Problem set 6 (optional)

due Friday December 15, 2023 at 11:59pm

⚠ Problem set 6 is fully optional. If you complete psets 1-5, we will drop your lowest of 5. If you complete pset 6, we will drop your lowest of 6.

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
```

As always, create a new colab R notebook. Please include the title "Problem set 6", your name, the date, and any collaborators somewhere at the top.

Dataset information

In this pset, we will return to the data from Problem set 3. To remind you, here is the data.

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)

Suppose you would like to predict the babies' retention ratio by day and age (as you did in pset 3).

Problem 1: Data exploration

- Load the dataset and perform the following exploratory data analysis
- Examine the dataset's structure (glimpse), summary statistics, and handle missing values if any.
- Visualize the pairwise relationship between the response variable and each of the three explanatory variables using appropriate plots (scatter plots, histograms, etc.).

Problem 2: Model specification

- Choose an appropriate type of model and explain why you've selected this model (e.g. regression, classification)
- Specify the functional form of the model with an equation in a colab text block (e.g. $y = \beta_0$).
- Specify the model in R code (e.g. `y ~ 1`)

Problem 3: Model fitting

- Fit the specified model.
- What does the model tell us? How does babies' retention ratio change with age and day?
- What is the meaning of the intercept of this model?

Problem 4: Conditions for inference

- Use the `check_model` function to test whether the conditions for inference are satisfied.
- Does the model satisfy the conditions for Linearity, Normality, and Equality? For each condition, describe what it refers to and explain how the plot returned by `check_model` suggests whether it does (or does not) satisfy the condition.
- Is the condition for Independence satisfied? Explain why or why not. If not, describe some options for fixing this.

Problem 5: Hypothesis testing models

- Perform a hypothesis test on your model with `anova`
- Is there a significant main effect of age or day? How do you know?
- Interpret what this means for the research question: How does babies' retention ratio change with age and day?

Problem 6: Model predictions

- Using the coefficients returned from your model in Problem 3 what is the model's predicted retention ratio for:
  - a baby who is 2 months old after a delay of 5 days?
  - a baby who is 3 months old after a delay of 20 days?
  - a baby who is 2 months old after a delay of -2 days? This one doesn't make sense (we can't have a delay of -2 days). Does the model generate a prediction in this case? How should we interpret this prediction?