

# Many Variables (Part 2)

EES 4891-06/5891-01

Bayesian Statistical Methods

Jonathan Magnolia Gilligan

Class #7: Wednesday February 4, 2026

# Masked Relationships

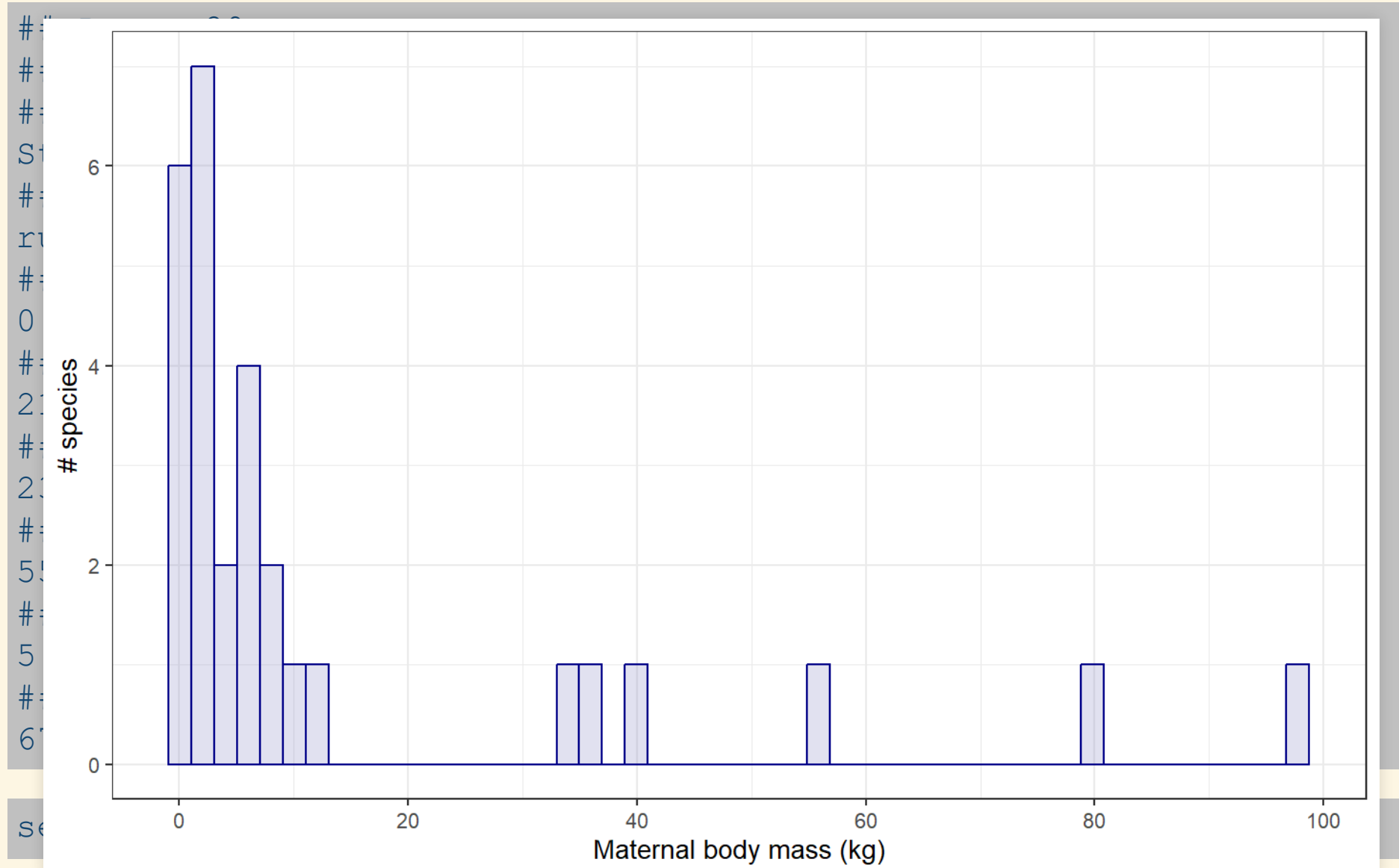
# Primate Milk Data

Primate Milk Data

# Masked Relationships

- Data from K. Hinde & L.A. Milligan, "Primate milk: Proximate mechanisms and ultimate perspectives," *Evolutionary Anthropology* 20, 9-23 (2011), doi:10.1002/evan.20289
- What we know have evolutionary forces shaped
  - Brains have different primate species
- Infants with large brains need more
  - 29 species, belonging to 4 clades.
- Hypothesis:
  - Species with large brains will produce higher mean maternal body mass (kg)
- Standardized data of total brain mass consisting
  - Note of neocortex tissue the log of mass.
- Clean data characteristics:
  - complete energy density (kcal) Remove rows with missing values of milk energy from fat, protein, and lactose

```
data(milk)
d <- milk
glucose(milk)
```



```
##                               mean      sd      5%      5%      0.1%      5%
d <- d %>% mutate(
  K = standardize(kcal.per.g),
  N = standardize(neocortex.perc),
  M = standardize(log(mass))
)
## neocortex.perc 67.5758824  5.9686117 58.4072 75.5872
##                               histogram
dcc <- drop_na(d, K, N, M)
## perc.fat
## perc.protein
## perc.lactose
## mass
## neocortex.perc
```

