Fitting Bayesian Regression Models

Stat 340, Fall 2021

In this exericse we'll attempt to model the ridership among registered members of the Capital Bikeshare service in Washington, D.C. using the temperature (in °F). The data can be loaded using the below code:

```r
# Use temp_actual for temperature
data("bikes", package = "bayesrules")
```

Based on past bikeshare analyses, suppose we have the following prior understanding of this relationship:

- On an average temperature day, say 65 or 70 degrees for D.C., there are typically around 5000 riders, though this average could be somewhere between 3000 and 7000.
- For every one degree increase in temperature, ridership typically increases by 100 rides, though this average increase could be as low as 20 or as high as 180.
- At any given temperature, daily ridership will tend to vary with a moderate standard deviation of 1250 rides.

Tuning priors

(a) Tune a simple linear regression model to match our prior understanding. To translate the prior information for the intercept, it is easier to work with a *centered* version of temperature: $x_c = x - \overline{x}$. The intercept then tells us about the expected ridership on an average temperature day (i.e., when $x_c = 0$).

Use careful notation to write out the complete Bayesian structure of this model. (Hint: Consider using an exponential prior for σ, rather than placing a gamma prior on the precision.)

(b) To explore our combined prior understanding of the model parameters, simulate the Normal regression prior model for 1000 iterations. You can simulate this directly in R using the `rnorm` and `rexp` functions.

(c) Plot 100 prior plausible model lines ($\beta_0 + \beta_1 x_c$) and four data sets simulated under the priors. Since you used a centered version of temperature, you can either stick with this centered predictor, or "back transform" the plausible lines. To back transform, notice that $\beta_0 + \beta_1 x_c = (\beta_0 - \beta_1 \overline{x}) + \beta_1 x$.

To plot the model for the **centered temperature**, x_c, run the following code:

```r
# I assume you have beta0 and beta1 objects from (b)
library(tidyverse)
prior_lines <- data.frame(beta0 = beta0, beta1 = beta1) %>%
  mutate(group = row_number())

ggplot(prior_lines[1:100,]) +
  geom_abline(aes(intercept = beta0, slope = beta1, group = group), alpha = 0.5) +
  lims(x = c(-50, 50),     # Adjust x-axis limits for reasonable x_c
   y = c(20, 10000))   # Adjust y-axis limits
```

To plot the model for the **uncentered temperature**, x, run the following code:

```r
# I assume you have beta0 and beta1 objects from (b)
xbar <- mean(bikes$temp_actual)
prior_lines <- data.frame(beta0 = beta0 - beta1 * xbar , beta1 = beta1) %>%
  mutate(group = row_number())

ggplot(prior_lines[1:100,]) +
  geom_abline(aes(intercept = beta0, slope = beta1, group = group), alpha = 0.5) +
  lims(x = c(20, 90),     # Adjust x-axis limits for reasonable x_c
   y = c(0, 10000))   # Adjust y-axis limits
```

To create four simulated data sets under the model, use the 500 values of the predictor variable from the data set and simulate one response for each predictor.

Describe our overall prior understanding of the relationship between ridership and humidity.

EDA

With the priors in place, it's time to examine the data.

(a) Plot and discuss the observed relationship between ridership and humidity in the bikes data.

(b) Does a linear regression model seem to be a reasonable approach to modeling this relationship? Explain.

Model fitting

Now, let's simulate our posterior model of the relationship between ridership and temperature, a balance between our prior understanding and the data.

(a) Use JAGS to simulate the Normal regression posterior model. Do so with 5 chains run for 8000 iterations each. I've included a JAGS template, but you need to fill in the blanks.

```
post_model <- "model{
  # Likelihood
  for(i in 1:n) {
y[i] ~ ___(___, phi)
mu[i] <- beta0 + beta1 * x[i]
  }

  # Priors
  beta0 ~ ___(___, pow(___, -2))
  beta1 ~ ___(___, pow(___, -2))
  sigma ~ dexp(___)
  phi <- pow(___, -2)
}"

post_data <- list(
  n = nrow(bikes),
  x = bikes$temp_actual - mean(bikes$temp_actual),
  y = bikes$rides
)

posterior <- run.jags(
  post_model,
  data = post_data,
  n.chains = ___,
  adapt = 1000,
  burnin = 5000,
  sample = ___,
  monitor = c("beta0", "beta1", "sigma")
)
```

(b) Perform and discuss some MCMC diagnostics to determine whether or not we can "trust" these simulation results.

(c) Plot 100 posterior model lines for the relationship between ridership and temperature. Compare and contrast these to the prior model lines from above.

The below code chunk will help you get started.

```
posterior_lines <- posterior$mcmc[[1]][1:100,c("beta0", "beta1")] %>%
  as.data.frame()
```

Posterior inference

Finally, let's dig deeper into our posterior understanding of the relationship between ridership and humidity.

(a) Provide a summary of your posterior model, including 95% credible intervals.

(b) Interpret the posterior median value of the σ parameter.

(c) Interpret the 95% posterior credible interval for the temperature coefficient, β_1.

(d) Do we have ample posterior evidence that there's a positive association between ridership and temperature? Explain.

Prediction

Tomorrow is supposed to be 67°F in Washington, D.C. What levels of ridership should we expect?

(a) Simulate two posterior models: the posterior model for the typical number of riders on 67°F days; and the posterior predictive model for the number of riders tomorrow.

(b) Construct, discuss, and compare density plot visualizations for the two separate posterior models in the previous part.

(c) Calculate and interpret an 80% posterior prediction interval for the number of riders tomorrow.