Report submissions are accepted in the following formats, one text document for the solutions, **name_proj**. *Also submit an email with your R-files, with a file named* **name_proj.R** *that can be used to run your analysis, remember to submit* **all** *of the files you have created, and also that the codes are possible to run. Email the files to* **jonas.wallin@stat.lu.se**.

Discussion is permitted (and encouraged), as long as your answers and code reflects your own work.

Deadline: Thursday 10/1 at 23.59

Projects

For the home project choose one of the four projects below. Set up a suggested model (or code) and present (to me) at the time will be specified on the two last occasions. If you have any project of you're own choosing contact me and I will let you know if it is suitable.

Tobit regression

In 1978 C. Fair developed a model for extramarital affairs using Tobit regression. In the article comes from magazine survey (which is very unreliable so no real conclusions can be made from the data). The data is one the homepage in the file Affair_data.txt The data is $y_i = \frac{q_i}{v_i}$, where q_i is the number of times the persons (only females) had sexual relation outside marriage, and v_i is the number of years in the marriage. For covariates we have the age of the persons, number of years married, happiness (scale 1-5) and intercept. The original article is also available at the live@lund homepage.

After this you are supposed to fit a Tobit regression (Tobit model) which means that you truncate all data above and below a certain level. To fit the model using R-stan read:

- Read the section about Estimating Censored values in Stan manual section 12.3, on page 188. And use this to implement the Tobit regression for the affairs data, with $a = 0$ and $b = 5$.

In this project you are supposed to compare the effect of using Tobit regression to ordinary regression. Which dependent variables are important what and why is there a difference?

Poisson time series

The data set nyc-east-river-bicycle-counts.csv contains number bicycle crossing over three bridges in New York.

We are going to assume each observation follows a Poisson distribution with mean

$$\mu_i = \exp\left(\beta_{b[i]} + \beta^t_{b[i]} x_{t[i]} + \beta_{temp} temp_i + \epsilon_i\right).$$

Here $b[i]$ is the bridge for observation i. Further $\epsilon_i \sim N(0, \sigma_b)$ and $x_{t[i]}$ is time variable x at position $t[i]$ (there are four measurements for each date). F

$$x_t \sim N(ax_{t-1}, \sigma_x).$$

- Fit the model without the time series component x. Report the posterior distribution of the parameters σ_b, β_0, and β_{temp}. For *temp* use any version of the temperature you want.

- Read the section about time series in Stan manual section 12.3, on page 87.

- Fit the model with time series. Also present the posterior distribution a, σ_x.

Finally, suggest an improvement of the model (you don't need to implement).

Mixture model

In this exercise you are supposed to build a Bayesian model to determine how cheated on a test. In an, imaginary, course there where 400 hundred student taking an exam. After the exam there where reports of cheating and thus a second exam was conducted. On the homepage the file cheating.dat contains the results for the exams.

You are supposed to build a Bayesian model that:

- Generate a posterior distribution of the probability that a student has cheated or not. Create a figure that on the x-axis has the probability of falsely accusing a student of cheating and on the y-axis has number of student you accuse for that probability. (So if $p = 0$ then you can't accuse any student and if $p = 1$ you accuse all students)

- The cheaters in the file cheating.dat is stored in a pattern which?

- Your model also should estimate how much a student gained by cheating. You shall also generate a posterior distribution of the gain generated by cheating.

Hint: The maximum score is 200 on both exams. A combination of mixture model and multilevel model is recommended. To implement mixture model in Stan see Stan manual Chapter 13.

Metropolis Hastings random walk (requires knowledge about Multivariate Normal)

In this exercise you are going to build your own MCMC algorithm, more specifically a Metropolis hasting random walk. Here you are supposed to sample from the posterior distribution of α, β in the following model:

$$\begin{aligned} h_i &\sim N(\mu_i, 9.36), \\ \mu_i &= \alpha + w_i\beta, \\ \alpha &\sim N(170, 100), \\ \beta &\sim N(0, 20). \end{aligned}$$

Here w_i is weight and h_i is height. The data is **Howell1** that is in the rethinking package.

Since we have two parameters the proposal from the random walk is given by

$$\begin{bmatrix} \alpha^* \\ \beta^* \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I}),$$

where the left hand side is a two dimensional normal distribution.

- Write a Metropolis-Hastings random walk (MHRW) algorithm to samples from the posterior distribution of your model given the data. Check if the samples from the algorithm are reasonable.
- Test the MHRW uses one parameter, σ, which be tuned to improve the mixing of the MCMC algorithm. Examine various values of σ see how it effects the mixing of the MCMC algorithm. Visualize with a figure the relation the σ parameter and the autocorrelation between the consecutive samples in your MCMC algorithm.
- Finally test replacing $\mathbf{I}$ with $\mathbf{X}^T\mathbf{X}$ where $\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ w_1 & w_2 & \dots & w_n \end{bmatrix}$. Again refit σ comment on the results.