Cvičení P04, od 6.12.2021:

Datový soubor `Cars2004.RData` (k přímému načtení do R) obsahuje údaje o automobilech dostupných v roce 2004 na trhu v USA. V dalším se budeme zabývat problémem vyhodnocení závislosti prodejní ceny (`price.retail`) na následujících charakteristikách jednotlivých aut: spotřeba ve městě (`cons.city`), spotřeba na dálnici (`cons.highway`), objem motoru (`engine.size`), koňská síla (`horsepower`), hmotnost (`weight`), obvod kola (`wheel.base`), délka (`length`), šířka (`width`), počet válců (`ncylinder`). Ze souboru předem vyřaďte vozy s hybridním, respektive rotačním motorem (`fhybrid` roven `"Yes"`, respektive `ncylinder` roven -1). Výsledná data by měla obsahovat údaje o $n = 423$ vozech.

Jako Y_i označme prodejní cenu $i$tého vozu v 1000 USD (tj. `price.retail` vydělené 1000), jako $\mathbf{X}_i = \big(X_{i,1}, \ldots, X_{i,p}\big)^\top$ $(p = 9)$ výše zmíněné charakteristiky. Uvažujte následující (hierarchický) model (primární parametry nejsou uváděny v podmínkách jednotlivých (podmíněných) rozdělení):

$$\begin{aligned} \mathbf{X}_i &\sim \mathcal{N}_p\big(\boldsymbol{\mu},\, \boldsymbol{\Sigma}\big), && i = 1, \ldots, n, \\ Y_i \,|\, \mathbf{X}_i &\sim \mathcal{N}\big(\beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta},\, \sigma^2\big), && i = 1, \ldots, n, \end{aligned}$$

$\big(Y_i,\, \mathbf{X}_i^\top\big)^\top$ nezávislé pro $i = 1, \ldots, n$. Primárním parametrům tedy odpovídají: β_0, $\boldsymbol{\beta} = \big(\beta_1, \ldots, \beta_p\big)^\top$, σ^2, $\boldsymbol{\mu} = \big(\mu_1, \ldots, \mu_p\big)^\top$, $\boldsymbol{\Sigma}$.

Jako apriorní rozdělení uvažte rozdělení založené na následujícím rozkladu (apriorní nezávislost):

$$p\big(\beta_0,\, \boldsymbol{\beta},\, \sigma^2,\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}\big) = p(\beta_0)\, p(\boldsymbol{\beta})\, p(\sigma^2)\, p(\boldsymbol{\mu})\, p(\boldsymbol{\Sigma}),$$

kde (náhodný hyperparametr λ je uveden explicitně v podmínce):

$$\beta_0 \sim \mathcal{N}(80,\, 100^2),$$

$$\boldsymbol{\beta} \,|\, \lambda \sim \mathcal{N}_p\big(\mathbf{0},\, \lambda^{-1}\, \mathrm{I}_p\big),$$
$$\lambda \sim \mathcal{G}(1,\, 0.005).$$

$$\mu_{\texttt{cons.city}} \sim \mathcal{N}(10,\, 10^2), \quad \mu_{\texttt{cons.highway}} \sim \mathcal{N}(10,\, 10^2), \quad \mu_{\texttt{engine.size}} \sim \mathcal{N}(3,\, 10^2),$$
$$\mu_{\texttt{horsepower}} \sim \mathcal{N}(200,\, 1\,000^2), \quad \mu_{\texttt{weight}} \sim \mathcal{N}(1\,500,\, 1\,000^2), \quad \mu_{\texttt{wheel.base}} \sim \mathcal{N}(250,\, 100^2),$$
$$\mu_{\texttt{length}} \sim \mathcal{N}(500,\, 100^2), \quad \mu_{\texttt{width}} \sim \mathcal{N}(10,\, 100^2), \quad \mu_{\texttt{ncylinder}} \sim \mathcal{N}(5,\, 10^2),$$

$$\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}_p(9,\, \mathrm{diag}(0.001, \ldots, 0.001)).$$

Pro reziduální rozptyl (či jeho přímou transformaci) uvažte postupně dvě možné apriorní volby:

(i) $\sigma^{-2} \sim \mathcal{G}(1,\, 0.005)$ (klasické gama rozdělení pro inverzní rozptyl);

(ii) $\sigma \sim \mathcal{U}(0.1,\, 100)$ (rovnoměrné rozdělení pro směrodatnou odchylku).

