STATISTICAL RETHINKING WINTER 2019
HOMEWORK, WEEK 10

When/how is homework due? This assignment is due Friday March 8. I'll probably only check them on Monday, March 11, however. So turning it in on the weekend would be fine.

There are two problems below. The first focuses on measurement error. The second focuses on imputation.

1. Consider the relationship between brain volume (`brain`) and body mass (`body`) in the `data(Primates301)`. These values are presented as single values for each species. However, there is always a range of sizes in a species, and some of these measurements are taken from very small samples. So these values are measured with some unknown error.

We don't have the raw measurements to work with—that would be best. But we can imagine what might happen if we had them. Suppose error is proportional to the measurement. This makes sense, because larger animals have larger variation. As a consequence, the uncertainty is not uniform across the values and this could mean trouble.

Let's make up some standard errors for these measurements, to see what might happen. Load the data and scale the the measurements so the maximum is 1 in both cases:

```
library(rethinking)
data(Primates301)
d <- Primates301
cc <- complete.cases( d$brain , d$body )

B <- d$brain[cc]
M <- d$body[cc]
B <- B / max(B)
M <- M / max(M)
```

Now I'll make up some standard errors for B and M, assuming error is 10% of the measurement.

```
Bse <- B*0.1
Mse <- M*0.1
```

Let's model these variables with this relationship:

$$B_i \sim \text{Log-Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta \log M_i$$

This says that brain volume is a log-normal variable, and the mean on the log scale is given by μ. What this model implies is that the expected value of B is:

$$\mathrm{E}(B_i|M_i) = \exp(\alpha)M_i^{\beta}$$

So this is a standard allometric scaling relationship—incredibly common in biology.

Ignoring measurement error, the corresponding `ulam` model is:

```
dat_list <- list(
    B = B,
    M = M )

m1.1 <- ulam(
    alist(
        B ~ dlnorm( mu , sigma ),
        mu <- a + b*log(M),
        a ~ normal(0,1),
        b ~ normal(0,1),
        sigma ~ exponential(1)
    ) , data=dat_list )
```

Your job is to add the measurement errors to this model. Use the divorce/marriage example in the chapter as a guide. It might help to initialize the unobserved true values of B and M using the observed values, by adding a list like this to `ulam`:

```
start=list( M_true=dat_list$M , B_true=dat_list$B )
```

Compare the inference of the measurement error model to those of `m1.1` above. Has anything changed? Why or why not?

2. Now consider missing values—this data set is lousy with them. You can ignore measurement error in this problem. Let's get a quick idea of the missing values by counting them in each variable:

```
library(rethinking)
data(Primates301)
d <- Primates301
colSums( is.na(d) )
```

```
          name                genus              species           subspecies
             0                    0                    0                  267
       spp_id             genus_id      social_learning      research_effort
             0                    0                   98                  115
         brain                 body           group_size            gestation
           117                   63                  114                  161
      weaning            longevity         sex_maturity maternal_investment
           185                  181                  194                  197
```

We'll continue to focus on just brain and body, to stave off insanity. Consider only those species with measured body masses:

```
cc <- complete.cases( d$body )
M <- d$body[cc]
M <- M / max(M)
B <- d$brain[cc]
B <- B / max( B , na.rm=TRUE )
```

You should end up with 238 species and 56 missing brain values among them.

First, consider whether there is a pattern to the missing values. Does it look like missing values are associated with particular values of body mass? Draw a DAG that represents how missingness works in this case. Which type (MCAR, MAR, MNAR) is this?

Second, impute missing values for brain size. It might help to initialize the 56 imputed variables to a valid value:

```
start=list( B_impute=rep(0.5,56) )
```

This just helps the chain get started.

Compare the inferences to an analysis that drops all the missing values. Has anything changed? Why or why not? Hint: Consider the density of data in the ranges where there are missing values. You might want to plot the imputed brain sizes together with the observed values.