Lab 7: Applied model specification

Not graded, just practice

Katie Schuler

2024-10-17

Practice your new *modeling* skills with these practice exam questions! Best to open a fresh Google Colab notebook and test things out! Refer to the study guide to find answers as well.

0.1 Primate brains

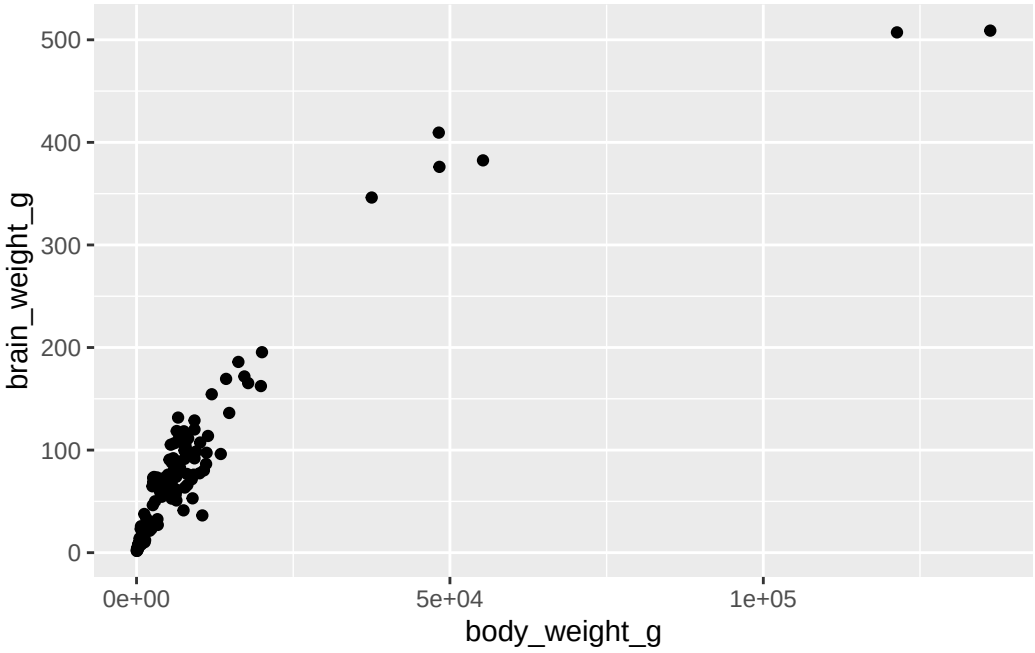
Primates have brains of varying sizes, and one possible explanation for this variation is differences in body size. Larger-bodied primates may tend to have heavier brains, but this relationship is not always straightforward. To investigate whether body size can reliably explain differences in brain weight across primate species, let's fit a model that predicts brain weight based on body size.

The data, in case you want to work with it yourself: primate brains

```
data <- read_csv("https://kathrynschuler.com/datasci/assests/csv/primate_brains.csv")
glimpse(data)
```

```
Rows: 144
Columns: 5
$ taxon          <chr> "Alouatta_caraya", "Alouatta_palliata", "Alouatta_pigra~
$ body_weight_g  <dbl> 5597, 6359, 8940, 6247, 1073, 870, 871, 239, 6409, 8034~
$ brain_weight_g <dbl> 52.72, 50.91, 52.97, 56.57, 21.41, 16.78, 17.21, 7.17, ~
$ diet_category  <chr> "Fol", "Fol", "Fol", "Fol", "Frug/Fol", "Frug", "Frug",~
$ group_size     <dbl> 6.68, 15.55, 5.93, 6.97, 3.00, 3.50, 3.51, 1.25, 16.40,~
```

```
ggplot(data, aes( x = body_weight_g, y = brain_weight_g)) +
    geom_point()
```



0.1.1 Type of model

(a) Is this a supervised or unsupervised learning problem?

- (A) Supervised
- (B) Unsupervised

(b) Is this regression or classification?

- (A) Regression
- (B) Classification

(c) Is the relationship between `brain_weight_g` and `body_weight_g` linear or nonlinear?

- (A) Linear
- (B) Nonlinear

(d) Is the nonlinear relationship linearizable or non-linearizable?

