

# Individual HW07

Your Name Here

```
library(bayesrules) # R package for our textbook
library(tidyverse) # Collection of packages for tidying and plotting data
library(janitor) # Helper functions like tidy and tabyl
library(rstan) # for MCMC
library(bayesplot) # for plotting
library(broom.mixed) # for tidy() for mcmc
```

## 1 BR 10.3

There are several instances in which data scientists might falsely consider themselves neutral. In this exercise, you will practice how to challenge the false idea of neutrality, an underappreciated skill for data scientists.

- (a) How would you respond if your colleague were to tell you “I’m just a neutral observer, so there’s no bias in my data analysis?”
- (b) Your colleague now admits that they are not personally neutral, but they say “my model is neutral.” How do you respond to your colleague now?
- (c) Give an example of when your personal experience or perspective has informed a data analysis.

## 2 BR 10.7 (mini pp check)

Suppose we have a small dataset where predictor  $X$  has values  $\vec{x} = (12, 10, 4, 8, 6)$  and response variable  $Y$  has values  $\vec{y} = (20, 17, 4, 11, 9)$ . Based on this data, we built a Bayesian linear regression model of  $Y$  vs  $X$ .

- (a) In our first MCMC draw,  $(\beta_0^{(1)}, \beta_1^{(1)}, \sigma^{(1)}) = (-1.8, 2.1, 0.8)$ . Explain/show how you would use these values, combined with the data, to generate a prediction for  $Y_1$ .
- (b) Using the first MCMC draw, generate predictions for  $(Y_1, Y_2, Y_3, Y_4, Y_5)$ . Comment on the difference between the predictions and the observed values

### 3 BR 11.5 (interaction terms)

*Note:* I don't expect you to do this in R! I recommend sketching on a piece of paper and inserting the image into your .rmd/.qmd

- (a) Sketch a model that would benefit from including an interaction term between a categorical and quantitative predictor.
- (b) Sketch a model that would not benefit from including an interaction term between a categorical and quantitative predictor.
- (c) Besides visualization, what are two other ways to determine if you should include interaction terms in your model?

### 4 BR 11.11

You will use the `penguins_bayes` data in the `{bayesrules}` package to build various models of penguin `body_mass_g`. Throughout, we'll utilize weakly informative priors and a basic understanding that the average penguin weighs somewhere between 3,500 and 4,500 grams. Further, one predictor of interest is penguin species: Adelie, Chinstrap, or Gentoo.

```
data(penguins_bayes)
```

Building from the previous exercise, our next goal is to model `body_mass_g` by `flipper_length_mm` and `species` with an interaction term between these two predictors.

- (a) Use `stan_glm()` to simulate the posterior for this model, with four chains at 10,000 iterations each. Provide diagnostics of your MCMC
- (b) Simulate and plot 50 posterior model lines. Briefly describe what you learn from this plot.
- (c) Produce a `tidy()` summary for this model. Based on the summary, do you have evidence that the interaction terms are necessary for this model? Explain your reasoning. **(d) Simulate, plot, and describe the posterior predictive model for the body mass of an Adelie penguin that has a flipper length of 197.** **(e) Construct and discuss a `ppc_intervals` plot.**

### 5 BR 11.3

Consider 4 separate models of `body_mass_g`:

1. `body_mass_g ~ flipper_length_mm`
2. `body_mass_g ~ species`
3. `body_mass_g ~ flipper_length_mm + species`

4. `body_mass_g ~ flipper_length_mm + bill_length_mm + bill_depth_mm`

- (a) Simulate these four models using the `stan_glm()` function.
- (b) Produce and compare the `pp_check()` plots for the four models.
- (c) Use 10-fold cross-validation to assess and compare the posterior predictive quality of the four models using the `prediction_summary_cv()`. *NOTE:* We can only predict body mass for penguins that have complete information on our model predictors. Yet two penguins have NA values for multiple of these predictors. To remove these two penguins, we `select()` our columns of interest before removing penguins with NA values. This way, we don't throw out penguins just because they're missing information on variables we don't care about:

```
penguins_complete <- penguins_bayes |>
  select(flipper_length_mm, body_mass_g, species, bill_length_mm, bill_depth_mm) |>
  na.omit()
```

- (d) Evaluate and compare the ELPD posterior predictive accuracy of the four models.
- (e) In summary, which of these four models is “best?” Explain.

## 6 BR 11.14 (slightly modified)

Build two additional models of penguin `bill_length_mm`.

- 1. The best predictive model you can
- 2. A model that balances prediction and inference goals

You get to choose which predictors to use, and whether to include interaction terms. Evaluate and compare these models to the ones you built in the previous problem. Which do you prefer and why?

## 7 BR 12.5

You will explore how the number of eagle sightings in Ontario, Canada has changed over time. Since this context is unfamiliar to us, we'll utilize weakly informative priors throughout. We'll balance this prior uncertainty by the `bald_eagles` data in the `{bayesrules}` package, which includes data on bald eagle sightings during 37 different one-week observation periods.

- (a) Construct and discuss a univariate plot of count, the number of eagle sightings across the observation periods.
- (b) Construct and discuss a plot of count versus year.

- (c) In exploring the number of eagle sightings over time, it's important to consider the fact that the length of the observation periods vary from year to year, ranging from 134 to 248.75 hours. Update your plot from part b to also include information about the observation length in hours and comment on your findings.

## 8 BR 12.7

Consider a Poisson regression model of  $Y$  versus  $X_1$ : year and  $X_2$ : hours.

- (a) In the bald eagle analysis, why might a Poisson regression approach be more appropriate than a Normal regression approach?
- (b) Simulate the posterior of the Poisson regression model of  $Y$  versus  $X_1$  and  $X_2$ . Check the `prior_summary()`.
- (c) Use careful notation to write out the complete Bayesian structure of the Poisson regression model of  $Y$  versus  $X_1$  and  $X_2$ .
- (d) Complete a `pp_check()` for the Poisson model. Use this to explain whether the model is “good” and, if not, what assumptions it makes that are inappropriate for the bald eagle analysis.

## 9 BR 12.8

The Poisson regression model of bald eagle counts ( $Y$ ) by year ( $X_1$ ) and observation hours ( $X_2$ ), was pretty good. Let's see if a Negative Binomial approach is even better.

- (a) Simulate the model posterior and use a `pp_check()` to confirm whether the Negative Binomial model is reasonable.
- (b) Use careful notation to write out the complete Bayesian structure of the Negative Binomial regression model of  $Y$  versus  $X_1$  and  $X_2$ .
- (c) Interpret the posterior median estimates of the regression coefficients on **year** and **hours**. Do so on the unlogged scale.
- (d) Construct and interpret a 95% posterior credible interval for the year coefficient.
- (e) When controlling for the number of observation hours, do we have ample evidence that the rate of eagle sightings has increased over time?