Pro každou z apriorních voleb proveďte následující kroky (pokud řešení některého z kroků nezávisí na volbě apriorního rozdělení pro parametr σ^2, stačí ho uvést jednou…).

1. Nakreslete (stačí rukou na papír) orientovaný acyklický graf (DAG) popisující uvažovaný model včetně apriorního rozdělení.

2. Odvoďte (stačí rukou na papír) plně podmíněnou hustotu (stačí tvar hustoty známý až na multiplikativní konstantu) pro vektorový parametr $\boldsymbol{\beta}$ (regresní koeficienty kromě absolutního členu), která by byla použita v rámci Gibbsova algoritmu, jestliže by vektor regresních koeficientů $\boldsymbol{\beta}$ byl generován najednou v rámci jednoho z kroků Gibbsova algoritmu. Je potřeba provádět dvě různá odvození při různých apriorních rozděleních pro reziduální rozptyl σ^2?

   Na základě odvozeného podiskutujte o roli parametru λ v uvažovaném modelu.

3. Implementujte výše uvedený model v JAGSu a vygenerujte dva markovské řetězce (pro každou z voleb apriorního rozdělení parametru σ^2), jejichž limitním rozdělením bude aposteriorní rozdělení pro uvažovaný model.

4. Nakreslete trajektorie (`traceplots`) pro parametry β_0, $\boldsymbol{\beta}$, σ, $\boldsymbol{\mu}$ a dále pro parametr λ a pro devianci modelu (kreslete oba řetězce do jednoho obrázku dvěma různými barvami). Nakreslete odhady autokorelačních funkcí pro regresní koeficienty β_0 a $\boldsymbol{\beta}$ (pro alespoň jeden z vygenerovaných řetězců).

   Posuďte, zda lze předpokládat konvergenci markovského řetězce k limitnímu rozdělení a zda řetězec vykazuje přijatelnou autokorelovanost.

5. Pro parametry β_0, všechny složky vektoru $\boldsymbol{\beta}$, parametr σ a parametr λ spočtěte základní charakteristiky aposteriorního rozdělení, 95% HPD věrohodnostní intervaly a nakreslete odhady aposteriorních hustot. Číselné hodnoty uveďte ve formě vhodné tabulky, ze které bude možné snadno porovnat výsledky při dvou apriorních volbách pro parametr σ^2. Taktéž aposteriorní hustoty kreslete tak, aby bylo možné snadno porovnávat výsledky získané při různých apriorních volbách.

   Které charakteristiky auta ovlivňují statisticky významně prodejní cenu auta, po očištění od možného vlivu zbývajících charakteristik?

6. Pro složky vektoru $\boldsymbol{\beta}$ uveďte klasické 95% intervaly spolehlivosti založené na odhadu normálního lineárního modelu metodou nejmenších čtverců. Liší se výrazně šířky některých/všech těchto intervalů od šířek HPD věrohodnostních intervalů? Jste schopni nalézt vysvětlení možných odlišností? Liší se nějak též závěry týkající se statistické významnosti vlivu jednotlivých charakteristik auta na prodejní cenu?

7. Spočtěte bayesovský bodový i intervalový (95%) odhad korelačního koeficientu mezi spotřebou ve městě a na dálnici, resp. mezi obvodem kola a délkou auta. Nakreslete odhady aposteriorních hustot pro oba tyto korelační koeficienty. Aposteriorní hustoty opět kreslete tak, aby bylo možné snadno porovnávat výsledky získané při různých apriorních volbách.

8. Pomocí metody Monte Carlo nakreslete marginální, tj. po vyintegrování charakteristik auta reprezentovaných náhodným vektorem $\mathbf{X}$, prediktivní hustotu prodejní ceny auta (reprezentované náhodnou veličinou Y). Spočtěte též související 95% HPD věrohodnostní interval (v tomto kontextu se jedná o analogii predikčního intervalu). Na základě posouzení prediktivní hustoty komentujte, zda uvažovaný model netrpí nějakými (zjevnými) nedostatky.

Deadline pro odevzdání vypracovaného úkolu (e-mailem na `komarek[AT]karlin.mff...`) je pátek 24.12. ve 12:24 CET.