Problem set 3

due Thursday, October 26, 2023 at 11:59pm

⚠ Similar exercises will be worked through in the R tutorial on Oct 19

Estimated time: 6 hours

Allocate about **1 hour per problem**, though some will take longer than others. You may need more time if programming is completely new to you, or less if you have some experience already.

Instructions Upload your `.ipynb` notebook to gradescope by 11:59pm on the due date.

- Note that each problem will be graded according to this rubric. Solutions that include packages or functions not covered in this course will recieve a score no higher than 2.
- You may collaborate with any of your classmates, but you must write your own code/solutions, understand all parts of the problem, and name your collaborators.
- You should also cite any outside sources you consulted, like Stack Overflow or ChatGPT, with a comment near the relevant lines of code (see example below). Recycled code that has not been cited will be considered plagerism and receive a zero.

```
# code here was inspired by user2554330 on stack overflow:
# https://stackoverflow.com/questions/69091812/is-everything-a-vector-in-r
```

Problem 0

> not graded

Create a new colab R notebook. Please include the title "Problem set 3", your name, the date, and any collaborators somewhere at the top.

Problem 1

The dataset below includes data simulated from work done by Carolyn Rovee-Collier. Dr. Rovee-Collier developed a new way to study very young babies' ability to remember things over time: the "mobile conjugate reinforcement paradigm". See a video of this paradim here and a nice description from Merz et al (2017) here:

> "In this task, one end of a ribbon is tied around an infant's ankle and the other end is connected to a mobile hanging over his/her crib. Through experience with this set-up, the infant learns the contingency between kicking and movement of the mobile. After a delay, the task is repeated, and retention is measured by examining whether the infant kicks more during the retention phase than at baseline (i.e., spontaneous kicking prior to the learning trials; Rovee-Collier, 1997). Developmental research using the mobile conjugate reinforcement paradigm has demonstrated that both the speed of learning and length of retention increase with age"

```
"https://kathrynschuler.com/datasets/rovee_collier_1989.csv"
```

The simulated dataset includes 4 variables:

1. `ratio` - the measure of retention
2. `day` - the delay in days (1 through 14)
3. `age` - the age group: 2 month olds or 3 month olds
4. `age_recoded` - the age group recoded as 0 (2 month olds) and 1 (3 month olds)

Explore these data with (at least) glimpse and a scatterplot. Include ratio on the y-axis, day on the x-axis, and color the dots by age. You may include any other explorations you wish to perform.

Problem 2

Suppose you have specified that you will use a linear regression model to predict the simulated Rovee-Collier babies' retention ratio by day and age. Your model can be represented by the following equation:

$y = w_0 + w_1x_1 + w_2x_2$, where:

- $y =$ ratio
- $x_1 =$ day
- $x_2 =$ age

Fit the specified model using ordinary least squares approach with each of the three different functions we learned in the tutorial: (1) with `lm`, (2) with `infer`, and (3) with `parsnip`. Did all three ways return the same parameter estimates? Explain why or why not.

Problem 3

Given the specified model and the parameters estimated in problem 2, compute the sum of squared error for the fitted model.

> Note: if you are stuck on Problem 2, you may proceed with this problem by using all zeros as your parameter estimates.

Problem 4

Expanding on problem 3, write a more general function that would allow you to compute the sum of squared errors for the model specified in problem 2. Your function should take two inputs: (1) the data and (2) the parameter estimates. Your function should return a single value as output. Test your function with each of the following parameter values:

1. 0, 0, 0
2. 2, -3, 5
3. 1, 2, 3

Which of these three options fit the data best? How do you know?

Problem 5

Use the `optimg` package to find the optimal parameter estimates for the model specified in problem 2 via gradient descent. Initialize your search with $b_0 = 0$, $b_1 = 0$, and $b_2 = 0$. How many iterations were necessary to estimate the parameters? Are the parameters estimated by your gradient descent the same as those returned by `lm()`? Explain why or why not.

Problem 6

The function given below finds the ordinary least squares estimate via matrix operations given two inputs: X, a matrix containing the input/explanatory variables, and Y, a matrix containing the output/response variable.

```
ols_matrix_way <- function(X, Y){
 solve(t(X) %*% X) %*% t(X) %*% Y
}
```

Use this function to estimate the free parameters of the model specified in problem 2. Are the parameters estimated by the matrix operation the same as those returned by `lm()`? Explain why or why not.