Problems for Data Scientists

1 Problem

This problem covers a few steps of data analysis and modeling that use a dataset in the accompanying file. The file contains measured data for a set of subjects and results from simulations of those subjects. The goal of this exercise is to perform analyses that will help determine whether the simulations are *consistent* with the data, meaning that statistical analysis cannot differentiate data from simulation, and write a brief summary for non-experts.

Setup

The file format is `pandas HDFStore`, and the contents are structured as:

- There are a series of tables, named `data`, `subject_0`, `subject_1`, ...
- The `data` table contains values for all subjects (the `DataFrame` index) and covariates (the `DataFrame` columns).
- The subject tables (e.g., `subject_5`) contain values for a single subject for a set of simulations (the `DataFrame` index) and all covariates (the `DataFrame` columns).

1.1 Data Loading and Transformation

1. Load the data from the file into two data structures: a single `DataFrame` for the data table, and a list of `DataFrame`s for the subjects.
2. For the list of subject `DataFrame`s, reorganize the data to get a list over simulations. That is, each entry in the list should be a `DataFrame` with values for a single simulation, with same layout as the `data` table.

1.2 Analysis

There are 3 covariates, a, b, and c. Compute and report:

- The mean and standard error of a for the `data`.
- The Pearson correlation coefficient between b and c for the `data`.

Next, for each simulation compute the mean of a and the Pearson correlation coefficient between b and c. Then report the empirical p-value of these statistics for the `data` under the distribution of statistics from the simulation.

1.3 Modeling

Write a function that will, for each simulation, evaluate via 5-fold cross validation the ability of a logistic regression model to differentiate the `data` from the simulation. Use accuracy as your metric; the function should report the mean ROC-AUC over the held-out fold across the 5 folds. Histogram the mean AUC values across simulations, and report the mean and standard deviation of this distribution.

1.4 Reporting

Write a 1-2 paragraph summary of your findings and what you would do next to further answer the central question. Assume it will be read by someone who understands the goal of the exercise but is a non-expert.