Homework 3

Handed out: Monday, September 24, 2018
Due: Wednesday, October 3, 2018 Midnight

Notes:

- We *highly* encourage typed (Latex or Word) homework. Compile as single report containing solutions, derivations, figures, etc.
- Submit all files including report pdf, report source files (e.g. .tex or .docx files), data, figures produced by computer codes and programs files (e.g. .py or .m files) in a **.zip** folder. Programs should include a Readme file with running instructions.
- Zipped folder should be turned in on Sakai with the following naming scheme: **HW3_LastName_FirstName.zip**
- Collaboration is allowed however all submitted reports, programs, figures, etc. should be an individual student's write ups. Direct copying could be considered cheating.
- Software resources for this Homework set can be downloaded from this link or on Sakai under the Resource folder.

In this homework, for problems $1-7$, we will be using the caterpillar regression problem discussed in Chapter 3 of Bayesian Core by Jean-Michel Marin and Christian Robert[1] using the caterpillar data-set original published by Tomassone in 1993[2]. We suggest reading the entirety of Chapter 3 since this homework involves directly reproducing the results found in the text.

Code Base: For this homework, we will provide the base code for the all the problems found in the links below:

- [Matlab]
- [Python 2]
- [R] from Bayesian Core

The goal is to have an in depth understanding of the algorithms implemented.

Problem 1-5pts

The problem of interest consists of a scalar output y (the logarithmic of the number of caterpillars colonies in a $500m^2$ area) and $k = 10$ explanatory variables x_i ($i = 1, .., 10$) which correlate to features such as altitude, landscape slope, vegetation count, etc. The exact meaning does not matter in this context. Using the caterpillar.mat data, plot the semi-log-y plots of the caterpillar colony (last column of the data file) versus each feature reproducing Figure 3.1 in Bayesian Core.

[1] Jean-Michel Marin and Christian Robert. *Bayesian core: a practical approach to computational Bayesian statistics*. Springer Science & Business Media, 2007.

[2] Richard Tomassone. "Biométrie Modélisation de phénomenes biologiques". In: (1993).

Problem 2-10pts

Reproduce Figure 3.2 in Bayesian Core. For this table you will need calculate four different quantities:

(a) *Derive* and compute the MLE of the regression coefficients including a bias term. Comment on the dimensionality of $\boldsymbol{X}$, $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{y}$.

(b) The standard significance:

$$\hat{\sigma}^2 = \frac{s^2}{n-k-1}, \quad s^2 = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^T, \tag{1}$$

which approximates the covariance matrix of $\hat{\boldsymbol{\beta}}$.

(c) The standard *t-statistic* against null hypothesis:

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2\omega_{ii}}} \sim \mathcal{T}(n-k-1, 0, 1), \quad \omega_{ii} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}|_{(i,i)}, \tag{2}$$

where $\hat{\sigma}^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ approximates the co-variance matrix of $\hat{\boldsymbol{\beta}}$. For this problem β_i can be taken as 0.

(d) The p-value:

$$p_i = \Pr(|T_i| > |t_{values(i)}|), \tag{3}$$

which which can be found by adding the left and right values of the CDF of $\mathcal{T}(n-k-1, 0, 1)$. In the frequentist setting, the $p-value$ is commonly used as a probability of the null-hypothesis. However, viewing this idea from a Bayesian perspective what is the logical fallacy frequentists are committing here?

Problem 3-20pts

Reproduce Tables 3.1 and 3.2 in Bayesian Core. For this problem you will need to compute the following marginal statistics of the posterior:

$$\mathbb{E}_\pi(\sigma^2|\boldsymbol{y}, \boldsymbol{X}), \quad \mathbb{E}_\pi(\beta_i|\boldsymbol{y}, \boldsymbol{X}), \quad \mathrm{Var}_\pi(\beta_i|\boldsymbol{y}, \boldsymbol{X}), \tag{4}$$

for $i = 0, 1..., 10$ where β_0 is the bias term. For this we will use the conjugate priors:

$$p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{X}) \sim \mathcal{N}\left(\widetilde{\boldsymbol{\beta}}, \sigma^2\boldsymbol{M}^{-1}\right), \tag{5}$$

$$p(\sigma^2|\boldsymbol{X}) \sim \mathcal{IG}(a, b), \tag{6}$$

in which $\boldsymbol{M} = \frac{\boldsymbol{I}}{c} \in \mathbb{R}^{(k+1)\times(k+1)}$ where c is a constant. Consider the likelihood of:

$$p(\boldsymbol{y}|\beta, \sigma^2, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2), \tag{7}$$

which is the same you used in Problem 2. What are the PDFs for the conditional and marginal posteriors (Can use the reference for help)? Use the hyper-parameter values of $\widetilde{\beta}_i = 0$, $k = 10$, $a = 2.1$ and $b = 2.0$ for computing the table values. c will be governed by which table you are reproducing.

