ECON 121, Applied Econometrics and Data Analysis **Prof. Tom Vogl**

Fall 2024 **University of California San Diego**

PROBLEM SET 3: SOCIOECONOMIC DETERMINANTS OF HEALTH

DUE BY 11:59 PM PDT ON THURSDAY 11/7

This problem set explores the socioeconomic correlates of health status in the United States. The dataset https://github.com/tomvogl/econ121/raw/main/data/nhis2010.rds contains a sample of adults from the 2010 National Health Interview Survey, with mortality follow-up to 2019. You will analyze two outcome variables: (1) mortality and (2) self-reported health status. Self-reported health status is based on the question: "On a scale of 1 (excellent) to 5 (poor), how would you rate your health?" The dataset contains many interesting covariates, including measures of socioeconomic status, race, health behaviors, and health conditions. You may notice that the dataset has sampling weights, but we will not use them here to keep matters simple.

To install R and *RStudio, follow this link. You are encouraged to work in a group of up to 4 members. You may write code together, but you must write verbal answers yourself. Please use a Markdown template for your code. Write verbal answers in the comments within the Markdown file, so that you produce a single PDF with code, results, and writing, which you will upload to Gradescope.*

1. List your group members.
2. Load packages and the dataset. Generate a binary variable that equals 1 if the respondent reports fair or poor health and 0 otherwise. Summarize the data and describe your findings in at most 4 sentences.
3. To get a sense of how self-reported health status relates to mortality risk over the lifecycle, compute mortality rates by age in two separate groups: (a) people who report being in fair-to-poor health and (b) people who report being in good-to-excellent health. Then draw separate line plots of the mortality-age relationship for the two groups in the same graph. How does the risk of death change with age? Do people with worse self-reported health status have higher risk of death? Answer in at most 4 sentences.
4. Draw four bar graphs to describe how average fair/poor health and mortality vary by socioeconomic status. In `ggplot()`, you should use `geom_bar(stat = "identity")`. For each of the four graphs, describe your results and take note of any unexpected patterns in 2-3 sentences.
   (a) Graph rates of mortality and fair/poor health by the level of family income.
   (b) Graph rates of mortality and fair/poor health by education level, with five categories of educational attainment: less than high school completion (<12), high school completion (12), some college (13-15), college completion (16), and post-graduate study (>16).

5. Age, income, education, and race/ethnicity are correlated, so we must use multiple regression to disentangle the relative importance of these variables in predicting health. For both mortality and fair/poor health, run linear probability models, probit models, and logit models with age, education, family income, and race/ethnicity as independent variables. Choose an appropriate functional form for age and education (linear, categorical, etc.), and be sure to motivate your choice in your written answer. (Remember that complicated functional forms are sometimes difficult to interpret, and interpretability is valuable. Sometimes it is useful to split a continuous variable into a series of dummy variables for different ranges.) For the probit and logit models, also compute the average marginal effects of the independent variables. For the logit model, further compute the odds ratios for the independent variables. Describe your results and take note of any expected or unexpected patterns. Are the LP, probit, and logit results similar? Answer in at most 5 sentences.
6. At the same age, who has a greater mortality risk? Is the difference in mortality risk statistically significant?

Group A: Asian adults with less than 12 years of education and family incomes less than $35k
Group B: Black adults with 16 years of education and family incomes over $100k

   Use your logit estimates from question (5) to answer. Do you think this model with no interaction terms is the best one for testing for differences between these groups? If not, how would you alter it? Answer in at most 4 sentences.
7. Should we think of the coefficients (or marginal effects or odds ratios) on family income as causal? Why or why not? Answer in at most 4 sentences.
8. Many wonder how much of the relationship between socioeconomic status and mortality reflects differences in health insurance or differences in health behaviors. Using a logit model and reporting odds ratios, explore how much of the mortality relationship these mediating variables can explain. Make sure you are able to interpret the results of the technique you use. Describe your results in at most 4 sentences.