- (A) Linearizable nonlinear

- (B) Non-linearizable nonlinear
- (C) Neither, the relationship is linear

(e) What function could we choose to linearize this relationships?

- (A) Quadratic polynomial
- (B) Cubic polynomial
- (C) Log transformation
- (D) None, the relationship is already linear
- (E) None, the relationship is non-linearizable

0.1.2 Model specification

Suppose we specify the following model for the primate brains data: $log\_brain\_weight = w_1 \cdot 1 + w_2 \cdot log\_body\_weight$

(a) What is the response variable?

- (A) brain_weight_g
- (B) body_weight_g
- (C) log_brain_weight
- (D) log_body_weight

(b) What is the explanatory variable(s)?

- (A) brain_weight_g
- (B) body_weight_g
- (C) log_brain_weight
- (D) log_body_weight

(c) True or false, the functional form of this model can be expressed as a weighted sum of inputs? $y = \sum_{i=1}^{n} w_i x_i$

- (A) True
- (B) False

(d) Which of the following model terms are included in the model specification above? Choose all that apply.

- (A) Intercept term
- (B) Main term
- (C) Interaction term
- (D) Transformation term

(e) Specify the model equation in R notation.

```
# like this (explicit intercept)
log_brain_weight ~ 1 + log_body_weight

# or like this (implicit intercept)
log_brain_weight ~ log_body_weight
```

0.1.3 Fitted model

Suppose you fit the model with `lm()` and return the following:

```
Call:
lm(formula = log(brain_weight_g) ~ 1 + log(body_weight_g), data = data)

Coefficients:
       (Intercept)  log(body_weight_g)
           -2.4649              0.7752
```

(a) Which of the following is w_1 in the model specification $log\_brain\_weight = w_1 \cdot 1 + w_2 \cdot log\_body\_weight$

- (A) 1
- (B) -2.4649

- (C) 0.7752
- (D) Not enough information to determine this

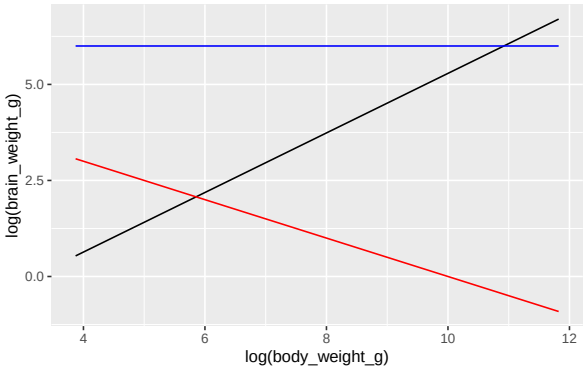
(b) Which of the following is w_2 in the model specification $log\_brain\_weight = w_1 \cdot 1 + w_2 \cdot log\_body\_weight$

- (A) 1
- (B) -2.4649
- (C) 0.7752
- (D) Not enough information to determine this

(c) Suppose a primate has a *log_body_weight* equal to 10. Which of the following would the model predict to be the primate's *log_brain_weight*?

- (A) 25.21
- (B) 5.29
- (C) -10.7752
- (D) Not enough information to determine this

(d) Which of the following figures could show the fitted model?