# Simple Regression

- Model

$$K \sim \text{Normal}(\mu, \sigma)$$

$$\mu = \alpha + \beta_N N$$

$$\alpha \sim \text{Normal}(0, 1)$$

$$\beta_N \sim \text{Normal}(0, 1)$$

$$\sigma \sim \text{Exponential}(1)$$

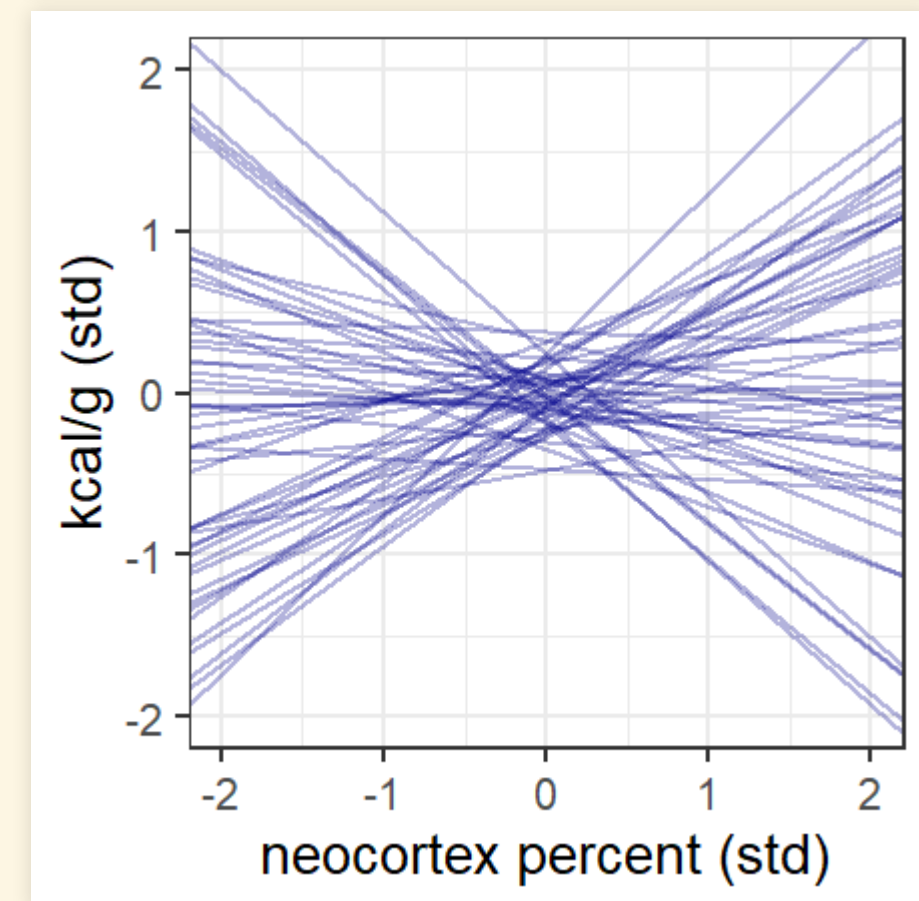
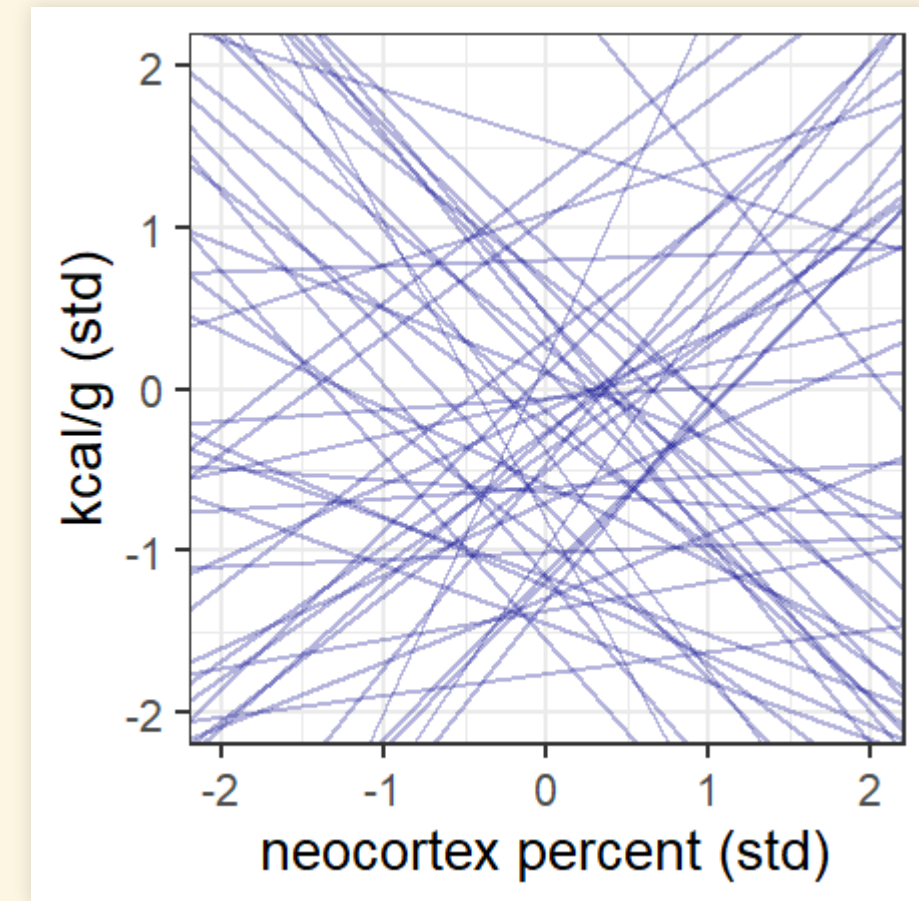
- Plot prior predictive distributions for variables.

- Prior predictions look absurd.

- Choose better priors:

$$\alpha \sim \text{Normal}(0, 0.2)$$

$$\beta_N \sim \text{Normal}(0, 0.5)$$



# Test Model with Simulated Data

- Can our model estimate parameters accurately?
- Test it with simulated data:
  1. Pick parameters  $\alpha$ ,  $\beta_N$ ,  $\sigma$  at random from the priors
  2. Generate simulated data with those parameters
  3. Use the model to estimate  $\alpha$ ,  $\beta_N$ , and  $\sigma$  from the simulated data.
  4. Compare the posterior distributions of  $\alpha$ ,  $\beta_N$ , and  $\sigma$  from the model to the actual values sampled in step 1.
- All three parameters lie within the 89% highest-density interval of the posterior
  - Success!

```
alpha  <- rnorm(1, 0, 0.2)
beta_N <- rnorm(1, 0, 0.5)
sigma  <- rexp(1, 1)
```

```
## Print the parameters
print(c(alpha = alpha, beta_N = beta_N, sigma = sigma))
```

```
##   alpha  beta_N  sigma
## -0.2168 -0.2577  0.0496
```

```
d_sim <- dcc |> select(clade, species, N) |>
  mutate(K = rnorm(n(), alpha + beta_N * N, sigma))
```

```
mdl <- quap(
  alist(
    K ~ dnorm(mu , sigma) ,
    mu <- a + bN * N ,
    a ~ dnorm(0 , 0.2) ,
    bN ~ dnorm(0 , 0.5) ,
    sigma ~ dexp(1)
  ), data=d_sim )
```

```
precis(mdl)
```

```
##           mean      sd   5.5%   94.5%
## a       -0.217 0.0125 -0.237 -0.197
## bN      -0.270 0.0129 -0.291 -0.249
## sigma   0.052 0.0088  0.038  0.066
```

# Examine Model

- Now, apply model to estimate  $\alpha$ ,  $\beta_N$ , and  $\sigma$  for the actual data.

