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

Winter 2020 **University of California San Diego**

PROBLEM SET 2: SOCIOECONOMIC DETERMINANTS OF HEALTH

DUE BY 11:59 PM ON MONDAY 2/10

This problem set explores the socioeconomic correlates of health status in the United States. The dataset contains a sample of adults from the 2000 National Health Interview Survey, with 5-year mortality follow-up. You will analyze two outcome variables: (1) mortality in the five years after the survey 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.

1. Generate a binary variable that equals one if the respondent reports fair or poor health. Summarize the data.
2. Now use bar graphs to describe the relationship between socioeconomic variables and health. (In Stata, the command `graph bar yvar,over(xvar)` will plot averages of *yvar* across the categories of *xvar*.) You will need to generate new socioeconomic variables that contain the right categories. You need not disaggregate by sex, but you may if you so desire. For each graph, describe your results and take note of any unexpected patterns.
   (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).
3. Age, income, education, and race/ethnicity are correlated, so we must use multiple regression to disentangle the relative importance of these variables in determining health. For both 5-year 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 write-up. (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, compute the marginal effects of the independent variables. Describe your results and take note of any expected or unexpected patterns. Are the LP, probit, and logit results similar?

4. Holding all else equal, do high-income African-Americans have higher or lower mortality risk than low-income whites? Use your estimates from one of the models in question (3) to run this test. Do you think this regression specification is appropriate for testing for differences between high-income African-Americans and low-income whites? If not, how would you alter it?
5. Should we think of the coefficients (or marginal effects) on family income as causal? Why or why not?
6. Many wonder how much of the relationship between socioeconomic status and health reflects differences in health insurance or differences in health behaviors. Using *one* of the above models (LP, probit, or logit), explore the role of these mediating variables. Make sure you are able to interpret the coefficients of the technique you use.
7. When we recategorized health status as a binary variable, we may have thrown out useful information. Does the five category version of self-reported health status predict mortality? Is the relationship monotonic (i.e., consistent across all five categories), or does mortality rise only after self-reported health becomes fair or poor?
8. Use an ordered probit to estimate the full relationship between socioeconomic variables and health status, as in question (3). Are the results similar to the results based on the binary health status variable? Which set of results in question (3) provides coefficients on a comparable scale?
9. Use your estimates from question (8) to generate predicted probabilities of being in each health status categories. Plot the distribution of predicted probabilities for whites and for blacks. How do these distributions differ? How do they compare with the unadjusted histogram of self-reported health status for blacks and whites?