- (A) the blue line
- (B) the red line

- (C) the black line
- (D) Not enough information to determine this

0.2 Social brain hypothesis

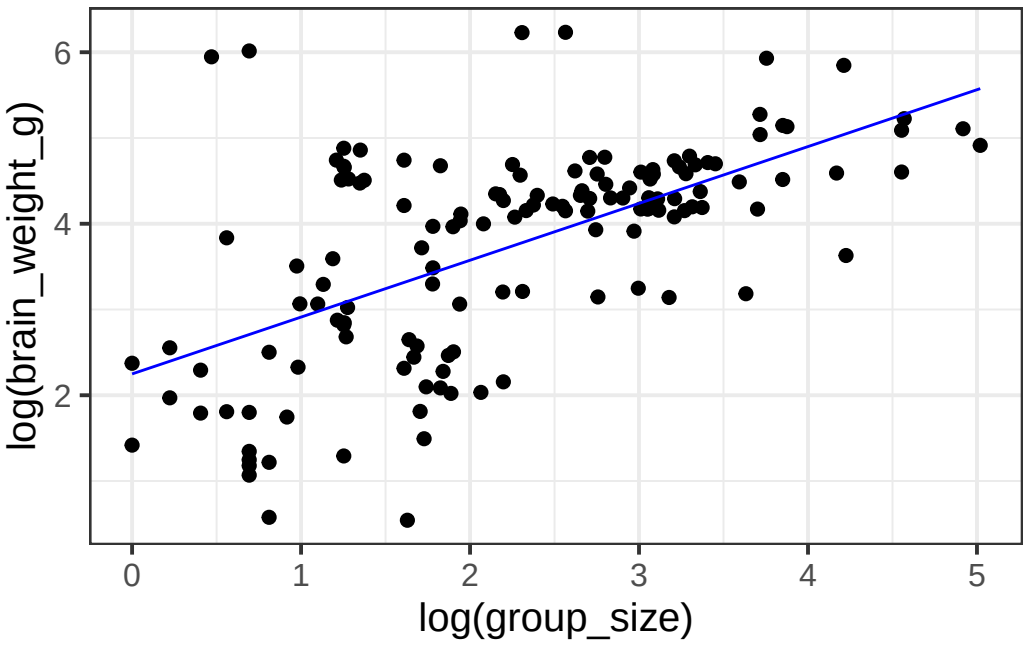
The **Social Brain Hypothesis** argues that the pressures of navigating increasingly complex social environments were a significant driver in the evolution of brain size and intelligence in humans and other primates.

Let's specify and fit this model in R.

```
model <- lm(log(brain_weight_g) ~ 1 + log(group_size),
    data = primate_brains)

primate_brains <- primate_brains %>%
    mutate(y_body_group = predict(model, primate_brains))

ggplot(primate_brains, aes(,
    y = log(brain_weight_g),
    x = log(group_size))
) +
geom_point(size = 2) +
geom_line(color = "blue", aes(y = y_body_group)) +
theme_bw(base_size = 15)
```



(a) Fill in the blank: how many inputs does this model have? _

0.3 b. Question

Specify the model as an equation

0.4 Answer

$\log(brain\_weight\_g) = w_1 \cdot 1 + w_2 \cdot \log(group\_size)$

or, if you created new columns in your data with the the log transformed data, for example:

```
data <- data %>%
    mutate(log_brain_weight = log(brain_weight_g)) %>%
    mutate(log_group_size = log(group_size))
```

then you could have written:

$log\_brain\_weight = w_1 \cdot 1 + w_2 \cdot log\_group\_size$

(c) Given the figure above, which of the following could be the free paramter estimate for w_1?

- (A) 1
- (B) 0.66
- (C) 2.25
- (D) 5
- (E) Not enough information to determine this

(d) Given the figure above, which of the following could be the free paramter estimate for w_2?

- (A) 1
- (B) 0.66
- (C) 2.25
- (D) 5
- (E) Not enough information to determine this

(e) Suppose we encounter a primate in a (log) group size of 4. What would the model predict for their (log) brain weight?

- (A) 3.5
- (B) 4.1
- (C) 4.9
- (D) 6.2
- (E) Not enough information to determine this

0.5 f. Question

Suppose we wanted to include log($body\_size\_g$) back into the model as an additional predictor of log($brain\_size\_g$). Specify the model in R.

