Session 5: Stan examples

1 References

- Stan documentation:
  - Stan user's guide
  - Stan Reference manual
  - Stan functions reference
  - Tutorials
- Stan user forum

2 Exercise - Discoveries:

The file *evaluation_discoveries.csv* contains data on the numbers of "great" inventions and scientific discoveries (Xt) in each year from 1860 to 1959. In this question you will develop a model to explain the variation in scientific inventions over time. The simplest model here is to assume that (a) one discovery is independent of all others, and (b) the rate of occurrence of discoveries is the same in all years (λ) [Lambert]. Since the data is discrete, these assumptions suggest the use a Poisson likelihood,

$$Y \sim \mathcal{P}\wr\rangle\int\int\wr\backslash(\lambda)$$

1. Open a text editor and create a file called "discoveries.stan" in your working directory. In the file create three parameter blocks:

```
data {

}

parameters {

}

model {

}
```

2. Fill in the data and parameter blocks for the above model.
3. Using a $\updownarrow\wr\}\mathcal{N}(2,1)$ prior for λ code up the model block; making sure to save your file afterwards.
4. From RStudio, load any packages necessary to use Stan.

5. Read the data (`readr::read_csv()`) then put it into a structure that can be passed to Stan.
6. Run your model using Stan, with 4 chains, each with a sample size of 1000, and a warm-up of 500 samples. Set seed=1 to allow for reproducibility of your results. Store your result in an object called "fit".
7. Diagnose whether your model has converged by printing "fit"" ".
8. For your sample what is the equivalent number of samples for an independent sampler?
9. Find the central posterior 80% credible interval for λ.
10. Draw a histogram of your posterior samples for λ.
11. Load the evaluation_discoveries.csv data and graph the data. What does this suggest about our model's assumptions?
12. Create a generated quantities block in your Stan file, and use it to sample from the posterior predictive distribution. (Hint: use the function `poisson_rng` to generate independent samples from your lambda).

A more robust sampling distribution is a negative binomial model:

$$Y_i \sim NB(\mu, \kappa)$$

where μ is the mean number of discoveries per year, and $var(Y) = \mu + \mu^2/\kappa$. Here κ measures the degree of over-dispersion of your model; specifically if κ increases then over-dispersion decreases.

13. Write a new Stan file called "discoveries_negbin.stan" that uses this new sampling model (Hint: use the Stan manual section on discrete distributions to search for the correct negative binomial function name; be careful there are two different parameterisations of this function available in Stan). Assume that we are using the following priors:

$$\mu \sim \updownarrow\}\mathcal{N}(2, 1)$$

$$\kappa \sim \updownarrow\}\mathcal{N}(2, 1)$$

14. Draw 1000 samples across 4 chains for your new model. Has it converged to the posterior?

3 Exercise - Hungover holiday regressions:

The data in file *hangover.csv* contains a series of Google Trends estimates of the search traffic volume for the term "hangover cure" in the UK between February 2012 to January 2016. The idea behind this problem is to determine how much more hungover are people in the "holiday season" period, defined here as the period between 10th December and 7th January, than the average for the rest of the year.

1. Graph the search volume over time, and try to observe the uplift in search volume around the holiday season.
2. The variable "holiday" is a type of indicator variable that takes the value 1 if the given week is all holiday season, 0 if it contains none of it, and $0 < X < 1$ for a week that contains a fraction X of days that fall in the holiday season. Graph this variable over time so that you understand how it works.

3. A simple linear regression is proposed of the form,

$$V_t \sim \mathcal{N}(\beta_0 + \beta_0\, h_t, \sigma)$$

where V_t is the search volume in week t and h_t is the holiday season indicator variable. Interpret β_0 and β_1 and explain how these can be used to estimate the increased percentage of hangovers in the holiday season.

4. Assuming $\beta_i \sim \mathcal{N}(0, 50)$ and $\sigma \sim \langle \dashv \updownarrow \{ - \mathcal{N}(0, 10)$ priors write a Stan model to estimate the percentage increase in hangoverness over the holiday period.