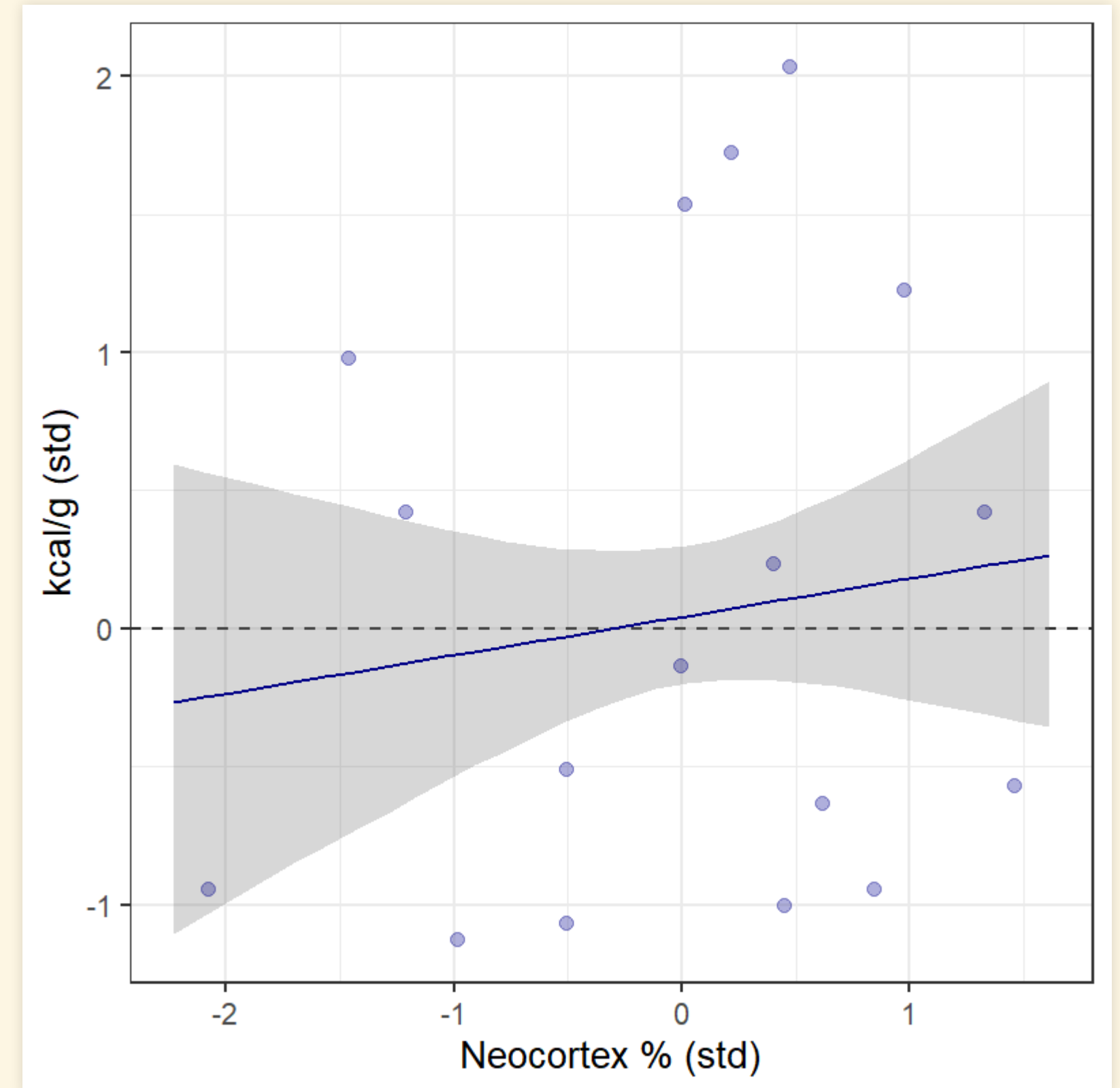
```
mdl_milk_1 <- quap(  
  alist(  
    K ~ dnorm(mu , sigma) ,  
    mu <- a + bN * N ,  
    a ~ dnorm(0 , 0.2) ,  
    bN ~ dnorm(0 , 0.5) ,  
    sigma ~ dexp(1)  
  ) , data=dcc )
```

- Examine the posterior estimate from the model:

```
precis_show(precis(mdl_milk_1))
```

##	mean	sd	5.5%	94.5%
## a	0.04	0.15	-0.21	0.29
## bN	0.13	0.22	-0.22	0.49
## sigma	1.00	0.16	0.74	1.26

- Both  $a$  and  $bN$  are consistent with zero. There isn't a strong relationship between  $N$  and  $K$ .



# Try A Different Model

- Model

$$K \sim \text{Normal}(\mu, \sigma)$$

$$\mu = \alpha + \beta_M M$$

$$\alpha \sim \text{Normal}(0, 1)$$

$$\beta_M \sim \text{Normal}(0, 1)$$

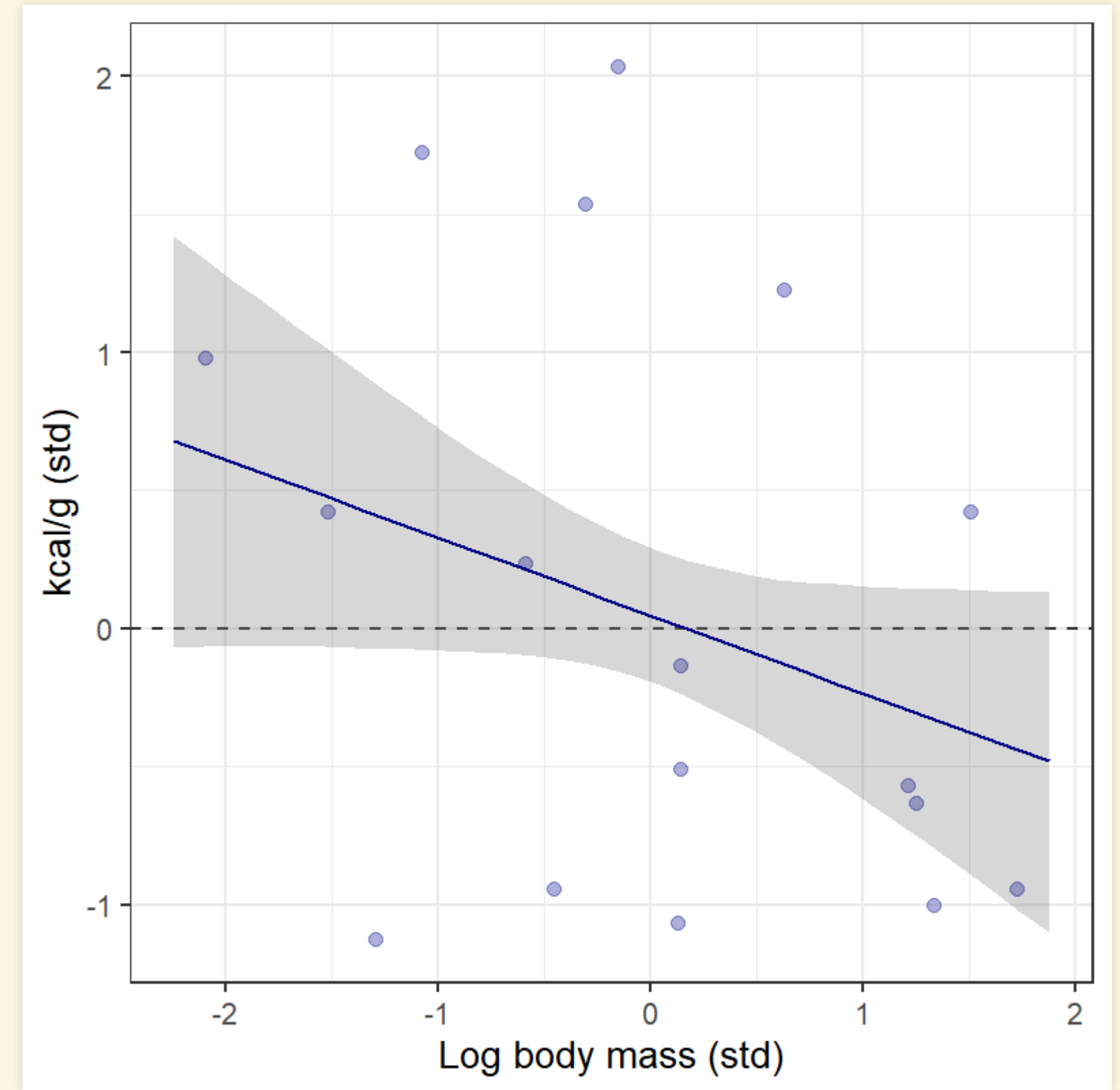
$$\sigma \sim \text{Exponential}(1)$$

- Examine the posterior:

```
precis_show(precis(mdl_milk_2))
```

##	mean	sd	5.5%	94.5%
## a	0.05	0.15	-0.20	0.29
## bM	-0.28	0.19	-0.59	0.03
## sigma	0.95	0.16	0.70	1.20

- Again,  $a$  and  $bM$  are consistent with zero.





