Many Variables (Part 2)

EES 4891-06/5891-01

Bayesian Statistical Methods

Jonathan Magnolia Gilligan

Class #7: Wednesday February 4, 2026

Masked Relationships

Masked Relationships

- Hypothesis:
  - Primates with larger brains produce higher-calorie milk so infant brains grow faster.
- Data:
  - Load data on characteristics of milk in different primate species.
    - `kcal.per.g`: Kilocalories energy per kg milk
    - `mass`: female body mass (kg)
    - `neocortex.perc`: percent of total brain-mass that is neocortex
- Standardize data
  - Note: we standardize the log of mass.
- Clean data
  - *complete-case analysis:* Remove rows with missing values

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

```
## Rows: 29
## Columns: 8
## $ 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.…
```

```
d <- d %>% mutate(
  K = standardize(kcal.per.g),
  N = standardize(neocortex.perc),
  M = standardize(log(mass))
)
```

```
dcc <- drop_na(d, K, N, M)
```

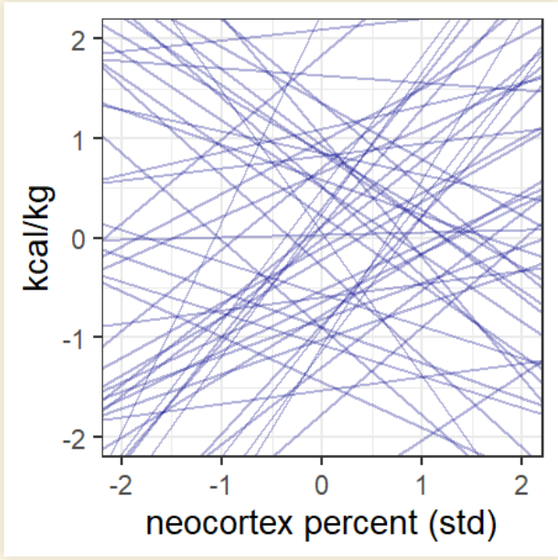
Simple Regression

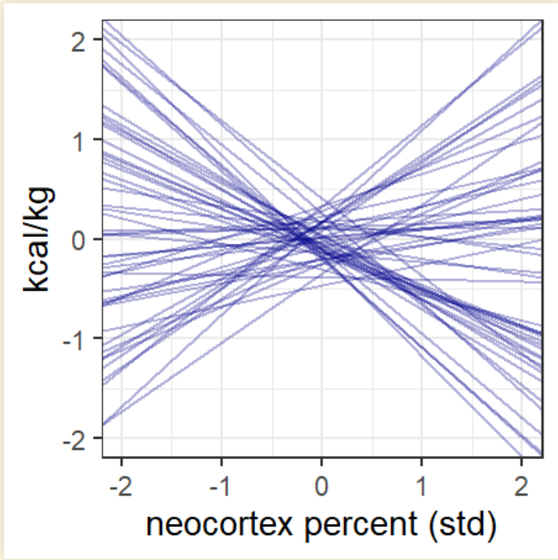
- Model

$$
\begin{aligned}
K &\sim \text{Normal}(\mu, \sigma) \\
\mu &= \alpha + \beta_N N \\
\alpha &\sim \text{Normal}(0, 1) \\
\beta_N &\sim \text{Normal}(0, 1) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

- Plot prior predictive distributions for variables.
  - Prior predictions look absurd.
- Choose better priors:

$$
\begin{aligned}
\alpha &\sim \text{Normal}(0, 0.2) \\
\beta_N &\sim \text{Normal}(0, 0.5)
\end{aligned}
$$



