| EEEB UN3005/GR5005 | Final Exam - Due 14 May 2019

Xun Zhao, xz2827

Final Exam Instructions: Complete this exam by writing code in the code chunks provided. If required, provide written explanations below the relevant code chunks. Replace "USE YOUR NAME HERE" with your name in the document header. When complete, knit this document within RStudio to generate a pdf. Please review the resulting pdf to ensure that all content relevant for grading (i.e., code, code output, and written explanations) appears in the document. Rename your pdf document according to the following format: EEEB3005_final_exam_firstname_lastname.pdf. Upload your final exam pdf document to CourseWorks by 5 pm on 14 May.

There are a few special instructions for this exam. First, **work alone on this exam**. You can, however, feel free to reference your own class notes, course materials posted on CourseWorks, and any other published resources (in print or online). Just please **don't consult with another person directly**. Second, note that partial credit will be considered for all exam problems. Please show your work where relevant even if you're not confident in every part of your solution, and make sure to try to address all parts of each problem. If you run into some significant issues with knitting your document because of code you can't completely figure out, comment out problematic code as needed. Finally, if you have clarifying questions about the exam, please don't hesitate to get in contact with me via email.

Good luck!

Problem 1 (10 points)

At the following link you'll find a research article by Kärvemo et al. in the journal *PLoS ONE* on factors affecting the prevalence of a fungal pathogen of pond-dwelling amphibians, *Batrachochytrium dendrobatidis* (*Bd*): https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0199852.

Specifically, the authors were interested in both local- (i.e., pond-level) and landscape-level environmental variables that might influence pathogen prevalence in six amphibian species in Sweden. They assessed the impact of these environmental variables by using a binomial multilevel model (i.e, a mixed effects model), with *Bd* infection as the outcome of interest. The specific pond-level characteristics analyzed were pH, perimeter (length of the pond shoreline), and amount of canopy cover shading the pond. In the Results section of the paper, the authors conclude: "One of the pond variables, pH, was positively associated with the detected *Bd* infection prevalence (Figs 2A and 3A), while Perimeter and Canopy had at most weak associations."

In a short paragraph (maximum of five sentences), use Fig. 2A as a point of reference to explain why the authors come to these specific conclusions. What aspects of the model results shown in Fig. 2A suggest a positive association between pH and *Bd* infection prevalence? Why do the authors suggest there is a positive association with pH and only "weak associations" with Perimeter and Canopy?

Answer:

$$\begin{aligned} \text{Infection} &\sim \text{Binomial}(n, p) \\ \text{logit}(p) &= \alpha + \beta \cdot x \end{aligned}$$

The β parameter shows the association between variables and infection, so the interval shown in Fig 2A is the interval of β parameter. The figure shows that the whole 95% Bayesian credible interval of β_{pH} is positive, so the author concludes that pH and infection are positively related. Besides, the mean value of $\beta_{\text{Perimeter}}$ and β_{Canopy} is positive, but part of the interval is negative. So, infection is still associated with perimeter and canopy, but the association is not so strong.

Data Description Interlude

All remaining data analysis problems for this exam will draw from the same livestock disease dataset concerning the mortality of large stock (Zebu cattle) and small stock (goats and fat-tailed sheep) due to a rinderpest epidemic across 11 different herds in East Africa. You can find this data on the class CourseWorks in the file `rinder.csv` or via the `rethinking` package using `data("Rinder")`. Small stock and large stock are pulled apart in the dataset, but are contained within the same herds in reality. Thus, you'll find 22 rows in the data (two for each herd) and four columns:

- `herd`: the identifying label for the herd
- `stock`: categorical value for large stock (cattle) or small stock (goats and sheep)
- `n`: number of animals before the rinderpest epidemic
- `mortality`: number of animals who died during the rinderpest epidemic