# Consider Both $M$ and $N$

- Model

$$K \sim \text{Normal}(\mu, \sigma)$$

$$\mu = \alpha + \beta_M M + \beta_N N$$

$$\alpha \sim \text{Normal}(0, 1)$$

$$\beta_M \sim \text{Normal}(0, 1)$$

$$\beta_N \sim \text{Normal}(0, 1)$$

$$\sigma \sim \text{Exponential}(1)$$

```
mdl_milk_3 <- quap(  
  alist(  
    K ~ dnorm(mu , sigma) ,  
    mu <- a + bM * M + bN * N,  
    a ~ dnorm(0 , 0.2) ,  
    bM ~ dnorm(0 , 0.5) ,  
    bN ~ dnorm(0 , 0.5) ,  
    sigma ~ dexp(1)  
  ) , data=dcc )
```

# Compare Models using Counterfactuals

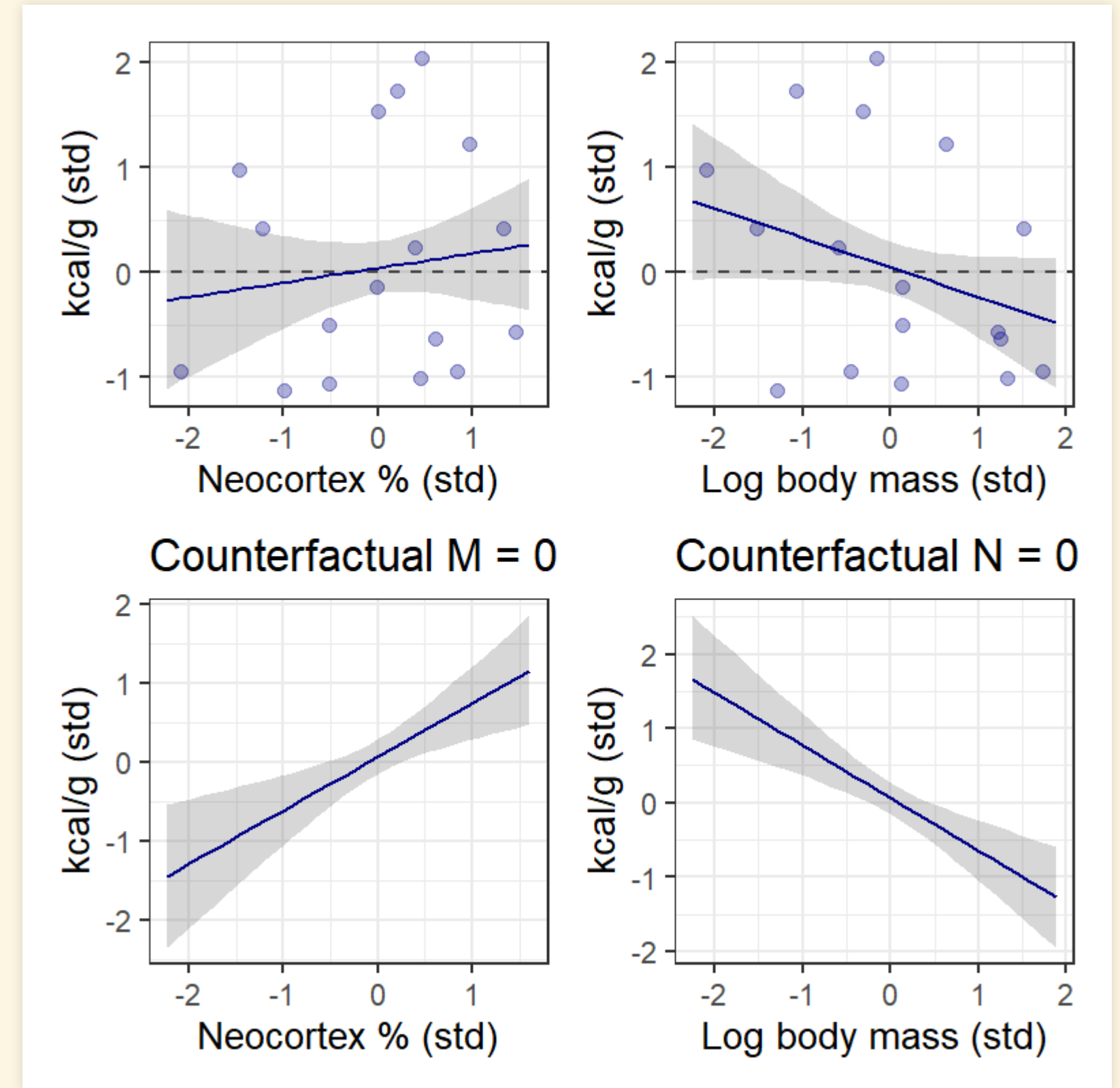
- Examine the posterior:

```
precis_show(precis(mdl_milk_3))
```

##		mean	sd	5.5%	94.5%
##	a	0.07	0.13	-0.15	0.28
##	bM	-0.70	0.22	-1.06	-0.35
##	bN	0.68	0.25	0.28	1.07
##	sigma	0.74	0.13	0.53	0.95

- $M$  and  $N$  have opposite effects, so they cancel out.

