MTH 228 Calculus and Probability for the Life Science II. Spring 2023.

Extra Credit Project.

Deadline: **June 12, 2023 11:59PM.**
i.e. Monday of Finals Week, same day as the Final Quiz.

Points: **This project contributes up to 20 points under the Projects rubric item.**
For this class, projects are worth 20% of the final grade with a 150 point total. Therefore, an additional 20 points contributes about 2.7 percentage points to the final grade (i.e. a little over a quarter of a letter grade)

1 Introduction.

An **allometric model** is a function in the form $y = ax^k$ for some constants a and k that relates two biological processes and/or characteristics [Shingleton, 2019]. For this project, we will be analyzing allometry data from the paper *The Allometry of Brain Size in Mammals* [Burger, George, Leadbetter, Shaikh 2019]. The data set is publicly available on Dryad, linked here. You are welcome to read the paper but you're not expected to do so. To do this project, we only need to work with the data set.

The paper claims that the brain-body allometry across mammals is given by $y = 10^{-1.26}x^{0.75}$ with x representing body mass in grams (g) and y that of brain mass in (g).* This is supported by a regression calculation on the brain-body mass data across 1522 mammalian species. The main goal of this extra credit project is to have you do the calculations yourself, using the tools introduced to you in this class (i.e. matrix algorithm/calculation). While you can use other resources to do the linear regression (e.g. Desmos, Excel), you do have to show the matrix calculation to get the extra credit.

* The abstract of the paper actually states that $a = -1.26$, i.e. the relationship is given by $y = -1.26x^{0.75}$. However, this is contradicted by the calculations in the paper. We've added the correction that $\log(a) = -1.26$ and $y = 10^{-1.26}x^{0.75}$ as stated above.

2 Matrix Calculation for Allometric/Linear Regression.

Given a set of allometric measurements $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, we aim to find the function $y = ax^k$ with constants a and k such that the squares of the differences between the actual values y_i and the expected values $y(x_i)$ are minimized. In other words, the function

$$\begin{aligned} D(a,k) &= (y_1 - y(x_1))^2 + (y_2 - y(x_2))^2 + ... + (y_n - y(x_n))^2 \\ &= (y_1 - a(x_1)^k)^2 + (y_2 - a(x_2)^k)^2 + ... + (y_n - a(x_n)^k)^2 \end{aligned}$$

is minimized with respect to a and k. This property is what makes the function the "best-fit" function. This is also why finding this function is sometimes called the least squares problem, with the resultant function being the least squares solution (LSS).

First, we assume that there does exist constants a and k (i.e. there does exist a function $y = ax^k$) such that the equation $y_i = a(x_i)^k$ is satisfied for all $i \in \{1, 2, ..., n\}$. This results in a systems of equations. However, that system is not linear and therefore, we can't apply our matrix algebra tools. We can fix that by working with the linearized version of $y = ax^k$ instead. Recall that we can get this linearized equation by applying log(...) to both sides of the equation, i.e.

$$\log(y) = \log(a) + k\log(x)$$

We should get the following linear system of equations:

$$\begin{aligned} k\log(x_1) + \log(a) &= \log(y_1) \\ k\log(x_2) + \log(a) &= \log(y_2) \\ &\vdots \\ k\log(x_n) + \log(a) &= \log(y_n) \end{aligned}$$

With some manipulation, it's possible to express this system as a matrix equation

$$\mathbf{M}\begin{pmatrix} k \\ \log(a) \end{pmatrix} = \mathbf{y}_{\log}$$

with $\mathbf{M}$ an $n \times 2$ matrix and $\mathbf{y}_{\log}$ a column vector with n entries.
Remark: We've intentionally left out an exact description of $\mathbf{M}$ and $\mathbf{y}_{\log}$ since that is part of the project.

Finally, it can be proven that one way to find the least squares solution and therefore, the best-fit function, is to use the following formula:

$$\begin{pmatrix} k \\ \log(a) \end{pmatrix} = (\mathbf{M}^\top \mathbf{M})^{-1} \mathbf{M}^\top \mathbf{y}_{\log}$$

You are expected to use this method for the extra credit project.

