Economics 128 **Professors Mariam Salloum and Angela Vossmeyer**
Data Science **Fall 2017**

Assignment 5
Due 11/06/2017

One of the most debated topics in Congress lately is health insurance, which includes programs like Medicaid (a major provider of coverage for children). This assignment investigates the effect childhood health status has on obtaining health insurance as an adult. A code book and data set are online, which are described below:

1. *PSID.txt, PSID.pdf*: These files are taken from the Panel Study of Income Dynamics (PSID). The PSID conducts an annual survey of a random sample of Americans. These are the raw data files exported from the website. The *.txt* file includes the data. The *.pdf* file is the code book that defines each variable, how it was measured, and details the survey question.

To begin, you'll need to clean the data set: (1) remove observations in which the respondent did not know the answer, refused to answer, or skipped the question, (2) if the variable is binary, specify it as 0 or 1, (3) if the variable is ordered or categorical, discretize the variable by creating indicator variables for each category, (4) transform large numbers to make them usable.

Once you have cleaned the data, specify a logit model where the outcome variable (constructed from ER64612) is:

$$y_i = \begin{cases} 1 & \text{without health insurance} \\ 0 & \text{with health insurance,} \end{cases}$$

and the covarates include an intercept, sex, race, age, employment (ER60167), adult health status (ER62366), childhood health status (ER62506), income, and education. Be careful with how you specify the controls that are ordered or categorical variables.
Recall, for the the logit model (in matrix notation), we care about the:

$$\Pr(y_i = 1|x_i, \beta) = \Lambda(x_i'\beta)$$

where Λ is the logistic CDF. Estimation of the model is done by maximum likelihood. Please use the clean data and logit model to analyze the effect of childhood health status on being without health insurance.

Please submit the following:

- All code and final data
- A brief discussion of the model and estimation procedure
- Tables of regression results
- An analysis and discussion of the findings.

Be sure to zip-up the files before submitting online.