- Masking



# Multiple Regression Model

- Model

$$K \sim \text{Normal}(\mu, \sigma)$$

$$\mu = \alpha + \beta_M M + \beta_N N$$

$$\alpha \sim \text{Normal}(0, 1)$$

$$\beta_M \sim \text{Normal}(0, 1)$$

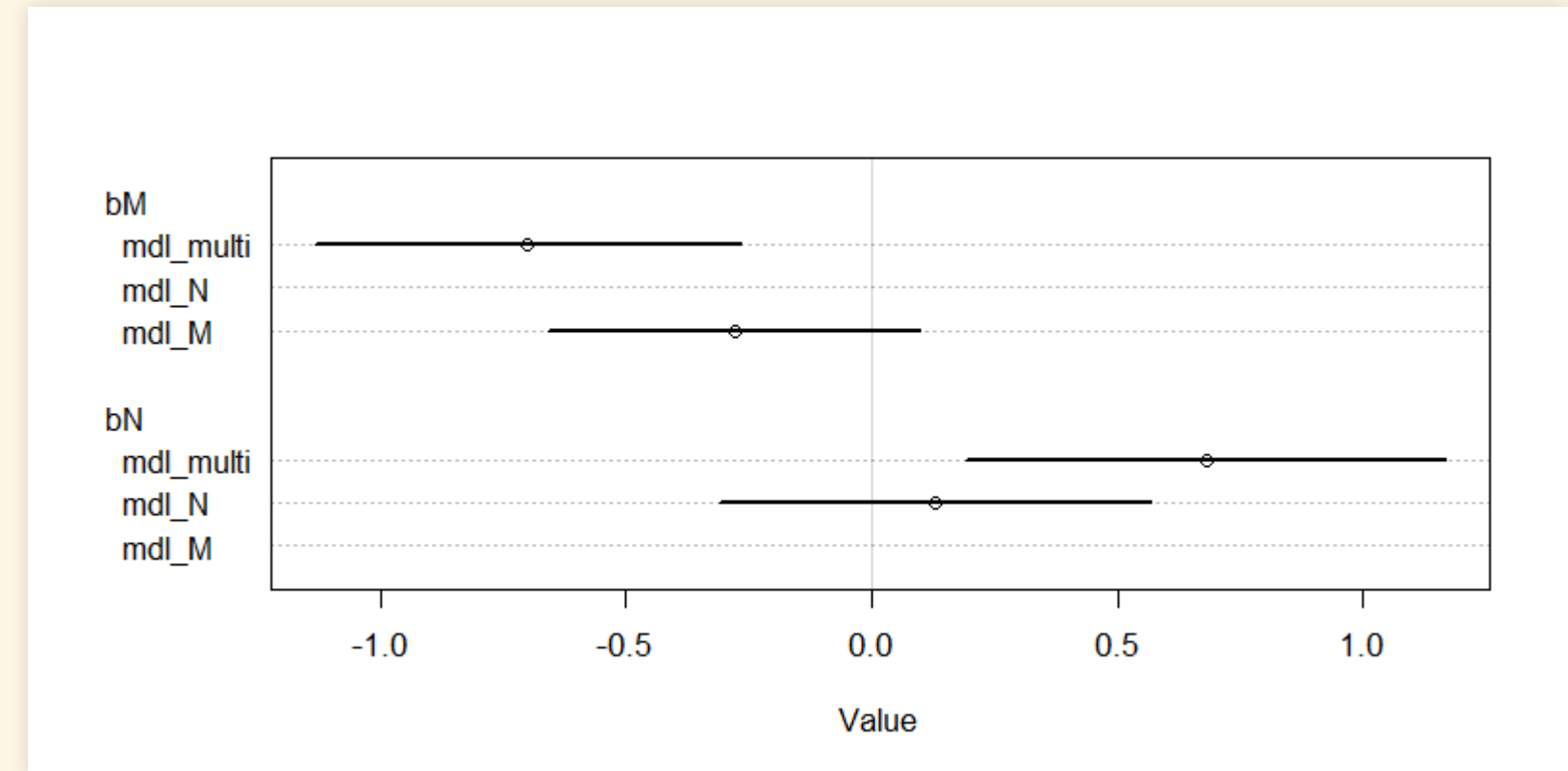
$$\beta_N \sim \text{Normal}(0, 1)$$

$$\sigma \sim \text{Exponential}(1)$$

```
precis_show(precis(mdl_milk_3, digits = 2))
```

```
##      mean   sd  5.5% 94.5%
## a      0.07 0.13 -0.15  0.28
## bM     -0.70 0.22 -1.06 -0.35
## bN      0.68 0.25  0.28  1.07
## sigma  0.74 0.13  0.53  0.95
```

```
coefplot(coefplot(mdl_M, mdl_N, mdl_multi),
         pars = c("bM", "bN"))
```



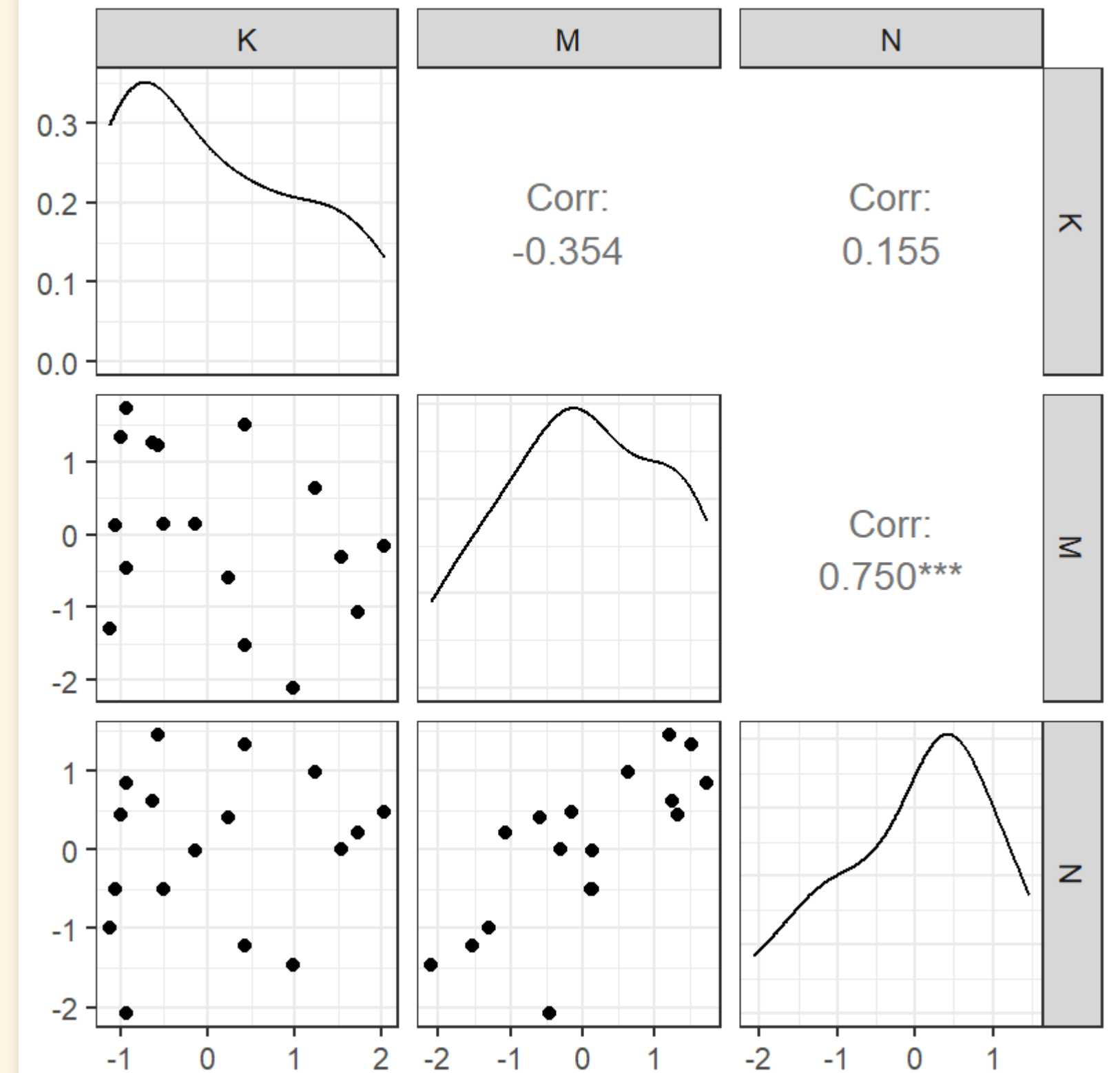
- This is the opposite of what we saw for divorce rates.
- The parameters for each predictor are consistent with zero for the single-predictor models
- When we include both predictors, the association with each is stronger.

