**Prediction Competition!**

Imagine you work for a large hospital as a data analyst. You have data on a representative sample of 100 past patients. The data set consists of 100 observations from patients admitted into an adult Intensive Care Unit (ICU) at an urban hospital. Our goal is to develop a model to predict whether a patient (a) has Medicare, and (b) will live long enough to be discharged from the hospital (recorded as died= 0) or die (recorded as died=1.) The data consists of a Case ID number in column A, the response variable live/die in column B, and other variables at time of admittance:

MEDICARE: Does Patient have Medicare?

AGE: Patient age

FEMALE: Gender (0 = male).

CANCER: Cancer history (1 = Yes)

SYS: Systolic blood pressure

HRA: Heart rate (beats/min)

PRE: Prev. recent ICU visit (0=No)

LOC: Consciousness status upon admittance … 0 is fully conscious, 1 is in condition of stupor, 2 is unconscious or in coma.

RACE: RACE/ETHNICITY (White, Black Hispanic)

HHINCOME: Household Income

**Part A. Predict Medicare**

1. In this activity, we will use the Medicare and Died variables as dependent variables. Take a look at the scatter plots and/or correlation matrix to check out the correlation between these and other available variables to help answer the following questions.
2. In real life, we tend to select variables for a regression equation based on theory/hypotheses and based on what we believe would affect the dependent variable. Select four variables that you believe would affect whether the patient have Medicare. Starting with these four variables, find a reasonable logistic regression model for whether the patient have Medicare and present it as an equation (can contain a subset of those four variables you chose, or if you want, you may also include interactions, dummies, quadratics, and/or logistic terms if you know how). You may re-iterate this model until it is to your satisfaction.
3. Compute the probability of Medicare for all 100 patients. Then, generate a variable that says how many predictions your model got “correct.” We’ll say it is “correct” if the probability is less than 0.5 and the Medicare variable is 0, or if the probability is greater than or equal to 0.5 and the Medicare variable is 1.
4. Write your name and anyone you worked with on a piece of paper, along with the model and the percent of predictions you got “correct.” Hand this piece of paper in. Prof Chan will use a test set to determine how many predictions your team got “correct.”
5. Ethically, how do you feel about predicting whether someone has insurance? For what purpose can you think of would this model be useful for?

**Part A. Predict Death**

1. Select another four variables for aspects you believe would affect whether the patient dies. Find another reasonable logistic regression model for whether the patient dies. Briefly explain which variables you chose to include in your model (can contain a subset of those four variables you chose, or if you want, you may also include interactions, dummies, quadratics, and/or logistic terms if you know how). You may re-iterate this model until it is to your satisfaction.
2. Compute the overall probability of death for all 100 patients. Then, generate a variable that says how many predictions your model got “correct.” We’ll say it is “correct” if the probability is less than 0.5 and the Died variable is 0, or if the probability is greater than or equal to 0.5 and the Died variable is 1.
3. Ethically, how do you feel about predicting whether someone will die? For what purpose can you think of would this model be useful for?