Problem set 4

due Monday, Nov 4, 2024 at noon

Run your code

Make sure your code can run before submission! Runtime > Run all

Instructions Upload your `.ipynb` notebook to gradescope by 11:59am (noon) on the due date. Please include your name, Problem set number, and any collaborators you worked with at the top of your notebook. Please also number your problems and include comments in your code to indicate what part of a problem you are working on.

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:

$y = w_1x_1 + w_2x_1 + w_3x_2$, where:

- y = ratio
- x_1 = 1 (constant)
- x_2 = day
- x_3 = age

Fit the specified model using ordinary least squares approach with each of the three different functions we learned in class: (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 any parameter estimates of the model specified in problem 2. Your function should have two arguments: (1) `data` and (2) 'par', which is a vector of the parameter estimates. Your function should return a single value as output. Test your function with each of the following parameter values:

1. 0.1, 0.2, 0.3
2. -10, -30, 5

3. 3, 2, 1

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 $b_0 = -100$, $b_1 = 100$, and $b_2 = 0.5$. How many iterations were necessary to estimate the parameters? Are the parameters estimated by your gradient descent the same as those returned by `lm()` in Problem 2? 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.

```r
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.

Problem 7

Estimate the accuracy of the specified model on the population using bootstrapping or k-fold cross validation (choose one, not both). Use the `collect-metrics()` function to return the R^2 value.

Challenge (optional)

No points awarded or removed for this question! Just for fun!

TBD.