# Interpreting Result

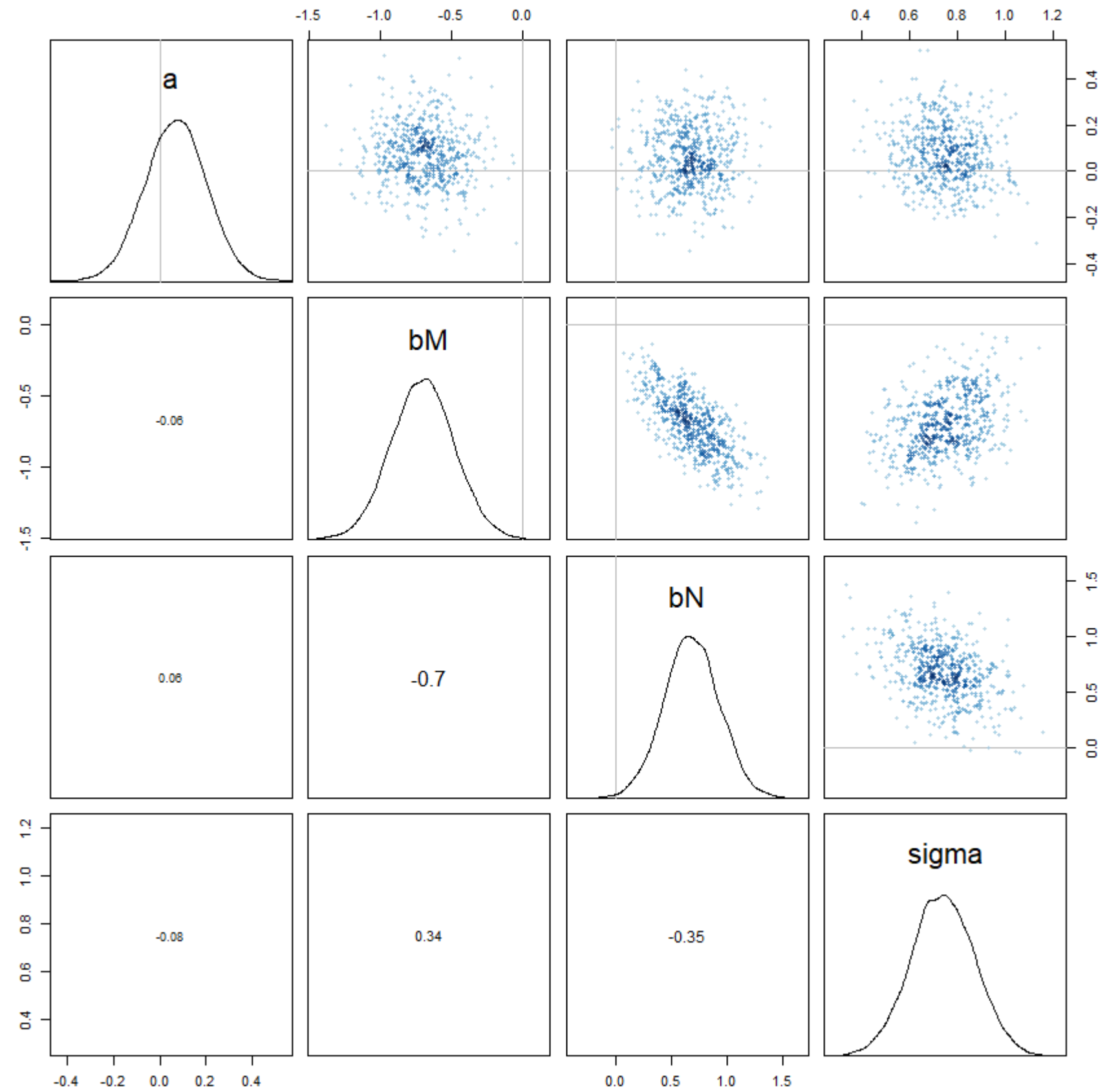
- No relationship between  $K$  and either  $M$  or  $N$ , if we ignore the relationship between  $M$  and  $N$
- Pairs plot shows relationships among  $K$ ,  $M$ , and  $N$ 
  - $M$  and  $N$  are strongly correlated
- Possible interpretations:
  - Species with high neocortex percent, relative to their body mass, have higher milk energy
  - Species with high body mass, relative to their neocortex percent, have higher milk energy

```
library(GGally)
```

```
ggpairs(dcc, columns = c("K", "M", "N"))
```

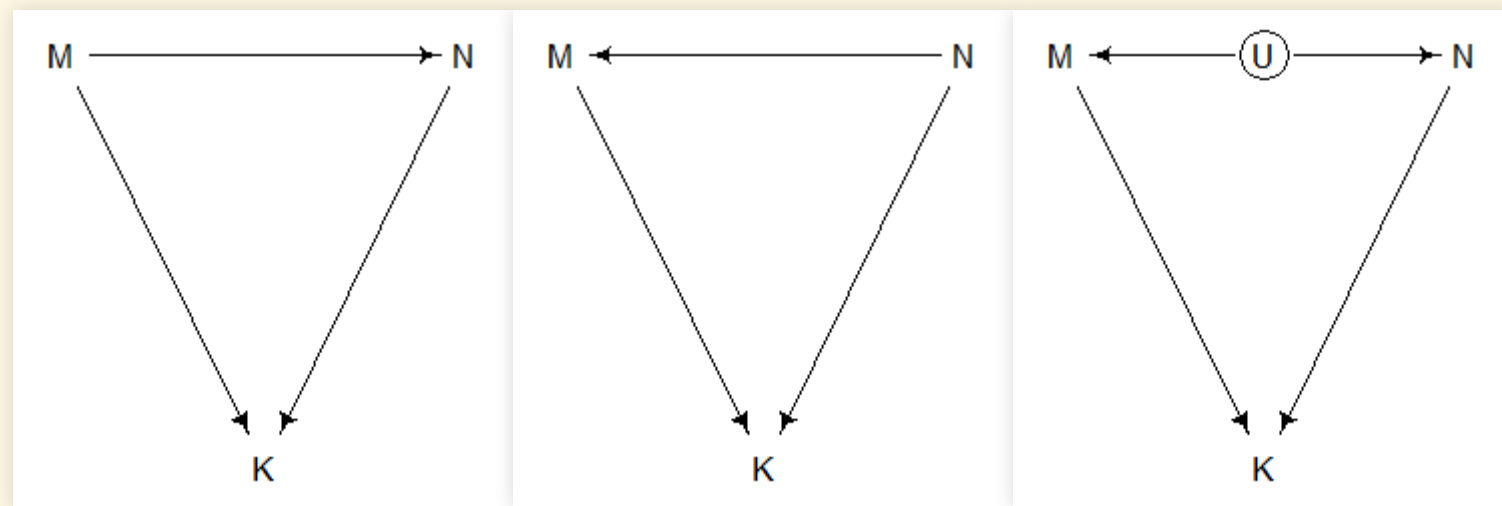


# Examining the Posterior



# Causal Possibilities

- Model results:
  - Bigger species (e.g., apes) tend to have lower-energy milk
  - Species with greater fraction of neocortex tend to have higher-calorie milk
  - But there's a relationship between body mass and neocortex percent
- There are 3 possible DAGs



1. Larger body mass causes greater neocortex percent
  2. Greater neocortex percent causes great body mass
  3. M and N are both determined by a third (latent) variable U that we didn't observe
    - More on latent variables in Ch. 6.
- Figuring out the right diagram is **hard**.
    - All three have the same *conditional independencies*.
    - Data alone won't solve this.
      - Our scientific knowledge can rule out absurd possibilities.

# Categorical Variables

# Categorical Variables

- Categories:
  - Discrete variables, describing a group that an individual falls into
  - Unordered:
    - Species: turtles, lizards, crocodiles, ...
    - Sex: male, female
    - Rock: granite, diorite, basalt, ...
  - Ordered:
    - Developmental status: infant, juvenile, adult
    - Geologic period: Permian, Triassic, Jurassic, Cretaceous, ...
    - Educational attainment: less than high-school, high school grad, some college, college grad, postgrad degree



# Milk Data

```
glimpse(d)
```

```
## Rows: 29
## Columns: 11
## $ clade      <fct> Strepsirrhine, Strepsirrhine, Strepsirrh...
## $ species    <fct> Eulemur fulvus, E macaco, E mongoz, E ru...
## $ kcal.per.g <dbl> 0.49, 0.51, 0.46, 0.48, 0.60, 0.47, 0.56...
## $ perc.fat   <dbl> 16.60, 19.27, 14.11, 14.91, 27.28, 21.22...
## $ perc.protein <dbl> 15.42, 16.91, 16.85, 13.18, 19.50, 23.58...
## $ perc.lactose <dbl> 67.98, 63.82, 69.04, 71.91, 53.22, 55.20...
## $ mass       <dbl> 1.95, 2.09, 2.51, 1.62, 2.19, 5.25, 5.37...
## $ neocortex.perc <dbl> 55.16, NA, NA, NA, NA, 64.54, 64.54, 67....
## $ K          <dbl> -0.9400408, -0.8161263, -1.1259125, -1.0...
## $ N          <dbl> -2.08019603, NA, NA, NA, NA, -0.50864129...
## $ M          <dbl> -0.4558357, -0.4150024, -0.3071581, -0.5...
```

```
table(d$clade)
```

```
##
##           Ape New World Monkey Old World Monkey
##           9                   9                   6
## Strepsirrhine
##           5
```

- Consider how the average milk energy varies by clade.