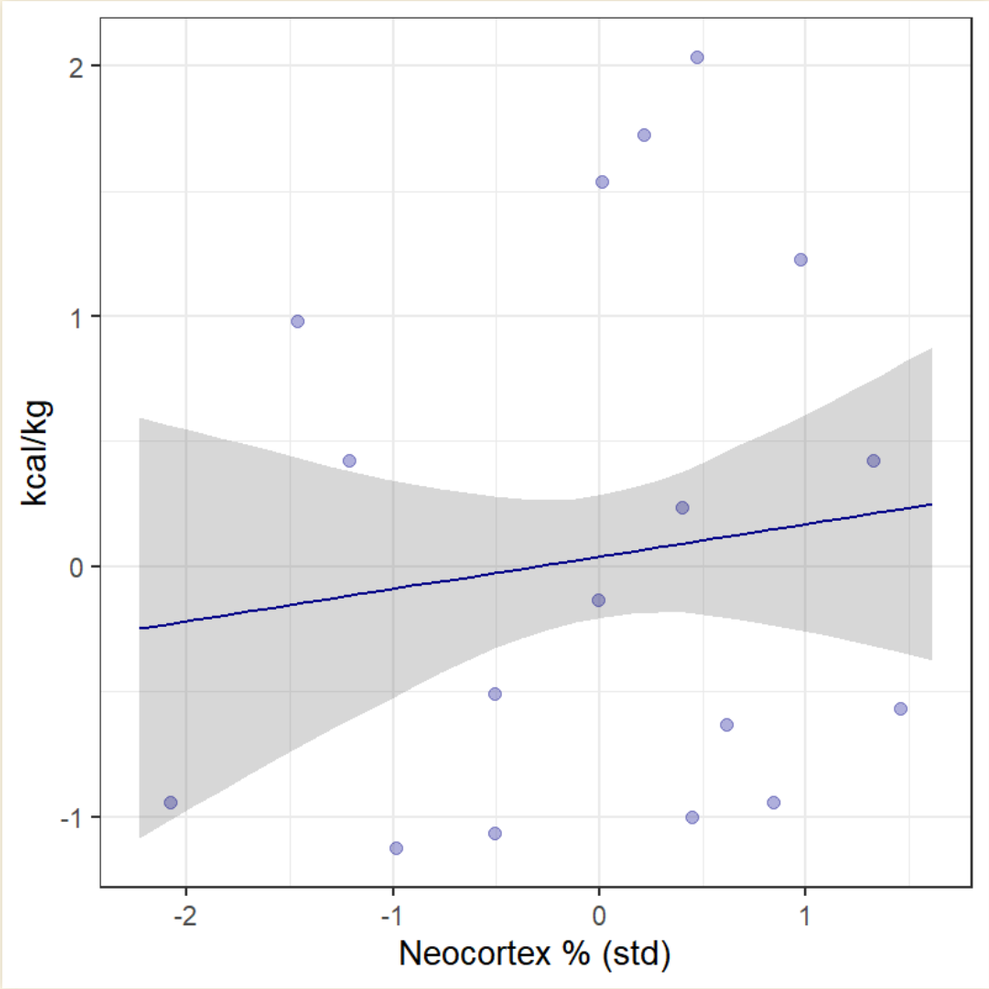


Examine Model

```
precis_show(precis(mdl_milk_1a, digits = 2))
```

```
##       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*.



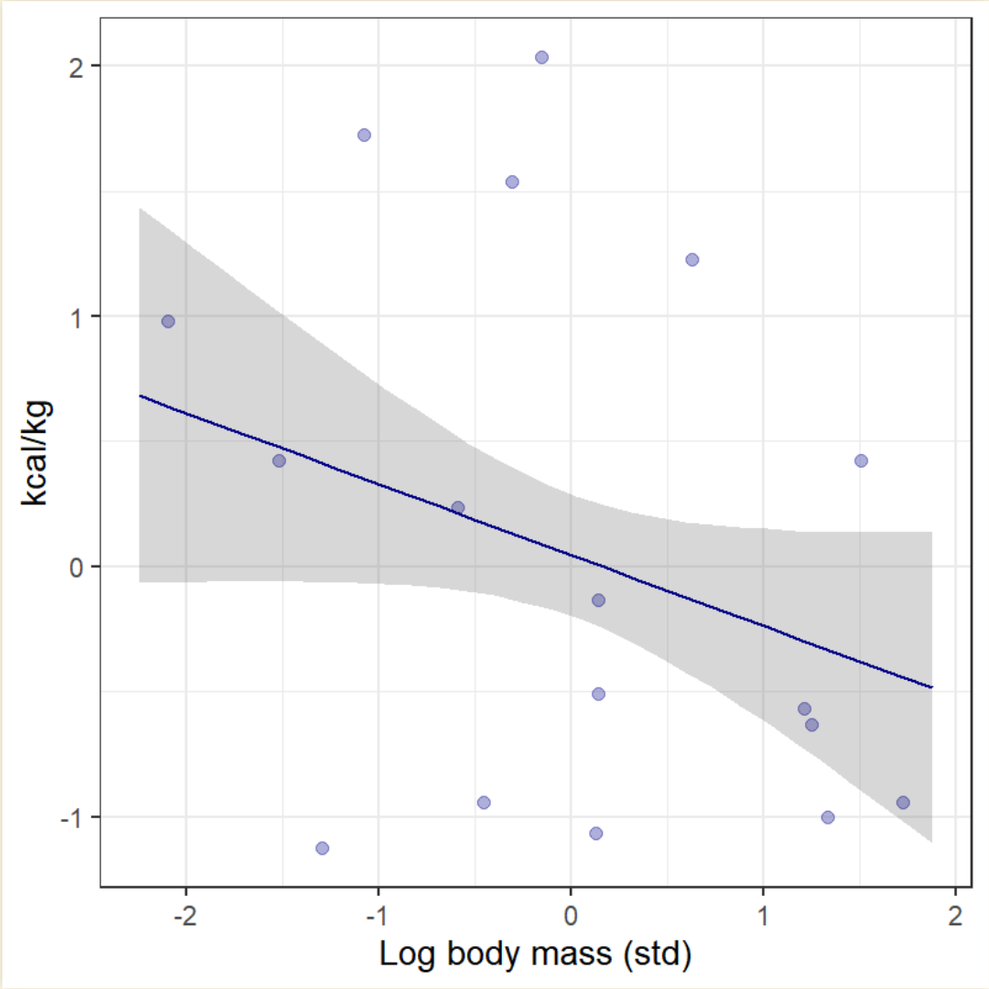
Try A Different Model

- Model

$$
\begin{aligned}
K &\sim \text{Normal}(\mu, \sigma) \\
\mu &= \alpha + \beta_M M \\
\alpha &\sim \text{Normal}(0, 1) \\
\beta_M &\sim \text{Normal}(0, 1) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

