jiným rozdělením náhodných chyb.
- Vysvětlovaná proměnná počet: Poissonovo rozdělení.
- Vysvětlovaná proměnná doba trvání: Weibullovo rozdělení.
- Modely volby logit: logistické rozdělení.
- Řádový uspořádaný logit (rank ordered logit), multinomiální logit (preferován při více alternativách – výpočetní nenáročnost).

Nezávislost irelevantních alternativ

- Předpoklad použití multinomiálního logitu (ne vždy splněná vlasnost)!
- Podíly šancí se s přidáním alternativy nemění.
- Dopravní příklad: auto ($Y = 0$), věřejná doprava ($Y = 1$), kolo ($Y = 2$).
- Porušení: auto ($Y = 0$), červený autobus ($Y = 1$), modrý autobus ($Y = 2$).
- Řešení skrze vnořený (nested) logit model: nejdříve auto $\times$ hromadná doprava $\rightarrow$ po volbě hromadné dopravy logit pro červený $\times$ modrý autobus.