- Model:

$$K \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{\text{Clade}[i]}$$

$$\alpha_j \sim \text{Normal}(0, 0.5) \text{ for } j = 1 \dots 4$$

$$\sigma \sim \text{Exponential}(1)$$

```
d <- d |> mutate(clade_id = as.integer(clade))
```

```
mdl_clade <- quap(
  alist(
    K ~ dnorm(mu, sigma),
    mu <- a[clade_id],
    a[clade_id] ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ), data = d
)
```

# Results

- Model:

$$K \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{\text{Clade}[i]}$$

$$\alpha_j \sim \text{Normal}(0, 0.5) \text{ for } j = 1 \dots 4$$

$$\sigma \sim \text{Exponential}(1)$$

```
d <- d |> mutate(clade_id = as.integer(clade))

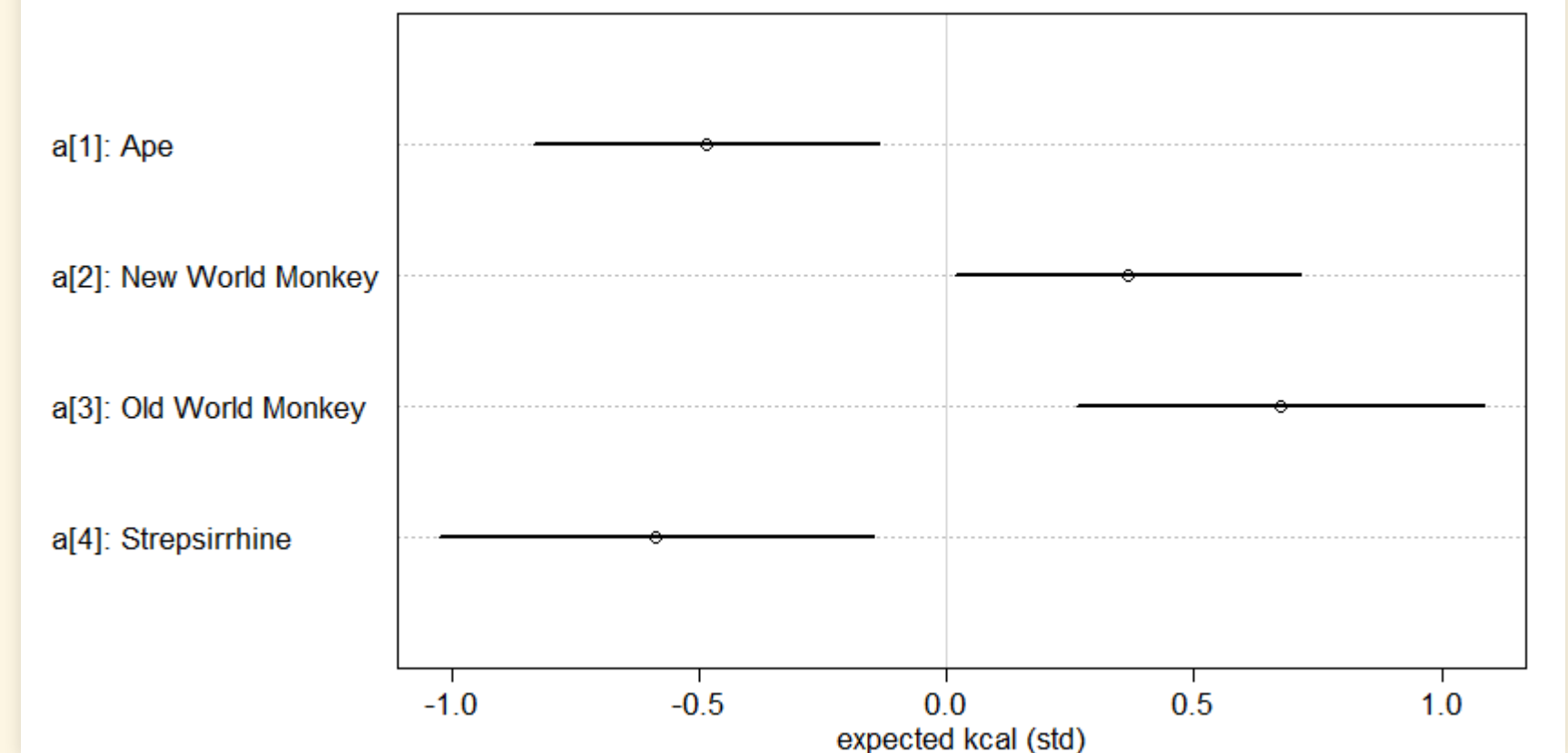
mdl_clade <- quap(
  alist(
    K ~ dnorm(mu, sigma),
    mu <- a[clade_id],
    a[clade_id] ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ), data = d
)
```

```
precis(mdl_clade, depth = 2)
```

```
##           mean      sd   5.5% 94.5%
## a[1]    -0.48 0.218 -0.832 -0.14
## a[2]     0.37 0.217  0.019  0.71
## a[3]     0.68 0.258  0.264  1.09
## a[4]    -0.59 0.275 -1.025 -0.15
## sigma   0.72 0.097  0.565  0.87
```

