

Bayesiánská analýza

III. Normální lineární regresní model s přirozeně konjugovanou apriorní hustotou (více vysvětlujících proměnných)

Obsah tématu

- 1 Lineární regresní model
- 2 Věrohodnostní funkce a apriorní hustota
- 3 Posteriorní hustota
- 4 Porovnání modelů
- 5 Intervaly nejvyšší posteriorní hustoty
- 6 Predikce
- 7 Monte Carlo integrace
- 8 Empirická ilustrace

- Koop (2003) – kapitola 3
- Normální lineární regresní model s více vysvětlujícími proměnnými.
- Použití maticové algebry.
- Přírozeně konjugovaný prior \Rightarrow analytické výsledky (podobně jako u modelu s jedinou vysvětlující proměnnou).

Obsah tématu

- 1 Lineární regresní model
- 2 Věrohodnostní funkce a apriorní hustota
- 3 Posteriorní hustota
- 4 Porovnání modelů
- 5 Intervaly nejvyšší posteriorní hustoty
- 6 Predikce
- 7 Monte Carlo integrace
- 8 Empirická ilustrace

LRM v maticovém vyjádření I

- k vysvětlujících proměnných x_{i1}, \dots, x_{ik} pro $i = 1, \dots, N$ a model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i.$$

- Úrovňová konstanta: $x_{i1} = 1$.
- Vektory $N \times 1$ a $k \times 1$:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_N \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \epsilon_N \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix}$$

LRM v maticovém vyjádření II

- Matice vysvětlujících proměnných rozměru $N \times k$

$$X = \begin{bmatrix} 1 & x_{12} & \cdot & \cdot & x_{1k} \\ 1 & x_{22} & \cdot & \cdot & x_{2k} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{N2} & \cdot & \cdot & x_{Nk} \end{bmatrix}$$

- Regresní model:

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

Předpoklady

- Předpoklady o ϵ a X determinují formu věrohodnostní funkce.
- Klasické předpoklady (později uvolněny):
 - 1 $\epsilon \sim N(0_N, h^{-1}I_N)$, kde $h = \sigma^{-2}$;
 - 2 všechny prvky X jsou nenáhodné veličiny.

Obsah tématu

- 1 Lineární regresní model
- 2 Věrohodnostní funkce a apriorní hustota**
- 3 Posteriorní hustota
- 4 Porovnání modelů
- 5 Intervaly nejvyšší posteriorní hustoty
- 6 Predikce
- 7 Monte Carlo integrace
- 8 Empirická ilustrace

Věrohodnostní funkce I

- Věrohodnostní funkce s využitím OLS odhadů:

$$\nu = N - k,$$

$$\hat{\beta} = (X'X)^{-1}X'y,$$

$$s^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{\nu}.$$

Věrohodnostní funkce II

- Věrohodnostní funkce:

$$p(y|\beta, h) = \frac{1}{(2\pi)^{\frac{N}{2}}} \left\{ h^{\frac{1}{2}} \exp \left[-\frac{h}{2} (\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) \right] \right\} \\ \times \left\{ h^{\frac{\nu}{2}} \exp \left[-\frac{h\nu}{2s^{-2}} \right] \right\}.$$

Apriorní hustota

- Normální-gama rozdělení, kdy $\beta|h$ z vícerozměrného normálního rozdělení, h je stále z gama rozdělení:

$$\begin{aligned}\beta|h &\sim N(\underline{\beta}, h^{-1}\underline{V}), \\ h &\sim G(\underline{s}^{-2}, \underline{\nu}).\end{aligned}$$

- Přirozeně konjugovaný prior pro β a h :

$$\beta, h \sim NG(\underline{\beta}, \underline{V}, \underline{s}^{-2}, \underline{\nu}).$$

- $\underline{\beta}$: k -rozměrný vektor apriorních středních hodnot pro k regresních koeficientů, β_1, \dots, β_k .
- \underline{V} : je pozitivně definitní apriorní část kovarianční matice rozměru $k \times k$.