3 Brain-Body Mass Data Set

The data set `BrainAllometry_Supplement_Data.csv` (available here) is a csv file containing 14 measurements from 1522 mammalian species. We're mainly concerned with the following fields:

1. `Binomial`: the binomial name, i.e. genus and species.
2. `order`: the taxonomic order.
3. `Mean_body_mass_g`: the mean (i.e. average) body mass in grams across samples of the same species.
4. `Mean_brain_mass_g`: the mean (i.e. average) brain mass in grams across samples of the same species.

4 Requirements for Extra Credit

We'll be using a subset of the data set for our calculations. That is, we won't be processing all 1522 measurements. Instead, we'll be focusing on specific orders in this data set.

1. **Afrosoricida.** 13 measurements.
2. **Didelphimorphia.** 16 measurements.
3. **Lagomorpha.** 15 measurements.
4. **Peramelemorphia.** 14 measurements.

You'll be responsible for filtering out the data that you need to use. For all calculations, assume that body mass is the independent variable and brain mass is the dependent variable. In other words, given the model $y = ax^k$, x represents mean body mass in grams and y the mean brain mass in grams.

For each order, you have to provide the following:

1. **A short description of the order.**
   This may include a short list of animals categorized under this order. Images are greatly encouraged.

2. **The list of measurements under the order.**
   Your list must include (1) the binomial name, (2) the mean body mass, and (3) the mean brain mass.
3. **The matrix equation to be solved explicitly stated.**
   That is, the entries of the matrices $\mathbf{M}$ and $\mathbf{y}$ must be identified and the equation, with the matrices written, must be included. For brevity, you may use ellipses if the other entries of the matrix are clear from context (based on what has been presented in your submission before this part).
4. **Tools used in the matrix calculation and how they were used.**
   You are not expected to do the calculation by hand. Instead, you must include the steps you've done to do the calculation. All measurements under each order must be used for full credit.

   For example, if you've used the Desmos matrix calculator or Symbolab, include what you've entered, what the calculator gave you explicitly, and how the result factors into your calculations. Screenshots are expected, if you were to use this method. If you've used code (e.g. Python), include the code that you've used in your submissions, including functions that you've defined and the relevant function calls.

   Remark 1: While you can have some software do the linear regression for you (as is the usual practice), that is not the point of this project. If just the values of a and k are given without supporting calculations, your submission will **not** be accepted. In fact, these constants are identified in the paper. You may use that to check your calculations and results.

   Remark 2: If you have some experience in Python, we recommend doing the calculations in a Jupyter notebook. This is not required, of course, but it makes the calculations easier since the calculations between orders are done very similarly; and it makes presentations of answers easier too. We also recommend that you work with the `pandas` and `numpy` packages.
5. **The exact expressions of both the linearized model and the allometric model.**

   Remark 3: Since we're doing calculations on a subset of data, the constants you get may not agree with the overall model. For example, over all measurements, that resultant allometric constant is $k = 0.75$. However, the constant for order Perissodactyla ($n = 11$ measurements) is calculated to be $k_{\text{Perissodactyla}} = 0.45$. This discrepancy is included and addressed in the paper.
6. **A scatter plot of the linearized version of the data points and the graph of the resultant linear model on the same figure.**
   Add labels to the plot (i.e. identifying what exactly the plot represents), the axes (i.e. what the values on the axes represent and the corresponding units), etc. Decide on appropriate ranges for the axes such that most of the data points are included and the pattern (in this case, linear) is visible.

   Remark 4: You may include the data and results for all four orders in one figure. However, points under different orders must be distinguishable (by color, for example) and a legend must be added.
7. **A scatter plot of the data points and the graph of the resultant allometric model on the same figure.**
   Same standards as the previous item.

Your submission must be presented in a clean and organized matter. Think of it as a lab report. While it is recommended that you type up your submission, handwritten work will also be accepted.

As counterexamples, the following will **not** be accepted: (1) scratch work or pages of calculations with ambiguous direction and/or without clear context; (2) images that are taken from a weird angle and/or are blurry; (3) a submission where the supposed order of content is unclear; (4) work where multiple lines or work are crossed out or stricken-through to eliminate them from the submission. Note that this list does not cover all scenarios.