Using Probabilistic Programming Languages for Bayesian Inference

Package Loading

```
Pkg.activate(@__DIR__)
Pkg.instantiate()

using CSV
using DataFrames
using Plots
```

```
  Activating project at `~/Teaching/simulation-data-analysis/labs/mcmc-lab`
```

Overview

In this lab, you will use a probabilistic programming language to fit a linear model for monthly mean tide gauge data from the Sewell's Point, VA tide gauge from 1927 through 2022, obtained from the Permanent Service for Mean Sea Level. The data (in `data/norfolk-monthly-tide-data.txt`) has been slightly cleaned by setting dates to the `yyyy-mm` format. We've left missing values as `-99999`; make sure to fix those as appropriate for your programming language.

First, let's load the data. In Julia, we will replace `-99999` with `missing`.

```
tide_dat = CSV.read("data/norfolk-monthly-tide-data.txt", DataFrame)
# replace -99999 with missing
tide_dat.gauge = ifelse.(tide_dat.gauge .== -99999, missing, tide_dat.gauge)
```

```
1145-element Vector{Union{Missing, Int64}}:
 6950
 6935
 6917
    missing
    missing
 6627
 6734
 6789
 6777
 6911
 6838
 6819
 6896

 7288
 7130
 7169
 7250
 7463
 7346
 7256
 7303
 7424
 7457
 7299
 7302
```

Now let's plot the data.

```
scatter(tide_dat.datetime, tide_dat.gauge, xlabel="Month", ylabel="Monthly
 ↪  Mean Sea Level (mm)", legend=false)
```

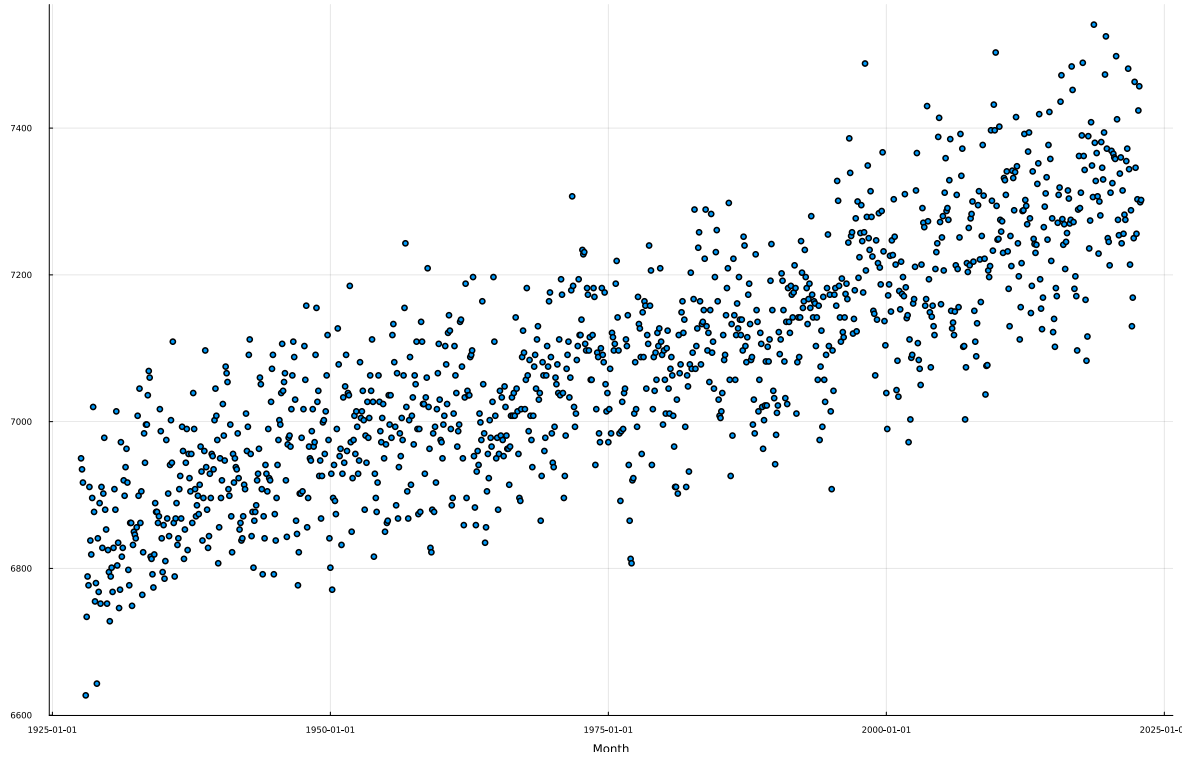

Figure 1: Plotted tide gauge data

We would like to quantify the uncertainty in the time-trend of this local sea level increase (which includes global mean sea level rise but also more local effects, such as subsidence). The plot in Figure 1 looks roughly linear, so let's use the following model (assuming the errors are independent and identically-distributed for simplicity):

$$y(t) = \alpha + \beta t + \varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, \sigma).$$

The lab involves the following steps:

1. First, write a model for the regression in () in the probabilistic programming language of your choice. You'll need to pick some priors for α, β, and σ.
2. Sample from the posterior with four chains (for convergence diagnostics).
3. Evaluate convergence. How many iterations did you use? What is the effective sample size?
4. Plot the posterior distributions. In particular, we are interested in uncertainty in the β coefficient, which reflects the mean increase in sea-level rise over time in mm/months.

5. Generate hindcasts by sampling from the posterior distribution and simulating data. If you plot the 95% posterior predictive distribution and the data, how does it look?