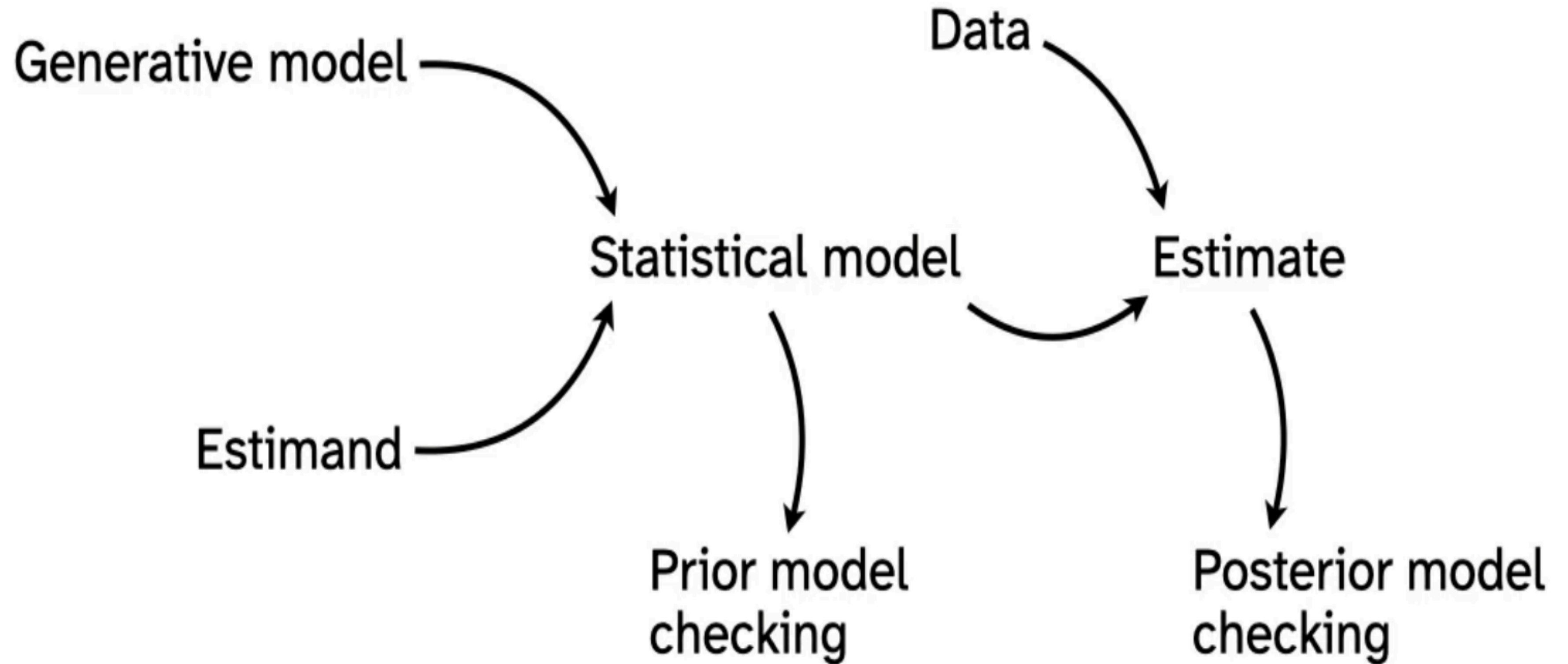
```
labels <- str_c("a[", 1:4, "]: ", levels(d$clade))

plot(precis(mdl_clade, depth = 2, pars = "a"),
     labels = labels, xlab = "expected kcal (std)")
```



# Review

# Bayesian Workflow



# Generative Models

- Consider the model

$$K \sim \text{Normal}(\mu, \sigma)$$

$$\mu = \alpha + \beta_M M + \beta_N N$$

$$\alpha \sim \text{Normal}(0, 1)$$

$$\beta_M \sim \text{Normal}(0, 1)$$

$$\beta_N \sim \text{Normal}(0, 1)$$

$$\sigma \sim \text{Exponential}(1)$$

- The generative part is in the first two lines:

$$K \sim \text{Normal}(\mu, \sigma)$$

$$\mu = \alpha + \beta_M M + \beta_N N$$

- Our observations of  $K$  are drawn from a random distribution with mean  $\mu$  and standard deviation  $\sigma$ .

- ■  $\mu$  is a linear function of  $M$  and  $N$ .
- There are four unknown parameters that describe the details of this generative process:  $\alpha$ ,  $\beta_M$ ,  $\beta_N$ , and  $\sigma$ .
- Where does the randomness come from in the first line of the model?
  - *Why* are our observations of  $K$  randomly distributed?
- 1. Sampling: milk properties vary from individual to individual and may vary over time.
  - What we measure depends on which individual we choose and when we measure its milk
- 2. Our laboratory assay has uncertainty.
- 3. We may make errors in collecting and storing the milk or in performing the assay

# Estimands

- Generative Model

$$K \sim \text{Normal}(\mu, \sigma)$$

$$\mu = \alpha + \beta_M M + \beta_N N$$

- We want to *estimate* the unknown parameters  $\alpha$ ,  $\beta_M$ ,  $\beta_N$ , and  $\sigma$ .
  - These are our *estimands*.
    - We use these *estimands* to answer research questions, such as whether either  $\beta_M$  or  $\beta_N$  is nonzero.

# Statistical Model

- We start with our *generative model*

$$K \sim \text{Normal}(\mu, \sigma)$$

$$\mu = \alpha + \beta_M M + \beta_N N$$

and our *estimands*  $\alpha$ ,  $\beta_M$ ,  $\beta_N$ , and  $\sigma$ .

- We want to use a *statistical model* to *estimate* our *estimands*.

- In Bayesian statistics, our *statistical model* combines:

1. Our *generative model*,

$$K \sim \text{Normal}(\mu, \sigma)$$

$$\mu = \alpha + \beta_M M + \beta_N N$$

2. *Prior* estimates for our *estimands*,

$$\alpha \sim \text{Normal}(0, 1)$$

$$\beta_M \sim \text{Normal}(0, 1)$$

$$\beta_N \sim \text{Normal}(0, 1)$$

$$\sigma \sim \text{Exponential}(1)$$

3. and *data*

to generate a *posterior estimate* of the *estimands*.

# Choosing Priors

- Choosing priors can be tricky
- Sometimes you know your priors from previous research.
  - You already have estimates of your *estimands* from your literature review, and you're trying to improve the accuracy of those estimates by collecting more data.
- Other times, theory can guide you.
  - Mass cannot be negative
  - The slope of a sand pile can't exceed a critical value or it will collapse
- **Uninformative Priors:** You know almost nothing about the *estimand*
  - Globe tossing: You don't know anything about the fraction of earth covered by water.
    - **Uniform prior:** all values from 0 to 1 are equally likely.
  - The posterior is almost entirely determined by the data
- **Weakly-informative prior:** You know something about the *estimand* but have a lot of uncertainty
  - The intercept for predicting height from weight is somewhere around 178 cm, but it could be anywhere from 138 to 218.
  - The posterior is a balance between the prior and the data
- **Strongly-informative prior:** You are very confident about the *estimand*, and can confidently rule out many possibilities, but you still want to improve the precision of your estimate.
  - People have been measuring the speed of light for more than 100 years, but you want to make it even more accurate.
  - The posterior is mostly determined by the prior, and new data only changes it a little.



# Prior Predictive Tests

- You often know more than you think.
  - Certain values of the *estimand* are just not believable.
    - Your priors should rule these out
- Prior predictive checks can help you find *weakly-informative* priors that rule out absurd values, without unduly constraining your analysis.
- *Strongly-informative* priors can be a problem if they are overconfident.
  - They can prevent your data from contributing to an improved *posterior* estimate.
- In most cases, *weakly-informative* priors are the best choice, and using prior predictive checks can help guide you to a sweet spot between too informative and not informative enough

# Applying Statistical Models

- After you have:
  1. Developed your *generative model*
  2. Chosen your *estimands*
  3. Chosen your *priors*
- It's time to apply your statistical model to your data to create a *posterior probability distribution* for your *estimand*
- After you apply your statistical model, you perform various *posterior* tests of the model to help determine how well you trust the results of your analysis.

# Bayesian Analysis and Scientific Method

- Science proceeds iteratively:
  - Each experiment or observation adds to the knowledge we already have.
- Bayesian statistical methods embody this
  - Previous knowledge determines your *priors*
  - New data from experiments or observations lets you create a *posterior* estimate that improves your knowledge of the *estimand* from what you knew before.
  - When you get new data, the old *posterior* becomes the *prior* for your next analysis

# Globe Tossing

# Sampling

- You have a globe and want to figure out what fraction of the earth's surface is water.
- Toss the globe in the air, catch it, and note whether your index finger is on water or land: outcomes are  $W$  and  $L$ .
- At every toss, use Bayes's theorem to update your estimate of the fraction that is water.



# Iteratively Improving Estimates of Water Coverage

