Co_2 concentration in Mauna Loa station

Week5-ex3, problem statement

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 mean(y)) and 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, 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

Total points 20. Each question gives 4 points as follows: 2 point for a solution that is towards the correct direction. 2 points more if the solution is correct. Note, in questions that require discussion, you should not give full points if the discussion is missing or if it is not adequate.