Obsah tématu

- 1 Lineární regresní model
- 2 Věrohodnostní funkce a apriorní hustota
- 3 Posteriorní hustota**
- 4 Porovnání modelů
- 5 Intervaly nejvyšší posteriorní hustoty
- 6 Predikce
- 7 Monte Carlo integrace
- 8 Empirická ilustrace

Posterior I

- Posteriorní hustota – proporcionální součinu věrohodnostní funkce a apriorní hustoty
- Obdobné NLRM s jedinou vysvětlující proměnnou (použití matic).
- Posteriorní hustota parametrů:

$$\beta, h|y \sim NG(\bar{\beta}, \bar{V}, \bar{s}^{-2}, \bar{\nu}).$$

- Přičemž platí:

$$\begin{aligned}\bar{V} &= (\underline{V}^{-1} + X'X)^{-1}, \\ \bar{\beta} &= \bar{V}(\underline{V}^{-1}\beta + X'X\hat{\beta}), \\ \bar{\nu} &= \underline{\nu} + N.\end{aligned}$$

- \bar{s}^{-2} je implicitně definováno následovně:

$$\bar{\nu}\bar{s}^2 = \underline{\nu}\underline{s}^2 + \nu s^2 + (\hat{\beta} - \underline{\beta})'[\underline{V} + (X'X)^{-1}](\hat{\beta} - \underline{\beta}).$$

Vlastnosti posteriorní hustoty I

- Marginální posteriorní hustota pro β má vícerozměrné t -rozdělení:

$$\beta|y \sim t(\bar{\beta}, \bar{s}^2 \bar{V}^2, \bar{\nu}).$$

- Z definice t -rozdělení:

$$E(\beta|y) = \bar{\beta},$$
$$\text{var}(\beta|y) = \frac{\bar{\nu} \bar{s}^2}{\bar{\nu} - 2} \bar{V}.$$

Vlastnosti posteriorní hustoty II

- Platí intuice z NLRM s jednou vysvětlující proměnnou (jen vektory resp. matice).
- $\hat{\beta}$ je vektor.
- Matice $(X'X)^{-1}$: podobná role jako skalár $\frac{1}{\sum x_i^2}$ v jednoduchém LRM.
- \bar{V} : matice rozměru $k \times k$.
- Posteriorní střední hodnota parametru β , $\bar{\beta}$: maticově vážený průměr apriorní a datové informace.

Neinformativní prior I

- Nastavení $\underline{\nu} = 0$ a \underline{V}^{-1} „blížící“ se hodnotou k nule.
- Obvykle $\underline{V}^{-1} = cI_k$, kde c je skalár blížící se nule.
- Nepravá apriorní hustota:

$$p(\beta, h) \propto \frac{1}{h}.$$

Neinformativní prior II

- Získáme OLS odhady:

$$\beta, h|y \sim NG(\bar{\beta}, \bar{V}, \bar{s}^{-2}, \bar{\nu}),$$

kde

$$\bar{V} = (X'X)^{-1},$$

$$\bar{\beta} = \hat{\beta},$$

$$\bar{\nu} = N,$$

$$\bar{\nu s}^2 = \nu s^2.$$

Obsah tématu

- 1 Lineární regresní model
- 2 Věrohodnostní funkce a apriorní hustota
- 3 Posteriorní hustota
- 4 Porovnání modelů**
- 5 Intervaly nejvyšší posteriorní hustoty
- 6 Predikce
- 7 Monte Carlo integrace
- 8 Empirická ilustrace

Omezení ve tvaru nerovnosti

- Předpokládáme omezení ve tvaru nerovnosti:

$$R\beta \geq r.$$

- R je známá matice rozměru $J \times k$ a r je známý J -rozměrný vektor.
- Porovnáváme dva modely:

$$M_1 : R\beta \geq r,$$

$$M_2 : R\beta \not\geq r.$$

- Značení v definici M_2 znamená, že jedno nebo více z J omezení v M_1 není splněno.

Posteriorní podíl šancí I

- Snadný výpočet posteriorního podílu šancí, použití neinformativních priorů není problém:

$$PO_{12} = \frac{p(M_1|y)}{p(M_2|y)} = \frac{p(R\beta \geq r|y)}{p(R\beta \not\geq r|y)}.$$

- Posteriorní hustota pravděpodobnosti pro β je z vícerozměrného t -rozdělení $\Rightarrow p(R\beta|y)$ je z vícerozměrného t -rozdělení.

Omezení ve tvaru rovnosti

- **Případ 1:** M_1 obsahující omezení $R\beta = r$ a M_2 , u kterého toto omezení neplatí (vnořené modely – nested models).
- **Případ 2:** $M_1 : y = X_1\beta_{(1)} + \epsilon_1$ vzhledem k $M_2 : y = X_2\beta_{(2)} + \epsilon_2$, kde X_1 a X_2 jsou matice obsahující i různé vysvětlující proměnné (nevnořené modely – non-nested).
- Oba případy v definicích modelů pro $j = 1, 2$:

$$M_j : y_j = X_j\beta_{(j)} + \epsilon_j.$$

- Porovnání nevnořených modelů zahrnuje $y_1 = y_2$
- Porovnání vnořených modelů definuje M_2 jako neomezenou regresi a M_1 obsahuje omezení $R\beta = r$ (vyžaduje předdefinování vysvětlované a vysvětlujících proměnných).