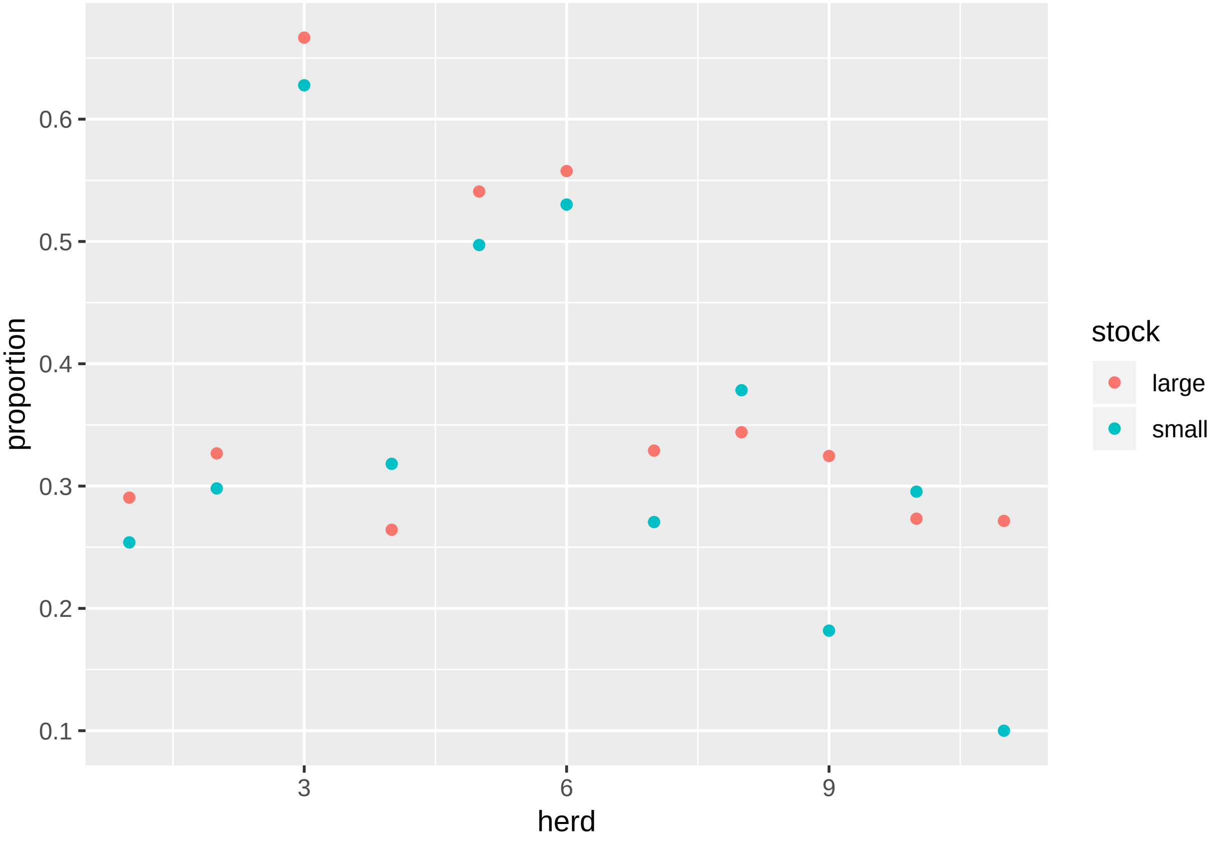
Consider the hypothesis that herders in East Africa keep small stock as insurance against risk, because goats and sheep survive better in the face of stressors like disease. But during good times, goats and sheep are a burden since they're less economically beneficial to keep as compared to cattle. To test this hypothesis, you'll be modeling rinderpest-driven mortality as a function of stock type.

Problem 2 (15 points)

First, import the rinderpest data into R using a method of your choice. Then, to better familiarize yourself with the data, generate a plot of the *proportion* of animals dying in an epidemic across herds, separated out by stock type. Conceptualize the plot this way: `herd` should appear on the x-axis and proportion of animals dying in an epidemic should appear on the y-axis. Represent the different `stock` types in the data using color. Since each herd contains some `stock` of each type, you should have two data points for each `herd` illustrated.

Using the plot to assist you, across most herds in the observed data, which stock type tend to suffer more mortality: small stock or large stock?

```
data(Rinder)
Rinder$proportion = Rinder$mortality / Rinder$n
graph = ggplot(Rinder, aes(x = herd, y = proportion, color = stock)) + geom_point()
plot(graph)
```



Answer:

In this graph, there are 8 herds where the mortality of large stock is greater than small stock, while there are only 3 herds that mortality of small stock is greater than large stock.

So from the graph, the small stock lives better.

Problem 3 (15 points)

Use `map()` to fit a binomial generalized linear model (with a logit link), modeling `mortality` as an outcome with `stock` size as a predictor variable.