```
precis_show(precis(mdl_milk_2, digits = 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
```

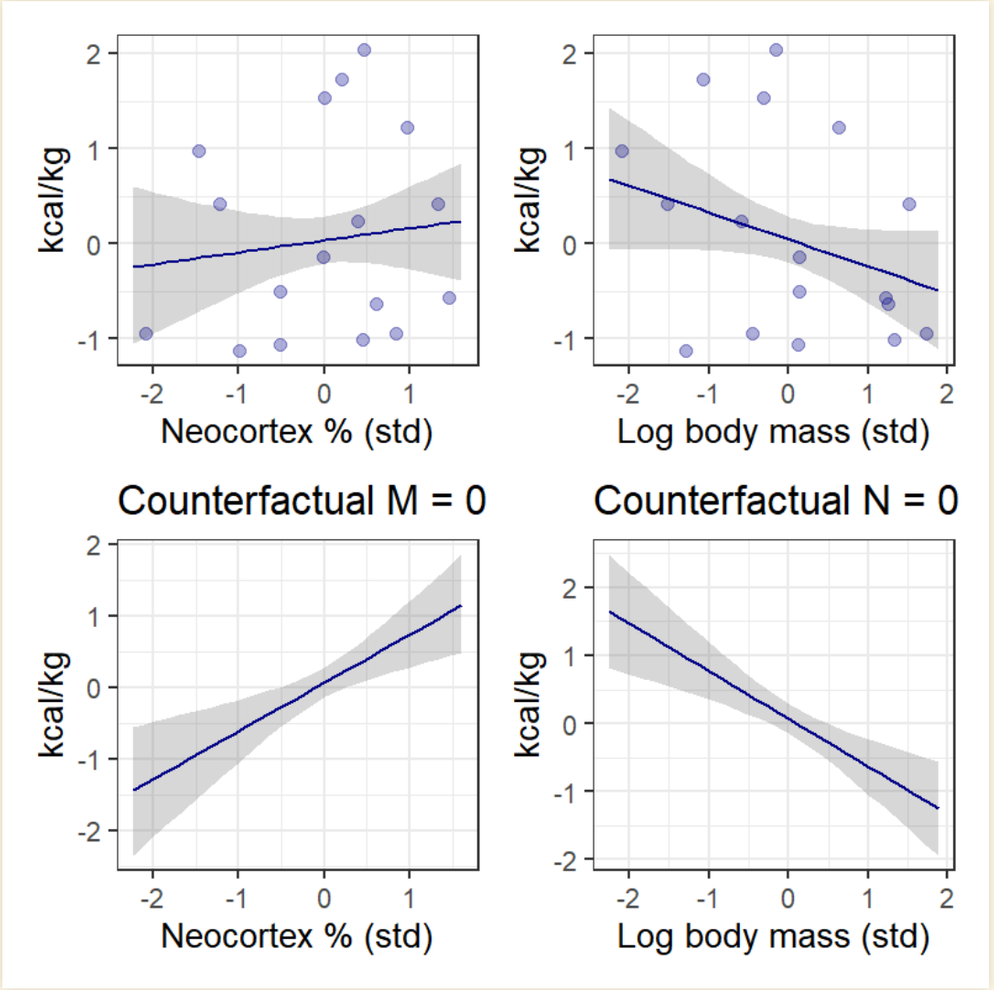


Compare Models using Counterfactuals

- Model

$$
\begin{aligned}
K &\sim \text{Normal}(\mu, \sigma) \\
\mu &= \alpha + \beta_M M + \beta_N N \\
\alpha &\sim \text{Normal}(0, 1) \\
\beta_M &\sim \text{Normal}(0, 1) \\
\beta_N &\sim \text{Normal}(0, 1) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

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



Multiple Regression Model

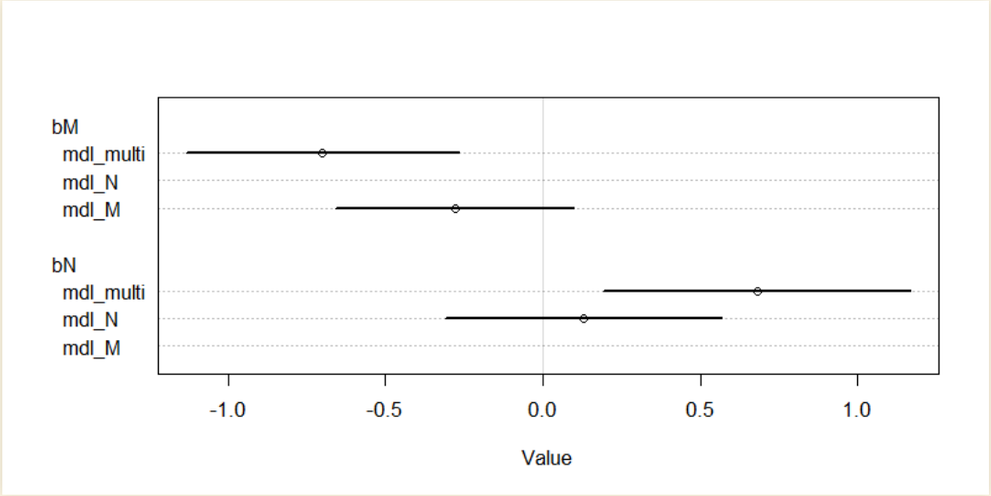
- Model

$$
\begin{aligned}
K &\sim \text{Normal}(\mu, \sigma) \\
\mu &= \alpha + \beta_M M + \beta_N N \\
\alpha &\sim \text{Normal}(0, 1) \\
\beta_M &\sim \text{Normal}(0, 1) \\
\beta_N &\sim \text{Normal}(0, 1) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

