# 2023 Data Analysis - Qualifying Exam
# Department of Statistics, UC Irvine

**Handed out:** Friday, June 16, 2023 at 1:00 pm
**Due:** Friday, June 23, 2023 at 5:00pm

**Turning In Your Exam:** Upload your complete solution, with you identifier number using the link provided to you by Department Manager Laura Swendson (laura.s@uci.edu) **no later than 5pm on Friday, June 23rd.** **LATE EXAMS WILL NOT BE ACCEPTED AND WILL NOT BE SCORED.**

## 1  Background information

Declining physical function is a common problem for older adults. Several factors are assumed to be associated with steeper declines in physical function over time. Here, we are interested in understanding whether **perceived discrimination is associated with greater decline in physical function in older adults**. To this end, you are provided with a dataset containing up to 5 repeated measurements for each of $n$=7,836 individuals, for a total of 20,059 observations. The dataset is a random subset from the *Chicago Health and Aging Project* [1], a longitudinal population-based study conducted over 20 years, from 1993 to 2012, in neighborhoods in the south side of Chicago. Individuals in the study were all adults over 64 years of age, who were assessed over time via interviews and underwent clinical exams. Data collection occurred in multiple waves, for a total of 6 waves each lasting 3 years.

The outcome of interest is a measure of *physical function* (`physical`), assessed at each visit through the Short Physical Performance Battery (SPBB) [2], a tool used to evaluate the level of functioning of lower extremities in older adults. The SPPB exam was administered by having the subject perform three activities: a chair stand test; a 3-part balance test with side-by-side, tandem and semi-tandem stance; and a gait speed test. Each activity was scored on a scale from 0 to 5, with higher scores indicating better physical function. The scores on each individual assessment test were then summed together to yield a summary score of physical function ranging from 0 to 15. A higher total score indicates better performance.

The primary exposure of interest, *perceived discrimination* (`discrim`), was measured only at the first visit and was quantified through the Everyday Discrimination Scale (EDS) [3]. EDS consists of 9 questions aimed at assessing whether a person believes they have experienced discrimination in different aspects of their day-to-day life because of their race, ethnicity, gender, age, appearance or sexual orientation. The answers to each question were coded as binary (0 or 1) and the scores to each questions were added to yield a total score ranging from 0 to 9. Higher scores indicate greater perceived discrimination.

The dataset also contains information on the subjects' age, gender, race, BMI, years of education, income, number of self-reported medical conditions, and perceived stress. The latter was measured with a perceived stress scale, containing 6 items each scored on a 0 to 3 scale, and summed to yield a total score. Also in this case, a higher total score is representative of greater perceived stress.

# 2 Scientific goal

The overall scientific goal of this study is to assess the impact of perceived discrimination on older adults' physical function.
Specifically, your data analysis should:

1. Hypothesize a potential mechanism for the association between perceived discrimination and physical function in older adults at baseline. Carry out a cross-sectional statistical analysis to test your hypothesis and examine this association. Quantify any differences in the association between perceived discrimination and physical function by race/ethnicity (Non-Hispanic White vs. Non-Hispanic Black).

2. Determine whether perceived discrimination is associated with a greater decline in physical function in Non-Hispanic Black compared to Non-Hispanic White older adults. Conduct an appropriate statistical analysis to assess this hypothesis. Provide the estimated profiles for physical function over time for different perceived discrimination levels with confidence intervals.

# 3 General instructions

You are to analyze the data to address the objectives listed above. You should properly justify your model and use appropriate statistical methods for estimating and quantifying uncertainty in associations.
Your final analysis should be presented in the form of a brief report (no more than 10 pages including relevant tables and figures). A font size of 11 points or larger must be used. Margins in all directions must be at least 1 inch. You may place additional information (e.g. diagnostic plots) in an Appendix if you feel it is necessary, however the Appendix should not be a mere copy-and-paste of computer code and computer output, nor it should be a collection of figures without captions.
The report should (at minimum) consist of the following sections:

- *Abstract* - A brief summary of your basic findings

- *Introduction* - A brief introduction/motivation to the problem at hand and what is to be addressed

- *Methods* - A discussion and justification of the statistical methods you have used to analyze the data and how you went about analyzing the data

- *Results* - A presentation and interpretation of the results of your analysis

- *Discussion* - A synopsis of your findings and any limitations your study may suffer from

Your report should appropriately cite ALL the software you used in the conduct of your analysis and in the drafting of your report. These citation should go into the Methods section of your report and should be specific about the purpose of each software tool utilized. For example, you might include a statement such as "All the statistical analysis provided were conducgted using `R` version

4.2 … (packages: `ggplot`, …, etc.). `Grammarly` was used for editing purposes; [..]". Please be exhaustive in listing all the software you have employed.

Your report should be well-written, succinct and to the point! It should be written in a language that is understandable to the scientific community while precisely interpreting your findings. Given that the purpose of this exam is to assess your statistical competency in analyzing data to address a given scientific goal, it is critical that: (1) the appropriateness of your modeling choices be clearly justified in your report; (2) the discussion of statistical methods be more technical than that provided to a non-statistical audience.

You may use tables, plots and figures to help explain your findings. You may use any written references that you wish, but you **cannot communicate (talk, email, etc.) with anyone about your analysis**.

On the cover page of your report, please write and sign the following statement:

"I attest that upon completion of this exam, I have destroyed the data file and I acknowledge that these data should not be used for any other purposes."

# 4    Data description

The dataset `DA2023.csv` contains a total of 20,059 observations on the following variables for 7,836 individuals :

| Variable Name | Description |
|---|---|
| id | Subject ID |
| physical | Physical function score. Range: 0-15. |
| discrim | Perceived discrimination score, assessed during first visit. Range: 0-9. |
| stress | Perceived stress score, assessed during first visit. Range: 0-18. |
| age | Age of subject (years), measured at each visit. |
| ed | Number of years of education, reported during first visit. |
| income | Self-reported income level, reported during first visit. Categorical variable with 10 classes: 1 through 10. Higher class indicates higher income level. |
| BMI | Subject's BMI at first visit |
| med | Self-reported number of medical conditions, reported during first visit. Five classes: 0 for "no known medical condition", 1 for "one known medical condition", 2 for "two known medical conditions", and so forth. The last class, 5 denotes "five or more known medical conditions". |
| male | 0: Female, 1: Male. |
| black | 0: Non-Hispanic White, 1: Non-Hispanic Black. |

You are allowed to use this dataset **only** for the purpose of this qualifying examination.

# 5 References

1. J. L. Bienias, L. A. Beckett, D. A. Bennett, R. S. Wilson, and D. A. Evans. (2003). Design of the Chicago Health and Aging Project (CHAP). *Journal of Alzheimer's Disease* **5(5)** 349-355.

2. J. M. Guralnik, E. M. Simonsick, L. Ferrucci, R. J. Glynn, L. F. Berkman, D. G. Blazer, P. A. Scherr, and R. B. Wallace. (1994). A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. *Journal of Gerontology* **49(2)**, M85-M94.

3. D. R. Williams, Y. Yu, J. S. Jackson, N. B. Anderson (1997). Racial differences in physical and mental health: socioeconomic status, stress, and discrimination. *Journal of Health Psychology* **2**: 335-351.