**MATH3871/MATH5960**
**Assignment 2**

**Note:** This assignment is due in Week 12 at the start of the lecture.

Name(s): ________________________________

I (We) declare that this assessment item has not been submitted for academic credit elsewhere, and acknowledge that the assessor of this item may, for the purpose of assessing this item:

- Reproduce this assessment item and provide a copy to another member of the University; and/or,
- Communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking).

I (We) certify that I (We) have read and understood the University Rules in respect of Student Academic Misconduct.

Signed: ________________________________

Date: ______________

Note the following instructions for completing the assignment.

1. You must choose a **single** question from the list below and submit its solution individually (not a group assignment). All questions are worth the same marks (even though they may not be of the same difficulties for any one person).
2. Any computer code must be submitted through Moodle. The rest of the assignment (comments, figures, mathematical manipulations etc) must be submitted at the beginning of the lecture in week 12 as a printed document no longer than 4 pages (which includes this cover sheet). The fonts used must be not smaller than the font used to typeset these instructions.
3. To typeset the theoretical part of the assignment, you may work with the package `http://www.latex-project.org/`. Once you have completed installation, you may use the latex template provided in the Assignment folder on Moodle.

1. **Independence Sampler for Logit.** Example 6.2 on page 231 in the *Handbook of Monte Carlo method* book (available on Moodle) shows how to implement the random-walk Metropolis sampler for the Logit regression model. The example uses a computer simulated dataset. Repeat the simulations and Bayesian analysis using the *extramarital affairs* dataset on Moodle.

   (a) Estimate the posterior mean and the posterior covariance of the parameter $\boldsymbol{\beta}$.

   (b) Comment on the output. For example, which covariates seems to be relevant to whether a person has an affair or not?

   (c) Compare the posterior mean with the posterior mode. Is the posterior mode in this problem the same as the maximum likelihood estimate?

2. **Gibbs for Probit.** In lectures (Example 1.8) we examined the Gibbs/slice sampler for simulating from the posterior of the Probit regression.

   (a) Run the algorithm on the *extramarital affairs* dataset (on Moodle) and display boxplots of the marginal densities of vector $\boldsymbol{\beta}$ and an auto-correlations plot (of your choosing).

   (b) Comment on the output. For example, which covariates seem to be important in determining whether a person has an affair or not?

3. **Gibbs sampling for Truncated Normal.** Write a Matlab/R program to replicate the numerical results (the two figures) for Example 1.7, Page 20, in the MCMC lecture notes. Estimate $\mathbb{E}_\pi X_1$ and compute the error of the estimator using the batch-means variance estimate (equation (1.2) in the notes) with $m = 10^4$. You will need to determine a suitable burn-in parameter $b$ via experimentation.

4. **Hierarchical Zero-inflated Poisson model.** In the zero-inflated Poisson (ZIP) model, random data $X_1, \ldots, X_n$ are assumed to be of the form $X_i = R_i Y_i$, where the $\{Y_i\}$ have a $\mathsf{Poi}(\lambda)$ distribution and the $\{R_i\}$ a $\mathsf{Ber}(p)$ distribution, all independent of each other. Given an outcome $\tau = (x_1, \ldots, x_n)$, the objective is to estimate both $\lambda$ and $p$. Consider the following Bayes model:

   - $p \sim \mathsf{U}(0,1)$ (prior for $p$);
   - $(\lambda \mid p) \sim \mathsf{Gamma}(a,b)$ (prior for $\lambda$);
   - $(R_i \mid p, \lambda) \sim \mathsf{Ber}(p)$, independently (from the model above);
   - $(X_i \mid \boldsymbol{r}, \lambda, p) \sim \mathsf{Poi}(\lambda\, r_i)$, independently (from the model above),

   where $\boldsymbol{r} = (r_1, \ldots, r_n)$ and $a$ and $b$ are known parameters. It follows that

$$\pi(\tau, \boldsymbol{r}, \lambda, p) = \frac{b^a \lambda^{a-1} \mathrm{e}^{-b\lambda}}{\Gamma(a)} \prod_{i=1}^{n} \frac{\mathrm{e}^{-\lambda\, r_i} (\lambda\, r_i)^{x_i}}{x_i!} \, p^{r_i} (1-p)^{1-r_i} \ ,$$

We wish to sample from the posterior pdf $\pi(\lambda, p, \boldsymbol{r} \mid \tau)$ using the Gibbs sampler.

(a) Find the missing parameters

i. $(\lambda \mid p, \boldsymbol{r}, \tau) \sim \mathsf{Gamma}(?, ?)$

ii. $(p \mid \lambda, \boldsymbol{r}, \tau) \sim \mathsf{Beta}(?, ?)$

iii. $(r_i \mid \lambda, p, \tau) \sim \mathsf{Ber}(?)$

(b) Generate a random sample of size $n = 100$ for the ZIP model using parameters $p = 0.3$ and $\lambda = 2$.

(c) Implement the Gibbs sampler, generate a large (dependent) sample from the posterior distribution and use this to construct 95% Bayesian confidence intervals for $p$ and $\lambda$ using the data in b). Compare these with the true values.

5. **Poisson Regression.** In the Poisson regression we have count data $y_i$ which, given the parameter $\boldsymbol{\beta}$ and the covariate $\boldsymbol{x}_i$, is modelled as

$$(Y_i \mid \boldsymbol{\beta}, \boldsymbol{x}_i) \sim \mathsf{Poi}(\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))$$

for all $i$. The Bayesian likelihood is then

$$\pi(\boldsymbol{y} \mid \boldsymbol{\beta}, \mathbf{X}) = \exp\left(\sum_i y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - \ln(y_i!)\right) .$$

Therefore, using a flat improper prior $\pi(\boldsymbol{\beta}) \propto 1$ yields

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{y}, \mathbf{X}) \propto \exp\left(\sum_i y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right) .$$

Implement a random-walk sampler to draw $m = 10^4$ samples from the posterior density using the `RandDPatent.csv` dataset. The first column of the dataset is $\boldsymbol{y}$ (number of patents) and the second column is the money spent — the explanatory variable $x_i$ in the Poisson model with mean: $\mathbb{E}[Y \mid \boldsymbol{x}_i, \boldsymbol{\beta}] = \exp(\beta_1 + x_i \beta_2)$. Construct a scatterplot ($\beta_1$ versus $\beta_2$) from the output of the MCMC sampler.