Report submissions are accepted in the following formats: One report file in pdf format denoted **name_lab4.pdf**. *Also submit an email with your R-files, with a file named* **name_lab4.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 between groups is permitted (and encouraged), as long as your answers and code reflects your own work.

Deadline: Wednesday 10/01 at 23.59

Bullets, indicates mandatory exercises, whereas star indicates voluntarily.

Advanced Bayesian models

1 The data contains number of stop and frisk done by the police in each precinct in New York. You have accesses to the data by race and population size in the precinct. For further details see Andrew Gelman blog.

We assume the following model:

$$\begin{aligned} Y_i &\sim Po(\theta_i), \\ \theta_i &= \exp(\log(pop_i) + \beta^p_{p_i} + \beta^e_{e_i} + \epsilon_i) \\ \epsilon_i &\sim N(0, \sigma^2), \\ \beta^p_j &\sim N(0, \sigma^2_{\beta^p}), \\ \beta^r_j &\sim N(0, \sigma^2_{\beta^e}). \end{aligned}$$

Here p_i is the precinct of the ith observation and β^p is the precinct effect. Further e_i is the ethnicity of the ith observations and thus β^e is the ethnicity effect. ϵ_i are an over dispersion effect.

- Download the data file stop.txt, which contains four columns:
  1. prec - precinct
  2. pop- population
  3. stops- number of stop and frisks
  4. eth- ethnicity where 1 is black, 2 is hispanic and 3 is white
- Fit a Poisson distribution without using the precinct, report the major factors. That is use Rstan to estimate the model:

$$\begin{aligned} Y_i &\sim Po(\theta_i), \\ \theta_i &= \exp(\log(pop_i) + \beta^e_{e_i}) \\ \beta^r_j &\sim N(0, \sigma^2_{\beta^e}), \end{aligned}$$

  with a vague prior on σ_{β^p}.
- Present the posterior distribution of β^e.
- Implement the full model.
- What is the distribution of β^e for the full model?
- In which district is it most likely and least likely to be stopped.

[10p]

2* 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. [10p]