Reminders:

- The problem set is handed in on Canvas.
- Assignments are due at the start of class on Tuesdays (11:00am). The Canvas website will not accept late assignments.
- You may upload one single file, or alternately one file with your written work and one file with your code—if you submit two files, each problem should be entirely contained in one of the files. For the code component, we recommend using R Markdown (via R Studio) to produce a single file with your code, plots/output, and written explanations/comments.
- If you're photographing/scanning handwritten work with a smartphone, we recommend the free CamScanner app to produce a single PDF file containing all the pages. Do not submit each page as a separate file.
- If you are having trouble uploading to Canvas and run out of time, please email your work to the instructor or TA by 11:00am as proof of completion.

---

1. The `gala` data set from the Faraway textbook counts the number of tortoise species on different Galapagos islands. Each data point is one island. We will be interested in how the number of endemic species (meaning, species that live only on that island) relates to the size of the island (its area). Here is code to load this data:

```
> library(faraway)
> data(gala)
> x = gala$Area
> y = gala$Endemic
```

   If you've never used the `faraway` package in R, you will need to install it first before you can run the code above:

```
> install.packages('faraway')
```

   (a) Make a scatterplot of the X and Y values. Do you feel that a linear model is appropriate for this data set? Are there any features of this data set that would make you question this model?

   (b) Now try replacing X with the log-area:

```
> x = log(gala$Area)
```

   Make a new scatterplot. Do you feel that a linear model is appropriate for this data set? Are there any features of this data set that would make you question this model?

   (c) From this point on we'll use log-area as our X. Compute the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ and the variance estimator $\hat{\sigma}^2$ without using the `lm` command or any other regression commands, i.e. show the raw calculations for computing $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$. (It's of course fine to use vector multiplication and other elementary operations, you don't need to add up numbers individually.)

   (d) According to your answer above, what's the predicted number of endemic species for an island whose area is 2.0?

2. This problem continues with the least squares regression of Y = number of endemic species on X = log-area, from the previous problem.

   (a) In R, compute the correlation between the vector of fitted values and the vector of residuals. Explain the answer you see—how does it relate to the properties of least squares?

   (b) In R, compute the correlation between the vector of residuals and the variable `gala$Nearest`, which is the distance from each island to its nearest island (i.e. it's large if the island is far from any other island). Explain the answer you see—why does it make sense in the context of the data?

3. Consider a data set consisting of X values $X_1, \dots, X_n$ and Y values $Y_1, \dots, Y_n$. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the output of OLS on this data set. Now define
$$\tilde{X}_i = c \cdot (X_i + d)$$
for each $i = 1, \dots, n$, where $c \neq 0$ and d are arbitrary constants. Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the output of OLS run on data $\tilde{X}_1, \dots, \tilde{X}_n$ and $Y_1, \dots, Y_n$. Write equations for $\tilde{\beta}_0$ and $\tilde{\beta}_1$ in terms of $\hat{\beta}_0$ and $\hat{\beta}_1$ (and in terms of the constants c and d), and prove that your answer is indeed the OLS estimator.

4. In this problem we'll examine the effect of the normality assumption on the validity of inference for OLS.

   (a) Generate a simulated data set of size $n = 100$ as follows:

   - Draw $X_i \overset{\text{iid}}{\sim} \text{Uniform}[-1, 1]$
   - Set $\beta_0 = \beta_1 = \sigma^2 = 1$ and $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

   Compute the OLS estimate of $\hat{\beta}_1$ and a 90% confidence interval for the true coefficient β_1. (You may use `lm(...)` and `summary(lm(...))` for this problem.) Repeat this 1000 times. What coverage rate do you observe empirically, i.e. what percent of the time does the confidence interval actually contain β_1?

   (b) Now repeat the same experiment, but now the data is generated in a way that violates the variance assumption: draw $\epsilon_i \sim N(0, \sigma_i^2)$ where $\sigma_i^2 = X_i^2$, i.e. the noise variance is larger for more extreme X values. What coverage rate do you observe empirically, i.e. what percent of the time does the confidence interval actually contain β_1? (To be clear, you should construct the interval using the same OLS method as in part (a), i.e. pretending that you did not know that the assumptions are violated.)

   (c) Again repeat the same experiment, but now $\sigma_i^2 = (1 - |X_i|)^2$, i.e. the noise variance is smaller for more extreme X values. What coverage rate do you observe empirically, i.e. what percent of the time does the confidence interval actually contain β_1?

   (d) Explain the results you observe in parts (a), (b), (c). In particular, you should explain why you observe opposite trends in (b) versus (c).

5. Suppose you have a data set where X takes only two values while Y can take arbitrary real values. To consider a concrete example, consider a clinical trial where $X_i = 0$ indicates that patient i received the placebo, while $X_i = 1$ indicates that patient i received the treatment, and Y_i is the real-valued outcome for patient i, e.g. blood pressure. Let $\bar{Y}_P$ and $\bar{Y}_T$ indicate the mean outcome values for the placebo group and for the treatment group, respectively. What will be the values of the OLS coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ in terms of these group means? Justify your answer (a short & simple proof is preferred, without long calculations).