Posterioční podíl šancí II

- Posterioční podíl šancí – využití marginálních věrohodností.
- Marginální věrohodnost:

$$p(y_j|M_j) = c_j \left(\frac{|\bar{V}_j|}{|\underline{V}_j|} \right)^{\frac{1}{2}} (\bar{\nu}_j \bar{s}_j^2)^{-\frac{\bar{\nu}_j}{2}},$$

pro $j = 1, 2$, kde

$$c_j = \frac{\Gamma\left(\frac{\bar{\nu}_j}{2}\right) (\underline{\nu}_j \underline{s}_j^2)^{\frac{\underline{\nu}_j}{2}}}{\Gamma\left(\frac{\underline{\nu}_j}{2}\right) \pi^{\frac{N}{2}}}$$

Posterioční podíl šancí III

- Posterioční podíl šancí porovnávající M_1 s M_2 :

$$PO_{12} = \frac{c_1 \left(\frac{|\bar{V}_1|}{|V_1|} \right)^{\frac{1}{2}} (\bar{\nu}_1 \bar{s}_1^2)^{-\frac{\bar{\nu}_1}{2}} p(M_1)}{c_2 \left(\frac{|\bar{V}_2|}{|V_2|} \right)^{\frac{1}{2}} (\bar{\nu}_2 \bar{s}_2^2)^{-\frac{\bar{\nu}_2}{2}} p(M_2)}.$$

- Posterioční podíl šancí závisí na apriorním podílu šancí, zvýhodňuje soulad modelu s daty, koherenci mezi apriorní a datovou informací a šetrnost pokud jde o počet vysvětlujících proměnných.

Porovnání modelů – neinformativní priority

- Neformální pravidlo: *použití neinformativních priorů pro společné parametry a čistě informativní priority pro všechny ostatní parametry.*
- Nastavením $\underline{\nu}_1 = \underline{\nu}_2 = 0$: posteriorní podíl šancí má rozumnou interpretaci (soulad modelu s daty, koherenci mezi apriorní a datovou informací, atd.)
- Rozumné užití neinformativního prioru pro přesnost chyby.
- Užití neinformativního prioru pro $\beta_{(j)}$ – vážné problémy v případě $k_1 \neq k_2$.
- Neinformativní prior pro $\beta_{(j)}$ založený na $\underline{V}_j^{-1} = cI_{k_j}$ a $c \rightarrow 0$.
 $|\underline{V}_j| = \frac{1}{c^{k_j}}$ se vzájemně nevyruší.
- Jesliže $k_1 < k_2$, PO_{12} je nekonečno, v případě $k_2 > k_1$, PO_{12} se blíží k nule bez ohledu na data.

Obsah tématu

- 1 Lineární regresní model
- 2 Věrohodnostní funkce a apriorní hustota
- 3 Posteriorní hustota
- 4 Porovnání modelů
- 5 Intervaly nejvyšší posteriorní hustoty**
- 6 Predikce
- 7 Monte Carlo integrace
- 8 Empirická ilustrace

Přípustné intervaly

- Volná analogie ke konfidenčním intervalům (testování hypotéz).
- Definice pro jediný regresní koeficient β_j .
- 95% přípustný interval pro β_j je interval $[a, b]$ takový že:

$$p(a \leq \beta_j \leq b|y) = 0.95.$$

- Existuje nekonečně mnoho přípustných intervalů.
- Např. $\beta_j|y \sim N(0, 1) \Rightarrow$ 95% přípustné intervaly $[-1.96, 1.96]$, $[-1.75, 2.33]$, $[-1.64, \infty]$ atd.

Highest Posterior Density Intervals

- 95% HPDI je 95% přípustný interval, který má nejmenší rozsah oproti jakémukoliv jinému 95% přípustnému intervalu.
- Např. $[-1.96, 1.96]$ je nejkratší přípustný interval.
- HPDI existuje vždy, když existuje posteriorní hustota pravděpodobnosti.
- Lze jej využít i při neinformativním prioru.