Report the 97% PIs of the fit model parameters. How do your estimates reflect on the hypothesis that small stock survive epidemics better?

```
Rinder$dummy_stock = ifelse(Rinder$stock == 'small', 0, 1)
model.1 = map(
    alist(
        mortality ~ dbinom(n, p),
        logit(p) <- a + b * dummy_stock,
        a ~ dnorm(0, 1),
        b ~ dnorm(0, 1)
    ),
    data = Rinder
)
precis(model.1, prob = 0.97)
```

```
##    Mean StdDev  1.5% 98.5%
## a  0.03   0.04 -0.06  0.12
## b -0.12   0.05 -0.23 -0.02
```

Answer:

$$\begin{aligned} p_{\text{small}} &= \text{logistic}(a + b \cdot 0) = \text{logistic}(0.03) = 0.507 \\ p_{\text{large}} &= \text{logistic}(a + b \cdot 1) = \text{logistic}(-0.09) = 0.478 \end{aligned}$$

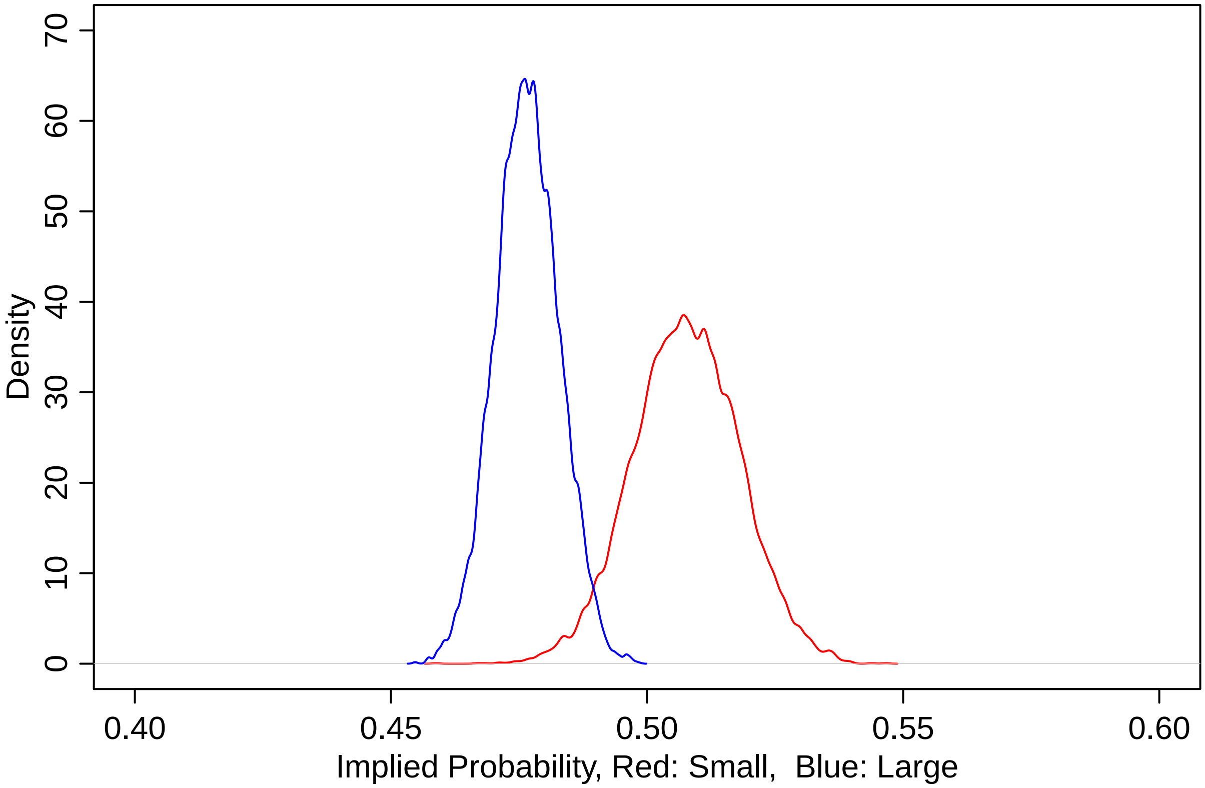
So, the probability of small stock to die is higher, which means small stock survive worse.

Problem 4 (15 points)

Plot the implied probability of mortality for small and large stock using the model you fit in Problem 3, making sure to incorporate full posterior uncertainty. You can approach this problem by using `dens()` to plot the implied probability of mortality values from the model posterior, showing values for small and large stock using different colors.

