Ray Yu

Draft

**Research Question**

This paper focuses on the utility and risk evaluation methods of a clinical examination dataset. In the paper both logistic and linear regression are used to examine how categorical and continuous variables like age, race, amount cigarette smoked associates with BMI, body weight mass index.

**Background/Significance of Question**

The dataset that this paper uses comes from The Framingham Heart Study, which is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. The study began in 1948 and 5,209 subjects were initially enrolled in the study. Participants have been examined biennially since the inception of the study and all subjects are continuously followed through regular surveillance for cardiovascular outcomes. The original data has included cardiovascular disease risk factors and markers of disease such as blood pressure, blood chemistry, lung function, smoking history, health behaviors, ECG tracings, Echocardiography, and medication use. Among these variables, those thought to be related to body weight are selected. The final dataset used in the project includes age, smoking behavior, heart rate, education level and sex.

As obesity is more and more recognized as a important issue, there would be more people, researchers as well as intruders, looking at data that includes information related to body weight. We should protect individuals from getting their identity disclosed out of theses body weight related datasets.

**Methods**

The dataset used in the project has been processed in various ways.

First of all, variables related to body weight, which are mentioned in the previous section, are hand selected out. Also, “N/A”’s and “Unknown”’s is filtered out for all variables. For the categorical variables, dummy variables are created for each category in each variable.

The research then uses jags to generate posterior draws, and with which two separate synthetic BMI datasets are created. With the two synthetic datasets created, empirical cdf, utility evaluation and identity disclosure risk are conducted and studied.

The utility of the synthetic data was calculated with propensity score; while for risk evaluation expected match risk, true match risk and false match risk are all evaluated.