Obsah tématu

- 1 Lineární regresní model
- 2 Věrohodnostní funkce a apriorní hustota
- 3 Posteriorní hustota
- 4 Porovnání modelů
- 5 Intervaly nejvyšší posteriorní hustoty
- 6 Predikce**
- 7 Monte Carlo integrace
- 8 Empirická ilustrace

Prediční hustota

- Chceme předpovědět:

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

- Předpověď je založena na:

$$p(y^*|y) = \iint p(y^*|y, \beta, h) p(\beta, h|y) d\beta dh.$$

- Lze ukázat:

$$y^*|y \sim t(X^* \bar{\beta}, \bar{s}^2 \{I_T + X^* \bar{V} X^{*'}\}, \bar{\nu}).$$

Obsah tématu

- 1 Lineární regresní model
- 2 Věrohodnostní funkce a apriorní hustota
- 3 Posteriorní hustota
- 4 Porovnání modelů
- 5 Intervaly nejvyšší posteriorní hustoty
- 6 Predikce
- 7 Monte Carlo integrace**
- 8 Empirická ilustrace

Monte Carlo integrace pro NLRM

- Porovnání modelů, predikci a posteriorní analýzu týkající se β lze provést analyticky (netřeba posteriorních simulátorů).

Theorem (Monte Carlo integrace)

Nechť $\beta^{(s)}$ pro $s = 1, \dots, S$ je náhodný výběr (vzorek) z $p(\beta|y)$ a $g(\cdot)$ je funkce a definujme

$$\hat{g}_S = \frac{1}{S} \sum_{s=1}^S g(\beta^{(s)}),$$

potom \hat{g}_S konverguje k $E[g(\beta)|y]$ pro S jdoucí k nekonečnu.

Algoritmus

- 1 Vygenerujeme náhodný výběr $\beta^{(s)}$ z posteriorní hustoty pro β použitím generátoru náhodných čísel pro vícerozměrné t -rozdělení.
- 2 Spočítáme $g(\beta^{(s)})$ a zachováme tento výsledek.
- 3 Opakujeme předchozí kroky S -krát.
- 4 Spočítáme průměr S náhodných výběrů $g(\beta^{(1)}), \dots, g(\beta^{(S)})$.

Těmito kroky získáme odhad $E[g(\beta)|y]$ pro jakoukoliv funkci parametrů, která nás zajímá.

Numerická standardní chyba

- Monte Carlo integrace: aproximace $E[g(\beta)|y]$ (volbou S řídíme velikost chyby aproximace).
- Číselné měřítko chyby aproximace – využitím centrální limitní věty:

$$\sqrt{S} \{ \hat{g}_S - E[g(\beta)|y] \} \rightarrow N(0, \sigma_g^2),$$

pro S jdoucí k nekonečnu, přičemž $\sigma_g^2 = \text{var}[g(\beta)|y]$.

- Aproximativní 95% konfidenční interval pro $E[g(\beta)|y]$

$$\left[\hat{g}_S - 1.96 \frac{\hat{\sigma}_g}{\sqrt{S}}, \hat{g}_S + 1.96 \frac{\hat{\sigma}_g}{\sqrt{S}} \right].$$

- Alternativně lze využít i **numerickou standardní chybu (NSE)** $\frac{\hat{\sigma}_g}{\sqrt{S}}$ (obsahuje stejnou informaci v kompaktnější podobě).

Obsah tématu

- 1 Lineární regresní model
- 2 Věrohodnostní funkce a apriorní hustota
- 3 Posteriorní hustota
- 4 Porovnání modelů
- 5 Intervaly nejvyšší posteriorní hustoty
- 6 Predikce
- 7 Monte Carlo integrace
- 8 Empirická ilustrace**

Reálná data

- Reálná data o prodeji domů ve Windsdoru, Kanada v roce 1987, $N = 546$.
- Proměnné:
 - y_i = prodejní cena i -tého domu (v Kanadských dolarech),
 - x_{i2} = rozloha i -tého domu (ve čtverečních stopách),
 - x_{i3} = počet ložnic v i -tém domě,
 - x_{i4} = počet koupelen v i -tém domě,
 - x_{i5} = počet pater v i -tém domě.
- Hlavní soubory *chapter03.m* + skript *chapter03_neinf.m* + řada dalších podpůrných funkcí.

Reálná data

- Předpokládámě vliv rozlohy mezi 0 a 20 dolary za stopu čtvereční $\Rightarrow \text{var}(\beta_2) = 25$; pro ostatní koeficienty volíme $\text{var}(\beta_3) = 2500^2$ a $\text{var}(\beta_4) = \text{var}(\beta_5) = 5000^2$.

