Problem 1. **(mammals dataset)** Consider the mammals dataset in the MASS package. We choose brain weight as the response variable and body weight as the predictor variable.

A. We work in log scale (both axes). Why is that? Illustrate with a plot or two, and offer some brief comments.

B. Produce a scatterplot. Make the plot nice, properly labeling it, and adding a title. Identify African Elephant, Asian Elephant, and Human, in the scatterplot.

C. Fit an affine function (i.e., a line) by least squares regression. Add the line to the scatterplot (in a color of your choice). What is the intercept and slope of the resulting line? Compute the Student 99% confidence intervals for them (without correcting for multiple inference).

D. Produce a plot of the Cook's distances. Offer some brief comments.

E. Find a species of mammal not represented in this dataset. Name the mammal and provide its (typical) body and brain weights, denoted $(x_{\text{new}}, y_{\text{new}})$ below. Indicate the source where you obtained this information (e.g., a URL).

F. Using the model you have fitted, build a 90% prediction interval for the expected brain weight of a mammal with body weight x_{new}. Offer some brief comments. Add this interval to the scatterplot.

Problem 2. (Collinearity among predictor variables) Consider a standard normal model where the sample is iid from the distribution $y = -1+2x_1-2x_2+\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 0.5^2)$ independent of (x_1, x_2), where $x_1 \sim \mathcal{N}(0, 1)$ and $x_2 = ax_1 + \sqrt{1-a^2}\, z$ with $a \in [-1, 1]$ and $z \sim \mathcal{N}(0, 1)$ independent of x_1 — so that $x_2 \sim \mathcal{N}(0, 1)$ as well and $\text{cor}(x_1, x_2) = a$. We consider $a \in \{0.8, 0.9, 0.99\}$ and sample size $n \in \{20, 50, 100\}$. For each combination (a, n), repeat the following $N = 1000$ times. Fit a simple linear model by least squares. Compute the 95% confidence intervals for the coefficients, and for each interval, recording its length and whether it contains the true value of the parameter. Also, record the variance inflation factors (one for each LS coefficient). With all the N repeats, for each confidence interval, produce the average length and the proportion of times it contained the true value of the parameter. Gather these across the combinations of (a, n) and produce a few plots. Separately, for each VIF, produce the average, and plot these across the combinations of (a, n). Offer some comments.

[Optional] Repeat with a setting that includes a larger number of variables and where one of the variables is correlated with not just one but several other variables.