

Fitting Bayesian Regression Models

Stat 340, Fall 2021

In this exercise 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:

```
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: $x_c = x - \bar{x}$. 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 .)

Solution: The prior information leads us to the following regression model:

$$\begin{aligned} Y_i | \beta_0, \beta_1, \sigma &\stackrel{\text{ind}}{\sim} \mu_i, \sigma \\ \mu_i &= \beta_0 + \beta_1 x_{ci} \\ \beta_0 &\sim \mathcal{N}(5000, 1000) \\ \beta_1 &\sim \mathcal{N}(100, 40) \\ \sigma &\sim \text{Expo}(.0008) \end{aligned}$$

- (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.

```
S <- 1000
beta0 <- rnorm(S, 5000, 1000)
beta1 <- rnorm(S, 100, 40)
sigma <- rexp(S, .0008)
```

Note: You can also simulate from the prior model using JAGS by passing in NAs for the response variable.

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

  # Priors
  beta0 ~ dnorm(5000, pow(1000, -2))
  beta1 ~ dnorm(100, pow(40, -2))
  sigma ~ dexp(.0008)
  phi <- pow(sigma, -2)
}"
```

```
prior_data <- list(
  n = nrow(bikes),
  x = bikes$temp_actual - mean(bikes$temp_actual),
  y = rep(NA, nrow(bikes))
)

prior_sim <- run.jags(
  prior_model,
  data = prior_data,
  n.chains = 1,
  sample = 1000,
  monitor = c("beta0", "beta1", "sigma", "y")
)
```

- (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 $\beta_0 + \beta_1 x_c = (\beta_0 - \beta_1 \bar{x}) + \beta_1 x$.

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

First, let’s plot the 100 prior plausible model lines. One way to do this is using `geom_abline` in `ggplot2`:

```
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),
       y = c(20, 10000))
```

An alternative way to do this is to fit a line through two points on the edge of plausible x_c values using `geom_segment()`. Let’s assume that centered temperature ranges from -50 to 50

```
prior_lines <- data.frame(beta0 = beta0, beta1 = beta1, x = -50, xend = 50) %>%
  mutate(y = beta0 + beta1 * x, yend = beta0 + beta1 * xend)

ggplot(prior_lines[1:100,]) +
  geom_segment(aes(x = x, y = y, xend = xend, yend = yend), alpha = 0.5)
```

Now, let’s simulate four data sets under the priors:

```
xc <- bikes$temp_actual - mean(bikes$temp_actual)
y1 <- rnorm(500, mean = beta0[1] + beta1[1] * xc, sd = sigma[1])
y2 <- rnorm(500, mean = beta0[2] + beta1[2] * xc, sd = sigma[2])
y3 <- rnorm(500, mean = beta0[3] + beta1[3] * xc, sd = sigma[3])
y4 <- rnorm(500, mean = beta0[4] + beta1[4] * xc, sd = sigma[4])

y1 <- prior_sim$mcmc[[1]][,4:503][3,]

prior_data <- bikes %>%
  select(rides, temp_actual) %>%
  mutate(
    xc = temp_actual - mean(temp_actual),
    prior_y1 = rnorm(length(rides), beta0[1] + beta1[1] * xc, sigma[1]),
    prior_y2 = rnorm(length(rides), beta0[2] + beta1[2] * xc, sigma[2]),
    prior_y3 = rnorm(length(rides), beta0[3] + beta1[3] * xc, sigma[3]),
    prior_y4 = rnorm(length(rides), beta0[4] + beta1[4] * xc, sigma[4]),
  )
```

```
p1 <- ggplot(NULL) + geom_point(aes(x = xc, y = y1))
p2 <- ggplot(NULL) + geom_point(aes(x = xc, y = y2))
p3 <- ggplot(NULL) + geom_point(aes(x = xc, y = y3))
p4 <- ggplot(NULL) + geom_point(aes(x = xc, y = y4))

p1 + p2 + p3 + p4 + plot_layout(nrow = 2)
```

Regardless of the plot, your discussion should center around whether the plot reflects your the prior understanding you were attempting to model.

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.

```
ggplot(bikes, aes(x = temp_actual, y = rides)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE)
```

- (b) Does simple Normal regression seem to be a reasonable approach to modeling this relationship? Explain.

Yes, there appears to be a postive linear association between temperature and ridership.

Model fitting

We can now 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.
- (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.

Posterior inference

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

- (a) Provide a tidy() 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 90% humidity 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 90% humidity 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