$$\text{var}(\beta) = \frac{\nu s^2}{\underline{\nu} - 2} \underline{V}.$$

- „Téměř“ neinformativní prior a informativní prior s $\underline{s}^{-2} = 4.0 \times 10^{-8}$, $\underline{\nu} = 5$ a

$$\underline{\beta} = \begin{bmatrix} 0.0 \\ 10 \\ 5000 \\ 10000 \\ 10000 \end{bmatrix} \quad \underline{V} = \begin{bmatrix} 2.40 & 0 & 0 & 0 & 0 \\ 0 & 6.0 \times 10^{-7} & 0 & 0 & 0 \\ 0 & 0 & 0.15 & 0 & 0 \\ 0 & 0 & 0 & 0.60 & 0 \\ 0 & 0 & 0 & 0 & 0.60 \end{bmatrix}.$$

Apriorní a posteriorní střední hodnoty pro β (směrodatné odchylky v závorkách)

	Prior	Posterior	
	Informativní	(Inf. prior)	(Neinf. prior)
β_1	0 (10000)	-4035.05 (3530.16)	-4009.55 (3590.11)
β_2	10 (5)	5.43 (0.37)	5.43 (0.37)
β_3	5000 (2500)	2886.81 (1184.93)	2824.61 (1210.43)
β_4	10000 (5000)	16965.24 (1708.02)	17105.18 (1728.18)
β_5	10000 (5000)	7641.23 (997.02)	7634.90 (1004.34)

Apriorní a posteriorní střední hodnoty pro h (směrodatné odchytky v závorkách)

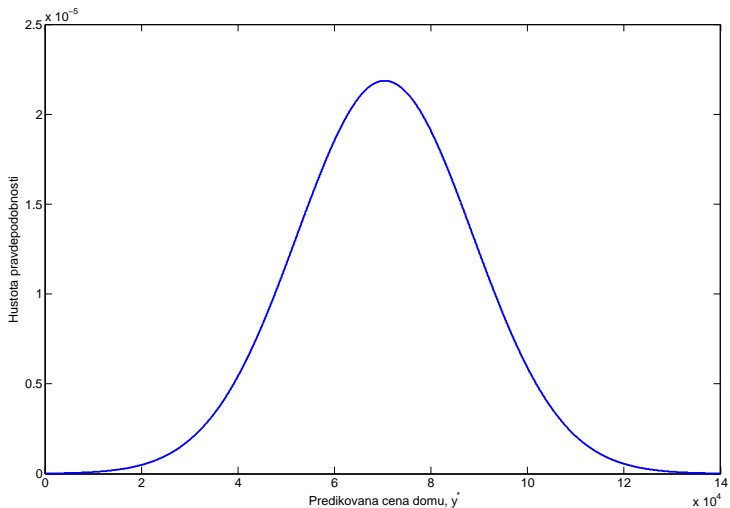
	Prior	Posterior	
	Informativní	(Inf. prior)	(Neinf. prior)
Stř. hodnota	4×10^{-8}	3.05×10^{-9}	3.03×10^{-9}
Sm. odchytky	2.53×10^{-8}	1.84×10^{-10}	1.83×10^{-10}

Porovnání modelů zahrnující parametr β

	$p(\beta_j > 0 y)$	95% HPDI	99% HPDI	Post. podíl šancí pro $\beta_j = 0$
<i>Informativní apriorní hustota</i>				
β_1	0.13	[-10969.27,2899.17]	[-13159.75,5089.64]	4.14
β_2	1.00	[4.71,6.15]	[4.49,6.38]	0.00
β_3	0.99	[559.29,5214.34]	[-175.96,5949.58]	0.39
β_4	1.00	[13610.20,20320.27]	[12550.37,21380.10]	0.00
β_5	1.00	[5682.82,9599.65]	[5064.16,10218.30]	0.00
<i>Neinformativní apriorní hustota*</i>				
β_1	0.13	[-11061.65,3042.54]	[-13289.45,5270.34]	—
β_2	1.00	[4.71,6.15]	[4.48,6.38]	—
β_3	0.99	[446.96,5202.27]	[-304.15,5953.38]	—
β_4	1.00	[13710.50,20499.85]	[12638.10,21572.25]	—
β_5	1.00	[5662.07,9607.73]	[5038.84,10230.96]	—

* Jako neinformativní apriorní hustota byla použita apriorní hustota blížící se čistě neinformativní hustotě. Podíl šancí tak bylo možno spočítat, ale není zde uváděn.

Predikční hustota pro dům o rozloze 5000 čtverečních stop.



Posterioční výsledky pro parametr β_2 spočítané alternativním způsobem

	Stř. hodnota	Směrodatná odchylka	NSE
Analyticky	5.43	0.37	—
Počet replikací			
$S = 10$	5.44	0.27	0.085
$S = 100$	5.43	0.38	0.038
$S = 1000$	5.44	0.36	0.011
$S = 10000$	5.44	0.36	0.004
$S = 100000$	5.43	0.37	0.001