

Bayesiánská analýza

IX. Modely kvalitativních a omezených vysvětlujících proměnných

Obsah tématu

- 1 Jednorozměrné modely
- 2 Model omezených dat – tobit
- 3 Model binární volby – probit
- 4 Uspořádaný probit
- 5 Multinomiální probit
- 6 Rozšíření

- 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

- 1 **Jednorozměrné modely**
- 2 Model omezených dat – tobit
- 3 Model binární volby – probit
- 4 Uspořádaný probit
- 5 Multinomiální probit
- 6 Rozšíření

Značení

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

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

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

- 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

- 1 Jednorozměrné modely
- 2 Model omezených dat – tobit**
- 3 Model binární volby – probit
- 4 Uspořádaný probit
- 5 Multinomiální probit
- 6 Rozšíření

Vztah nepozorovaných a pozorovaných dat

- Příklad požadovaných investic.

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

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

Posteriorní hustota

- Nezávislost latentních proměnných (stejně jako pozorovaná):

$$y_i = y_i^* \quad \text{pokud } y_i^* > 0$$

$$y_i = 0 \quad \text{pokud } y_i^* \leq 0$$

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

$$y_i^* = y_i \quad \text{pokud } y_i > 0$$

$$y_i^* | y_i, \beta, h \sim N(x_i' \beta, h^{-1}) 1(y_i^* < 0) \quad \text{pokud } y_i = 0$$

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

Empirická ilustrace

- BUDE ČASEM DOPLNĚNO!

Obsah tématu

- 1 Jednorozměrné modely
- 2 Model omezených dat – tobit
- 3 Model binární volby – probit**
- 4 Uspořádaný probit
- 5 Multinomiální probit
- 6 Rozšíření

Úvod

- Předpoklad rozhodování mezi dvěma alternativami.
- U_{ij} užitek jednotlivce i (pro $i = 1, \dots, N$) 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 užitech:

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

Probit model

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

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

- 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 $y_i \rightarrow p(y_i^*|y_i, \beta, h)$:

$$y_i^*|y_i, \beta, h \sim N(x_i'\beta, h^{-1})1(y_i^* \geq 0) \quad \text{pokud } y_i = 1$$

$$y_i^*|y_i, \beta, h \sim N(x_i'\beta, h^{-1})1(y_i^* < 0) \quad \text{pokud } y_i = 0$$

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 (jen identifikace $\beta \sqrt{h}$).
- Řešení: nastavení $h = 1$ (preferováno) nebo některý z β na 1 (apriorní kladný vliv této proměnné na pravděpodobnost!).

Empirická ilustrace

- Viz Koop (2003).

Obsah tématu

- 1 Jednorozměrné modely
- 2 Model omezených dat – tobit
- 3 Model binární volby – probit
- 4 Uspořádaný probit**
- 5 Multinomiální probit
- 6 Rozšíření

Úvod

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

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

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

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

- 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 $\gamma_0, \dots, \gamma_J$ 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. $\gamma_0, \gamma_1, \gamma_2$ a γ_3).
- x_i jen úrovněová 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 $j = 2, \dots, J - 1$.
- Označení $\gamma_{(-j)}$: vektor γ bez prvku γ_j .

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

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 argumentech podmíněné hustoty není přítomna žádná další informace o γ_j .

- Rovnoměrné rozdělení:

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

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

$$\bar{\gamma}_{j-1} = \max \{ \max \{ y_i^* : y_i = j \}, \gamma_{j-1} \}$$

$$\bar{\gamma}_{j+1} = \min \{ \min \{ y_i^* : y_i = j + 1 \}, \gamma_{j+1} \}$$

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

Obsah tématu

- 1 Jednorozměrné modely
- 2 Model omezených dat – tobit
- 3 Model binární volby – probit
- 4 Uspořádaný probit
- 5 Multinomiální probit**
- 6 Rozšíření

Úvod

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

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

pro $j = 1, \dots, J$.

- 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 užiteků.
- 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 $i = 1, \dots, N$, 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:

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

- $\max(y_i^*)$ 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 :

$$y_i^*|y_i, \beta, H \sim N(X_i'\beta, H^{-1})1(\max(y_i^*) < 0) \quad \text{pokud } y_i = 0$$

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

- 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_j odpovídá $N(0, \Sigma)$.
- Rozdělení vektoru ϵ_j do podoby

$$\epsilon_j = \begin{bmatrix} \epsilon_{1j} \\ v_j \end{bmatrix}$$

kde $v_j = (\epsilon_{2j}, \dots, \epsilon_{Jj})'$.

- Rozdělení matice Σ :

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

Bayesiánská analýza III

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

$$\begin{aligned}\epsilon_{1i} &\sim N(0, \sigma_{11}) \\ v_i|\epsilon_{1i} &\sim N\left(\frac{\delta}{\sigma_{11}}\epsilon_{1i}, \Phi\right),\end{aligned}$$

- $\Phi = \Sigma_v - \frac{\delta\delta'}{\sigma_{11}}$.
- 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} .

$$p(\delta, \Phi^{-1}) = p(\delta)p(\Phi^{-1})$$

$$p(\delta) = f_N(\delta | \underline{\delta}, \underline{V}_\delta)$$

$$p(\Phi^{-1}) = f_W(\Phi^{-1} | \underline{\nu}_\Phi, \underline{\Phi}^{-1})$$

- Podmíněné posteriorní hustoty:

$$p(\delta | y^*, \Phi, \beta) = f_N(\delta | \bar{\delta}, \bar{V}_\delta)$$

$$p(\Phi^{-1} | y^*, \delta, \beta) = f_W(\Phi^{-1} | \bar{\nu}_\Phi, \bar{\Phi}^{-1})$$

Bayesiánská analýza III

- Posteriorní parametry:

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

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

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

$$\bar{v}_\Phi = \underline{v}_\Phi + N$$

- Podmíněné hustoty $\rightarrow \epsilon_i = (\epsilon_{1i}, v_i')$ 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

- Viz Koop (2003).

Obsah tématu

- 1 Jednorozměrné modely
- 2 Model omezených dat – tobit
- 3 Model binární volby – probit
- 4 Uspořádaný probit
- 5 Multinomiální probit
- 6 Rozšíření**

Varianty probit a tobit

- Panelová data pro probit:

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

- Metody z části věnované panelovým datům.
- Panelový multinomiální 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 vlastnost)!
- Podíly šancí se s přidáním alternativy nemění.
- Dopravní příklad: auto ($Y = 0$), veřejná 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.