Finally, what is the expected difference in probability of mortality between small stock and large stock, according to your model?

```
samples.1 = extract.samples(model.1, n = 10000)
dens(logistic(samples.1$a), col = 'red',
     xlim = c(0.4, 0.6), ylim = c(0, 70),
     xlab = 'Implied Probability, Red: Small,  Blue: Large')
dens(logistic(samples.1$a + samples.1$b), add = TRUE, col = 'blue')
```



Problem 5 (15 points)

Now use `map2stan()` to fit a binomial multilevel model (i.e., generalized linear mixed model) to the rinderpest data. More specifically, this model should include `mortality` as an outcome, with `stock` size as a predictor variable and varying intercepts by `herd`.

Report the 97% HPDIs of *all* fit model parameters. Interpret the estimates, including the standard deviation among herds. How do your estimates reflect on the hypothesis that small stock survive epidemics better?

```
model.2 = map2stan(
    alist(
        mortality ~ dbinom(n, p),
        logit(p) <- a[herd] + b * dummy_stock,
        a[herd] ~ dnorm(0, 5),
        b ~ dnorm(0, 5)
    ),
    chains = 4,
    data = Rinder
)
```

```
## code for methods in class "Rcpp_stan_fit4model2e6f247d4246_mortality___dbinom_n__p_"
## code for methods in class "Rcpp_stan_fit4model2e6f247d4246_mortality___dbinom_n__p_"

## Warning: There were 1 divergent transitions after warmup. Increasing adapt_delta abov
## http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup

## Warning: Examine the pairs() plot to diagnose sampling problems

## Computing WAIC

## Constructing posterior predictions

## Aggregated binomial counts detected. Splitting to 0/1 outcome for WAIC calculation.
```

```
precis(model.2, prob = 0.97, depth = 2)
```

```
##        Mean StdDev lower 0.97 upper 0.97 n_eff Rhat
## a[1]  -1.05   0.11      -1.29      -0.80  3604    1
## a[2]  -0.88   0.10      -1.08      -0.65  3872    1
## a[3]   0.53   0.06       0.41       0.64  2517    1
## a[4]  -1.13   0.13      -1.40      -0.82  3639    1
## a[5]   0.00   0.06      -0.13       0.13  2843    1
## a[6]   0.09   0.06      -0.05       0.21  2683    1
## a[7]  -0.89   0.11      -1.12      -0.65  3336    1
## a[8]  -0.77   0.12      -1.02      -0.51  3583    1
## a[9]  -0.97   0.11      -1.20      -0.72  3672    1
## a[10] -1.11   0.13      -1.40      -0.83  3598    1
## a[11] -1.20   0.18      -1.60      -0.82  3913    1
## b      0.16   0.05       0.05       0.26  1509    1
```

Answer:

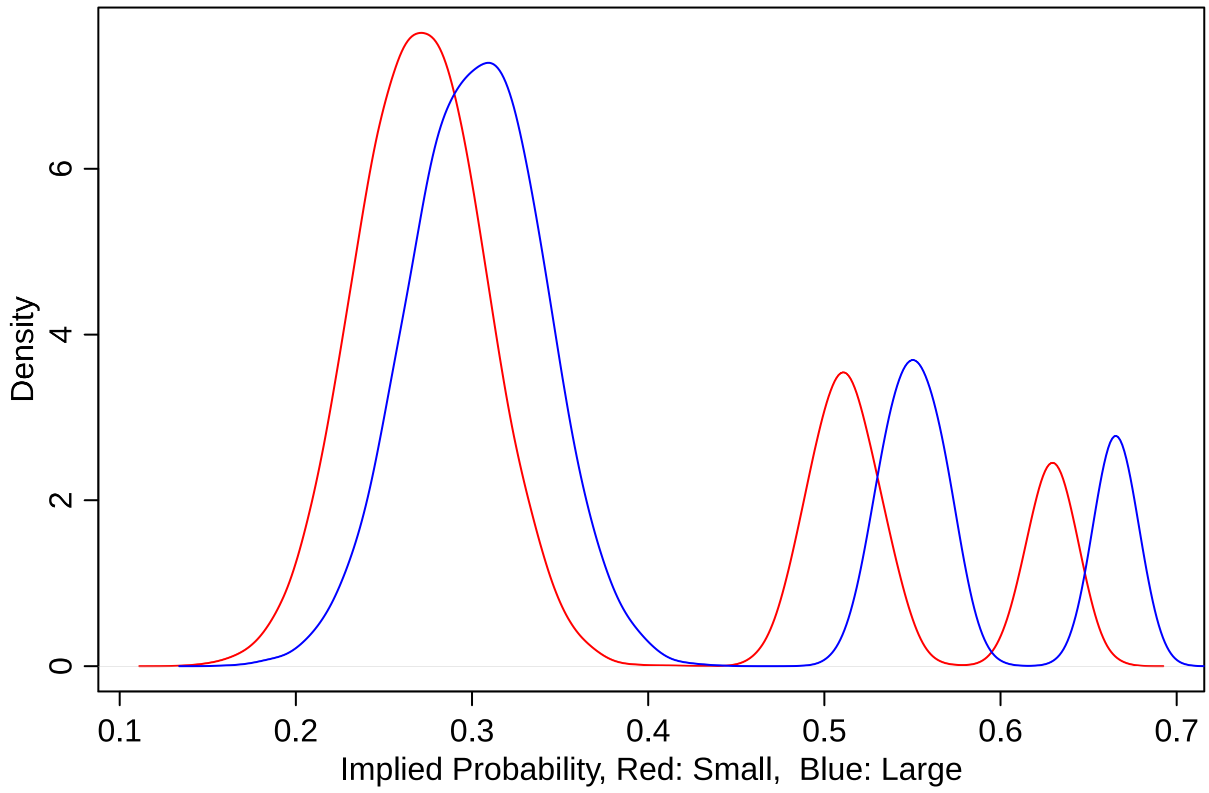
$$p_{\text{small}} = \text{logistic}(a + b \cdot 0)$$
$$p_{\text{large}} = \text{logistic}(a + b \cdot 1)$$