0.6 Answer

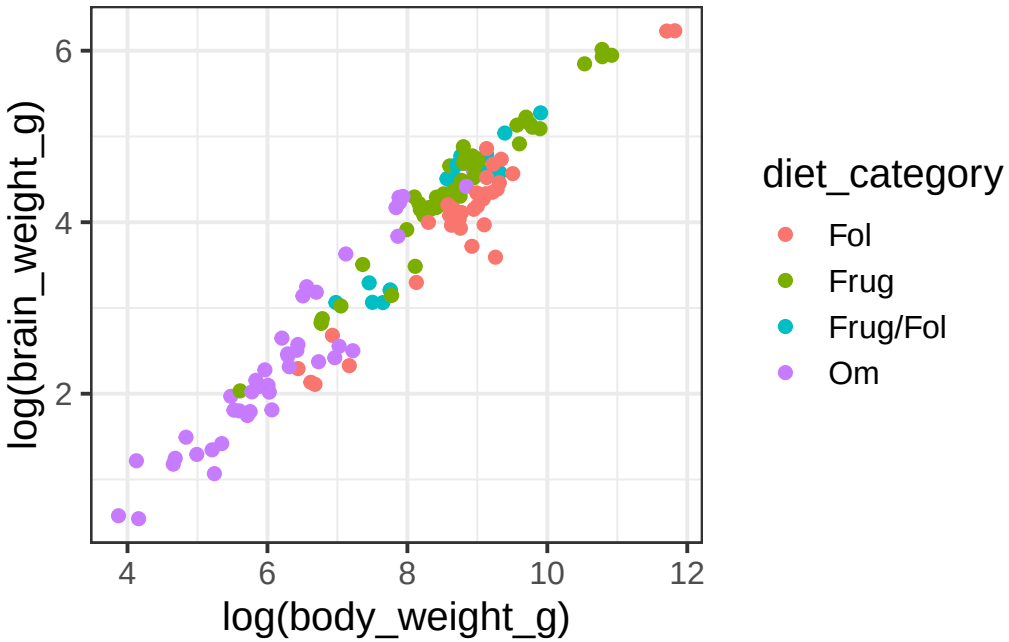
```
log(brain_size_g) ~ 1 + log(group_size) + log(body_size_g)
```

0.7 Fruit v Leaf eaters

Diet may influence the relationship between brain and body size in primates because the type of food a species consumes can impact its ability to meet the energy demands of a larger brain. Fruit-eating primates have access to energy-rich, easily digestible food, which could support the metabolic costs of both a large body and a larger, more complex brain.

Let's begin by adding `diet_category` to our plot mapped to the color aesthetic.

```
primate_brains %>%
    ggplot(aes(
        y = log(brain_weight_g),
        x = log(body_weight_g),
        color = diet_category
    )) +
    geom_point(size = 2) +
    theme_bw(base_size = 15)
```

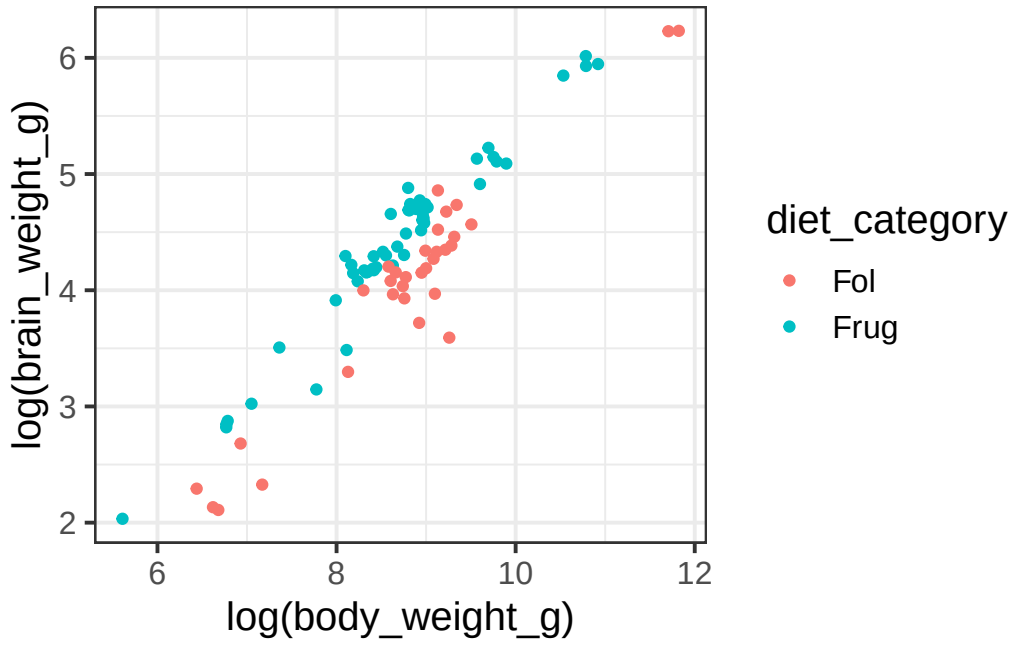


Frugivorous ("Frug") primates primarily eat fruit, while folivorous ("Fol") primates primarily consume leaves. The "Frug/Fol" category refers to primates that combine both fruit and leaf