Problem 4-20pts

Reproduce Tables 3.3 and 3.4 in Bayesian Core using Zellner's informative G-prior:

$$p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{X}) \sim \mathcal{N}\left(\widetilde{\boldsymbol{\beta}}, c\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right), \tag{8}$$

$$p(\sigma^2|\boldsymbol{X}) \propto \sigma^{-2} \quad \text{(Improper Jeffery's Prior)}, \tag{9}$$

where we will again take $\widetilde{\beta}_i = 0$ and consider the cases of $c = 100$ and $c = 1000$. *Derive* and compute the following posterior statistics for $\beta_i, i = 1, ..., 10$:

$$\mathbb{E}_\pi(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{c+1}\left(\widetilde{\boldsymbol{\beta}} + c\hat{\boldsymbol{\beta}}\right) \tag{10}$$

$$\mathrm{Var}_\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) = \frac{c}{c+1}\frac{\left(s^2 + \frac{(\widetilde{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\widetilde{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}})}{c+1}\right)}{n-2}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}. \tag{11}$$

What form does the marginal posterior $p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y})$ take?

Additionally compute the *log*10 of the Bayes factors for the exclusion of each feature independently:

$$B_{10}^\pi = \frac{f(\boldsymbol{y}|\boldsymbol{X})}{f(\boldsymbol{y}|\boldsymbol{X}_0, H_0)} = \frac{(c_0+1)^{(k+1-q)/2}}{(c+1)^{(k+1)/2}}$$
$$\times\left[\frac{\boldsymbol{y}^T\boldsymbol{y} - \frac{c_0}{c_0+1}\boldsymbol{y}^T\boldsymbol{X}_0\left(\boldsymbol{X}_0^T\boldsymbol{X}_0\right)^{-1}\boldsymbol{X}_0^T\boldsymbol{y} + \frac{1}{c_0+1}\widetilde{\boldsymbol{\beta}}_0^T\boldsymbol{X}_0^T\boldsymbol{X}_0\widetilde{\boldsymbol{\beta}}_0 - \frac{2}{c_0+1}\boldsymbol{y}^T\boldsymbol{X}_0\widetilde{\boldsymbol{\beta}}_0}{\boldsymbol{y}^T\boldsymbol{y} - \frac{c}{c+1}\boldsymbol{y}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} + \frac{1}{c+1}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widetilde{\boldsymbol{\beta}} - \frac{2}{c+1}\boldsymbol{y}^T\boldsymbol{X}\widetilde{\boldsymbol{\beta}}}\right]^{n/2}, \tag{12}$$

where f denotes a density function and H_0 denotes the null-hypothesis for a specific model variable (You need not to derive this yourself).

Interpret what the Bayes factor is telling you, and *discuss* if you can pull any conclusions regarding certain features. What is the potential pitfall of using an improper prior in this setting?

Problem 5-5pts

Reproduce Table 3.5 by computing the 90% high posterior density (HPD) intervals for $\beta_i, i = 1, ..., 10$ (Bayesian Core wrongly indicates 95% HPD). Use again Zellner's information G-prior with $c = 100$.

Problem 6-20pts

Reproduce Table 3.6 by now using the uninformed Jeffery's prior:

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{X}) \propto \sigma^{-2}, \tag{13}$$

$$p(c) = c^{-1}\boldsymbol{I}_{N*}(c), \tag{14}$$

where we have assigned the hyper-parameter c with a simple prior. Again using $\widetilde{\beta}_i = 0$, *verify* the marginal takes the following form:

$$f(\boldsymbol{y}|\boldsymbol{X}) = \sum_{c=1}^{\infty} f(\boldsymbol{y}|\boldsymbol{X}, c)c^{-1} \propto$$

$$\sum_{c=1}^{\infty} c^{-1}(c+1)^{-(k+1)/2} \left[\boldsymbol{y}^T\boldsymbol{y} - \frac{c}{c+1}\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\right]^{-n/2}. \quad (15)$$

What is the advantage of using the uniformed prior?

Problem 7-20pts

Reproduce Table 3.7−3.11 by computing the most likely models using both Zellner's G-prior (with $c = 100$) and Zellner's uninformative G-prior. This is completed through Gibbs sampling for variable selection, a Markov chain Monte Carlo method, the follows the algorithm below:

Algorithm 1 Gibbs Sampler for Variable Selection

Initialization :← $\boldsymbol{\gamma}^0 = \left(\gamma_1^0, ..., \gamma_p^0\right)$
for t=1; t≤T; t++ **do**
    Given $\left(\gamma_1^{t-1}, ..\gamma_p^{t-1}\right)$
    1. γ_1^t according to $\pi_1(\gamma_1|\gamma_2^{t-1}, ..\gamma_p^{t-1})$
    2. γ_2^t according to $\pi_2(\gamma_2|\gamma_1^t, \gamma_3^{t-1}, ..\gamma_p^{t-1})$
    3. γ_3^t according to $\pi_3(\gamma_3|\gamma_1^t, \gamma_2^t, \gamma_4^{t-1}..\gamma_p^{t-1})$
    ...
    p. γ_p^t according to $\pi_p(\gamma_p|\gamma_1^t, ..\gamma_{p-1}^t)$
end for

where $\gamma_i = \{0, 1\}$ is a binary indicator that variable x_i is included in the model (I.e. $\boldsymbol{\gamma} = \{1, 1...1, 1\}$ indicates a *full* model where all model variables are used). What exactly is this algorithm doing? How do we determine if a model parameter is to be turned on or off? Explain the Gibbs sampler and its implementation in your own words.