Since b is positive, so $p_{\text{small}} < p_{\text{large}}$ and the small stock dies with lower porbability, which means they survive better.

Problem 6 (10 points)

Plot the implied probability of mortality for small and large stock in an *average* herd using the model you fit in Problem 5, making sure to incorporate full posterior uncertainty.

```r
samples.2 = extract.samples(model.2, n = 4000)
dens(logistic(samples.2$a), col = 'red', xlab = 'Implied Probability, Red: Small,  Blue:
sampleb = rep(samples.2$b, dim(samples.2$a)[2])
dim(sampleb) = dim(samples.2$a)
dens(logistic(samples.2$a + sampleb), add = TRUE, col = 'blue')
```



Problem 7 (10 points)

How do the estimates and inferences of the model in Problem 5 differ from those produced by the model in Problem 3? Can you explain the differences? In answering this question, feel free to reference any aspects of the raw data and/or make any additional calculations or plots you want.

Answer:

`Model.1` in Problem 3 indicates than large stock lives better, but `Model.2` indicates that small stock lives better.

```r
summarize(group_by(Rinder, stock), prop = sum(mortality) / sum(n))
```

```
## # A tibble: 2 x 2
##   stock  prop
##   <fct> <dbl>
## 1 large 0.476
## 2 small 0.508
```

Here, we can see that when considering the sum of all herds, the mortality of small stock is larger.

This might be because that for small stocks, herds with higher mortality relatively have larger size, and herds with lower mortality realtively have smaller size. For example, herds 3, 5, 6, whose mortality are greater than 49%, have herd size larger than 500, a really big number compared with others. So these big numbers make the mean mortality higher than 50%.

Conversely, for large stocks, the size of different herds are similar to each other, which makes the average mortality slightly lower than small stock.

In Question.3 the model considers all herds as a whole group, so the death proportion of small stock is mainly decided by big size herds, whose mortality is higher. However, in Question.5, we use different intercepts for different herds, avoiding the influence between herds. So Question.5 shows the real pure model of herds, that is, the small stock lives better.

Problem 8 (10 points)

One reason for choosing a multilevel model, like the one you were just asked to fit, is because they help us balance the risk of generating statistical models that underfit or overfit the data. In a short paragraph (maximum of five sentences), explain in general terms what it means for a statistical model to underfit or overfit the data. Why do we want to avoid underfitting and why do we want to avoid overfitting? Finally, what is it about multilevel models (let's say a varying intercepts model if it helps you to think about specifics) that help us to navigate underfitting/overfitting?

Answer:

Underfitting means we fit a dataset without enough parameters, which causes the high bias between estimate and real data. Overfitting means that we fit a dataset with too many parameters, which causes high variance among estimate data itself. Underfitting cannot fit current dataset well due to high bias, and overfitting cannot generalize well when dealing with new data due to high variance. In the multilevel model we used above, we used more parameters in Question.5 than Question.3, helping the model to avoid underfitting, because more parameters provide more information about the difference among these herd (analyzing these herds respectively).