4 Week 4

4.1 Newcomb's speed of light

R-template `ex_speed_of_light.Rmd`.
Data file `ex_speedOfLight.dat`.

(Here we redo the analysis from page 66 in BDA3.)

Simon Newcomb conducted experiments on speed of light in 1882. He measured the time required for light to travel a certain distance and here we will analyze a data recorded as deviations from $24{,}800$ nanoseconds. The model used in BDA3 is

$$\begin{aligned} y_i &\sim N(\mu, \sigma^2) \\ p(\mu, \sigma^2) &\propto \sigma^{-2}. \end{aligned}$$

where y_i is the i'th measurement, μ is the mean of the measurement and σ^2 the variance of the measurements. Notice that this prior is improper ("uninformative"). This corresponds to widely used uniform prior for μ in the range $(-\infty, \infty)$, and uniform prior for $\log(\sigma)$ (BDA3 pp. 66, 52, and 21). Both priors are improper and cannot be found from Stan. You can use instead

$$\begin{aligned} p(\mu) &\sim N(0, (10^3)^2) \\ p(\sigma^2) &\sim \text{Inv-}\chi^2(\nu = 4, s^2 = 1) \end{aligned} \tag{4}$$

In this exercise your tasks are the following:

1. Write a Stan model for the above model and sample from the posterior of the parameters. Report the posterior mean, variance and 95% central credible interval for μ and σ^2.
2. Additionally draw samples from the posterior predictive distribution of hypothetical new measurement $p(\tilde{y}|y)$. Calculate the mean, variance and 95% quantile of the posterior predictive distribution.
3. How does the posterior predictive distribution differ from the posterior of μ and Why?
4. Which parts of the model could be interpreted to correspond to aleatory and epistemic uncertainty? Discuss whether this distinction is useful here.
5. Instead of Inverse-χ^2 distribution the variance parameter prior has traditionally been defined using Gamma distribution for the precision parameter $\tau = 1/\sigma^2$. By using the results in Appendix A of BDA3 derive the analytic form of a Gamma prior for the precision corresponding to the prior (4). This should be of the form Gamma(α, β), where α and β are functions of ν and s^2.

Note! Many common distributions have multiple parameterizations, for which reason you need to be careful when interpreting others' works. The variance/precision parameter and their priors are notorious for this. The reason is mainly historical since different parameterizations correspond to different analytical solutions.

Grading: 2 points from correct answer for each of the above steps.

4.2 Co_2 concentration in Mauna Loa station

R-template `ex_linear_regression_MaunaLoaCO2_template.Rmd`.
Data file `maunaloa_data.txt`.

This is an example of linear regression and we will analyse the Mauna Loa CO2 data[1]. The data contains monthly concentrations adjusted to represent the 15th day of each month. Units are parts per million by volume (ppmv) expressed in the 2003A SIO manometric mole fraction scale. The "annual average" is the arithmetic mean of the twelve monthly values where no monthly values are missing.

We want to construct and infer with Stan the following model:

$$\begin{aligned} y_i &= \mu(x_i) + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2) \\ \mu(x_i) &= a + bx_i \\ p(a) &= p(b) \propto 1 \\ \sigma^2 &\sim \text{Inv-Gamma}(0.001, 0.001) \end{aligned}$$

where $y_i, i = 1, \cdots, n$ are the reported CO2 values, x_i is time, measured as months from the first observation, a is an intercept, b is the linear weight (slope) and σ^2 is the variance of the "error" terms, ϵ_i, around the linear mean function.

In practice, it is typically advisable to construct the model for standardized observations $\dot{y}_i = (y_i - \text{mean}(y))/\text{std}(y)$ where $\text{mean}(y))$ and $\text{std}(y)$ are the sample mean and standard deviations of y_i values. Similar transformation should be done also for covariates x. You should then sample from the posterior of the parameters $(\dot{a}, \dot{b}, \dot{\sigma}^2)$ corresponding to the standardized data $\dot{y}_i$ and $\dot{x}_i$. After this you have to transform the samples of $\dot{a}, \dot{b}, \dot{\sigma}^2$ to the original scale.

Your tasks are the following:

1. Solve the equations to transform samples of $\dot{a}, \dot{b}, \dot{\sigma}^2$ to the original scale a, b, σ^2.

2. Sample from the posterior of the parameters of the above model using the Maunaloa CO2 data. (You can do this either with transformed or original data so if you didn't get step 1 right you can still proceed with this.) Check the convergence of model parameters and report the results of convergence tests. Visualize the marginal posterior distribution of the model parameters and report their posterior mean and 2.5% and 97.5% posterior quantiles.

3. Discuss how you would interpret the linear mean function $\mu(x)$ and how you would intepret the error terms ϵ_i.

4. Plot a figure where you visualize

   - The posterior mean and 95% central posterior interval of the mean function $\mu(x)$ as a function of months from January 1958 to December 2027.

[1] `http://cdiac.esd.ornl.gov/ftp/trends/co2/maunaloa.co2`

- The posterior mean and 95% central posterior interval of observations y_i as a function of months from January 1958 to December 2027. In case of historical years, consider the distribution of potential replicate observations that have not been measured but could have been measured.
- plot also the measured observations to the same figure

5. Visualize
   - the Posterior predictive distribution of the mean function, $\mu(x)$ in January 2025 and in January 1958 and the difference between these.
   - the Posterior predictive distribution of observations, $\mathbf{y}_i$ in January 2025 and in January 1958 and the difference between these.
   - Discuss why the distributions of $\mu(x_i)$ and y_i differ

See the R-template for additional instructions.
Grading: This exercise is worth 20 points so that each step gives 4 points.