```
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
```

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



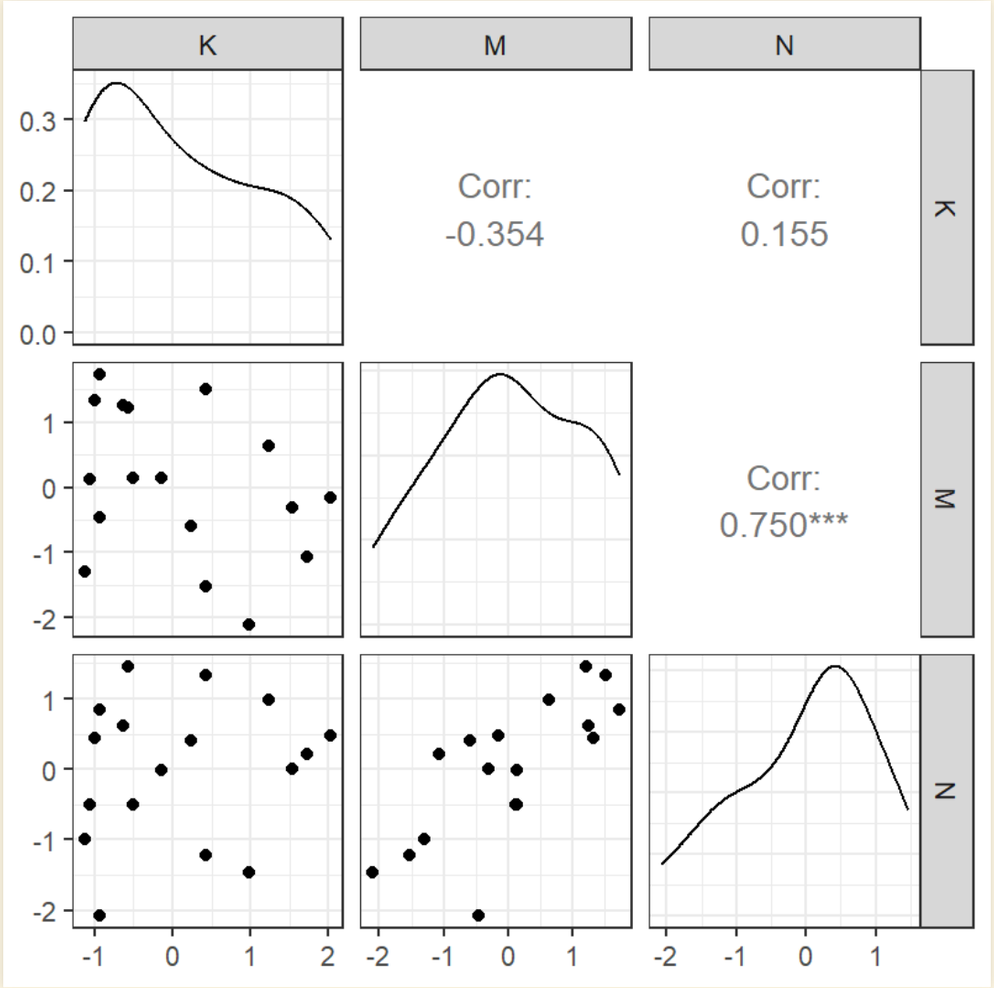
- This is the opposite of what we saw for divorce rates.
- The parameters for each predictor are consistent with zero for the single-predictor models
- When we include both predictors, the association with each is stronger.

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

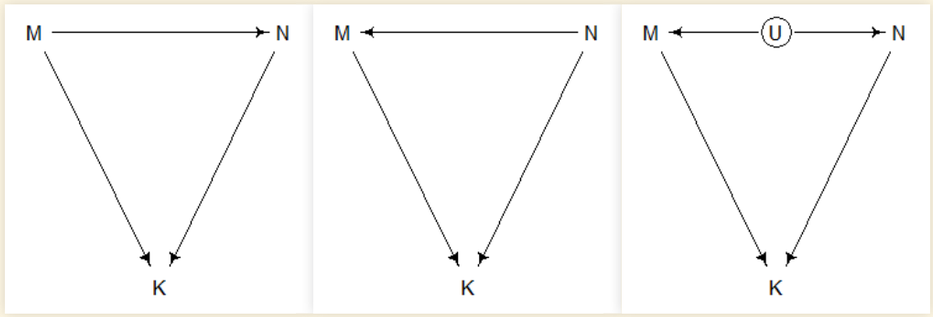
```
library(GGally)

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



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
    - Geologgic period: Permian, Triassic, Jurassic, Cretacious, ...
    - 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
```

- Model:

$$K \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha_{\text{Clade}[i]}$$
$$\alpha_j \sim \text{Normal}(0, 0.5) \text{ for } j = 1 \ldots 4$$
$$\sigma \sim \text{Exponential}(1)$$

```
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:

$$K \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha_{\text{Clade}[i]}$$
$$\alpha_j \sim \text{Normal}(0, 0.5) \text{ for } j = 1 \ldots 4$$
$$\sigma \sim \text{Exponential}(1)$$

```
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 -1.024 -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)")
```