consumption in their diet. "Om" stands for omnivores, which we might suspect is similar to "Frug/Fol" with more variation in diet. To simplify things, let's focus our analysis on just the Fol and Frug categories.

```
fruit_v_leaves <- primate_brains %>%
    filter(diet_category %in% c("Fol", "Frug"))

fruit_v_leaves %>%
    ggplot(aes(
        x = log(body_weight_g),
        y = log(brain_weight_g),
        color = diet_category
    )) +
    geom_point() +
    theme_bw(base_size = 15)
```



Suppose we specify a model that predicts brain weight by body size and diet category.

```
model <- lm(log(brain_weight_g) ~ log(body_weight_g) + diet_category, data = fruit_v_leaves)

model
```

```
Call:
```

```
lm(formula = log(brain_weight_g) ~ log(body_weight_g) + diet_category,
    data = fruit_v_leaves)

Coefficients:
       (Intercept)  log(body_weight_g)   diet_categoryFrug
           -2.8047              0.7778              0.4576
```

0.8 a. Question

Specify the model with a mathematical expression.

0.9 Answer

$\log(brain\_size\_g) = w_1 1 + w_2 \log(body\_size\_g) + w_3 diet\_category$

0.10 b. Question

Notice we did not include an interaction term between body weight and diet category. Why might a modeler make this decision?

0.11 Answer

A modeler might choose not to include an interaction term between body size and diet category based on exploratory visualization if the data suggests that the relationship between body size and brain weight is consistent across both diet categories. For instance, if scatter plots show parallel trends for frugivorous and folivorous primates, this could indicate that body size influences brain weight similarly, regardless of diet.

(c) True or false: the `diet_category` variable is categorical, so this is a classification problem.

- (A) True
- (B) False

0.12 c. Question

Write the *fitted model* as a mathematical expression.

0.13 Answer

$\log(brain\_size\_g) = -2.8047 \times 1 + 0.7778 \times \log(body\_size\_g) + 0.4576 \times diet\_category$

(d) Based on the fitted model returned by `lm()` above, which level of `diet_category` is the reference level?

- (A) Fol
- (B) Frug
- (C) Not enough information to determine this

0.14 e. Question

What is the model's prediction for a primate with a (log) body weight of 7 who eats leaves? Write your answer as a mathematical expression without simplifying it.

0.15 Answer

$\log(brain\_size\_g) = -2.8047 \times 1 + 0.7778 \times \mathbf{7} + 0.4576 \times \mathbf{0}$

(f) Which of the following could be a plot of the model?

0.16 Matching plots to equations

Plots A-F

(a) $y = w_1 1$ _____
(b) $y = w_1 1 + w_2 x$ _____
(c) $y = w_1 1 + w_2 z$ _____
(d) $y = w_1 1 + w_2 x + w_3 z$ _____
(e) $y = w_1 1 + w_2 x + w_3 z + w_4 a \times b$ _____
(f) $y = w_1 1 + w_2 x + w_3 z^2$ _____

0.17 Polynomials

1. What is the purpose of including polynomial terms in a linear model?
   - (A) To improve model interpretability
   - (B) To model nonlinear relationships
   - (C) To reduce overfitting in the model
   - (D) To ensure that residuals are normally distributed
2. Which of the following is an example of a quadratic polynomial term in a linear model?
   - (A) x
   - (B) x^2
   - (C) sqrt(x)
   - (D) log(x)
3. Why might higher-degree polynomial terms lead to overfitting in a linear model?
   - (A) Higher-degree terms make the model too simple
   - (B) Higher-degree terms force the model to fit the noise in the data
   - (C) Polynomial terms always reduce the model's flexibility
   - (D) Polynomial terms make the model biased
4. Which of the following models includes both linear and quadratic terms
   - (A) y = 0 + 1x
   - (B) y = 0 + 1x^2
   - (C) y = 0 + 1x + 2x^2
   - (D) y = 0 + 1x + 2x^3