STAT 4880 # Group Lab 6 – Regression Week 11

The data `bayesrules::bikes` has information on daily ridership for registered riders in Washington DC's "Capitol Bikeshare" program, and is featured in Chapter 9. `bikes` is derived from the "Bike Sharing" data set at the U.C. Irvine machine learning repository. The BayesRules! data was cut down to 500 entries, so in this lab we'll instead use the full data `fullbikes.csv`, available on our course web page.

I. Bike Share Data

1. Make a time plot of registered rides with `date` on the x-axis. You will need to convert the `date` field to an actual date with `as.Date`. Color your points by the temperature, and if you want your graph to look good change the color scale with: `scale_color_viridis_c(option = "turbo")`
2. This lab explores the relationship between the temperature (`temp_actual`) and the total registered rides (`registered`) for that day. Make a scatterplot showing registered rides as a function of temperature. How do you explain the apparent bands of data?
3. Create a frequentist regression model (`lm`) to explain registered rides as a function of temperature. What is the equation of the regression line? Is the relationship significant?

II. Model with rstan

The linear regression model:

$$\begin{aligned} Y_i|\beta_0, \beta_1, \sigma &\sim N(\beta_0 + \beta_1 X_i, \sigma^2) \\ \beta_0 &\sim N(m_0, s_0^2) \\ \beta_1 &\sim N(m_1, s_1^2) \\ \sigma &\sim \text{Exp}(\lambda) \end{aligned}$$

where Y_i is the number of registered rides, X_i is the daily temperature, and β_0, β_1, σ are parameters.

1. The priors on β_0, β_1, σ require you to choose hyperparameters m_0, s_0, m_1, s_1 and λ. Make reasonable choices based on your work in part I.
2. Code and run this model in rstan.
3. Make a traceplot and check model diagnostics.
4. Compare your estimated parameter means to the values from the frequentist model in part I.3. Estimate $P(\beta_1 > 0|Y_i)$ and relate to the p-value from part I.3.

III. Model with rstanarm

1. If you use all the defaults, the syntax for regression in `rstanarm` is exactly the same as the frequentist `lm`, except you use the function `stan_glm`, a "stan generalized linear model". Create a model `bikes_mod` of registered rides on temperature using `stan_glm`.
2. Use `prior_summary` to see what priors were chosen. The "Adjusted prior" is the one that actually gets used. Check that the adjustment is multiplication by s_y/s_x for the slope parameter, and multiplication or division by s_y for the other two parameters.
3. Make a traceplot with `plot(bikes_mod, "trace")` Use `summary` to check the model diagnostics.

4. Use rstanarm's `posterior_predict` to generate a sample from the posterior predictive distribution for a 50-degree day and make a histogram of it.

5. Estimate $P(Y > 4000|Y_i, X = 50)$, the posterior probability of more than 4000 riders on a 50-degree day.