Problem set 5

due Monday, October 20, 2025 at 11:59am (noon!)

Instructions

Upload your `.ipynb` notebook to **Gradescope** by **11:59 AM (noon)** on the due date. At the top of your notebook, include your **name**, **problem set number**, and any **collaborators**. Please include comments in your code to indicate which problem you are working on.

This problem set focuses on **model specification** — writing equations and model formulas that express relationships between variables.

Problem 1

Suppose you are studying songbird brains. You measured neuron density in regions involved in song learning for both **juvenile** and **adult** birds. Your dataset, songbird_neurons.csv, includes an identifier for each subject, an age group (`juvenile` or `adult`), a brain region (`HVC`, `RA`, or `Area X`), the number of neurons per cubic millimeter, and the total number of distinct syllables in each bird's song.

Your first question is about development: *How might you express a model that predicts neuron density as a function of age group?* Express your model symbolically using LaTeX notation, and then write the same model using R's formula syntax. Briefly explain what each term represents in your model.

Problem 2

Your next question concerns the relationship between brain structure and behavior. You wonder whether birds with denser neural circuits produce more complex songs. How could you specify a model that expresses **Song_Complexity** as a function of **Neuron_Density**? Write the model first as a LaTeX equation, then as an R formula. Describe what the intercept and slope represent in this context.

Problem 3

The dataset includes several potential explanatory variables for neuron density: **Brain_Region**, **Song_Complexity**, and **Age_Group**. Suppose you want to model neuron density as a weighted combination of these variables. Write down a linear model that includes all of them. Then consider whether any of these predictors might be redundant or uninformative, and propose a simpler model that balances parsimony and interpretability. Express both models in LaTeX and in R formula syntax, and explain why you might prefer the simpler one.

Problem 4

Imagine you are studying plant growth under different light conditions. In your dataset, polynomial_plants.csv, you have measured **Plant_Height** and **Light_Exposure** across different plant species. You suspect that the relationship between light and height might be nonlinear. Write three models expressing plant height as a function of light exposure: a linear, a quadratic, and a cubic polynomial. Represent each as a mathematical equation (for example, $y = w_0 + w_1x + w_2x^2$), and then as an R formula (for example, `Plant_Height ~ Light_Exposure + I(Light_Exposure^2) + I(Light_Exposure^3)`). Explain what adding higher-order terms allows the model to capture.

Problem 5

Finally, return to the familiar dataset animal_brain_body_size.csv. You want to model **Brain_Size** as a function of **Body_Size**. Write one model using the raw variables and another using log-transformed variables. Present each as a mathematical equation and as an R formula. Explain why a log transformation might be appropriate for comparing brain and body size across species.

Challenge (optional)

In Problem 3, you considered several predictors of neuron density. Suppose you now suspect that the effect of age group depends on brain region — that is, the two variables interact. How could you express this relationship using an interaction term? Write the model both in LaTeX and in R formula notation, and briefly describe what kind of relationship such an interaction term represents.