Homework 3

Due: Friday Feb. 27, 2025 at 10pm

This assignment has a total of 26 pts possible. Your score out of 24 will noted and scaled to 5 points (maximum of 5). This assignment is due *after* the exam, but I recommend doing as much as you can *before* the exam, since the practice will be helpful.

Question 1 – Conceptual (1 pt each)

- **Part A** (Study Design) What do we need to be true of a sample in order to generalize results to a population?
- **Part B** (Study Design) Why does randomization of treatments to experimental units let us make causal conclusions?
- **Part C** What form of relationship does Pearson's correlation measure?
- **Part D** If a scatterplot has a value of r = .9, does that mean there must be a linear relationship between the variables? Explain.
- **Part E** What does the phrase 'correlation $\neq$ causation' mean in your own words?
- **Part F** What form of relationship needs to exist between quantitative variables to do linear regression?

Question 2 – Cat Regression (9 points)

The problem includes a dataset with 144 cats, included with each observation is the sex of the cat, as well as body weight (kg) and heart weight (g).

```r
## Read in cat data
cats <- read.csv("https://collinn.github.io/data/cats.csv")
```

Part A: Use `lm()` to create a linear model in R predicting the weight of a cat's heart using body weight as an explanatory variable. Write the formula for the regression line in context.

Part B: Interpret the slope in context.

Part C: Interpret the intercept in context. Is the intercept meaningful?

Part D: What is the predicted heart weight of a cat that has a body weight of 3kg?

Part E: What is the residual for a cat with a body weight of 3kg that actually has a heart weight of 12g. Is this an under- or over-prediction?

Question 3 – Cherry Trees (6 points)

The dataset below includes information on 31 black cherry trees felled in the Allegheny National Forest, Pennsylvania. For each tree, it includes three variables, one for each diameter (in), height (ft), and volume (cubic ft).

```r
## Cherry tree data
cherry <- read.csv("https://collinn.github.io/data/cherry.csv")
```

Part A: Create two scatterplots of the data comparing diameter with volume and height with volume, in each case letting volume be the response variable. Based on these plots, which variable do you think would be a better predictor of volume?

Part B: Create two linear models, ones for each of the plots created in Part A (that is, with volume as a response variable in both models). Based on the `summary()` output, which of these models has a higher R^2 value? Is this consistent with what you decided in Part A?

Part C: Using the model with the highest R^2 in Part B, write the linear equation for predicting a tree's volume. Interpret both the slope and the intercept. Is the intercept meaningful in this case?

Question 4 – Matching Correlations (1pt)

IMS - Section 7.5, Question 7

Question 5 – Study Design Practice (4pts)

IMS - Section 2.5, Question 15 – This question is a review of the study design content we covered in week 4, but it will be extremely helpful for the exam.