**Stats 506 Group 1 Project Proposal**

**Weijie Pan**

**Jingxian Chen**

**Eric Hernandez-Montenegro**

**Question**

**Do people’s eating habits have the same effect on their diabetes status with or without insurance?**

**Data**

**2015-2016 Demographic Variables and Sample Weights**

**Variables:**

**SEQN** - **Respondent sequence number,**

**RIAGENDR - Gender,**

**RIDAGEYR - Age (in years),**

**INDFMIN2 - Annual Family Income**

**2015-2016 Health Insurance**

**Variables:**

**SEQN - Respondent sequence number,**

**HIQ011 - Covered by health insurance**

**2015-2016 Diabetes**

**Variables:**

**SEQN - Respondent sequence number,**

**DIQ010 - Doctor told you have diabetes**

**2015-2016 Dietary Interview - Total Nutrient Intakes, First Day**

**2015-2016 Dietary Interview - Total Nutrient Intakes, Second Day**

**Variables:**

**WTDRD1 - Dietary day one sample weight**

**WTDR2D - Dietary two-day sample weight**

**DR1TIRON - Iron**

**DR1TCALC - Calcium**

**DR1TZINC - Zinc**

**DR1TSODI - Sodium**

**DR1TATOC - Vitamin E**

**DR1TVARA - Vitamin A**

**DR1TALCO - Alcohol**

**DR1TVC - Vitamin C**

**DR1TTFAT - Total fat**

**DR1TFIBE - Dietary fiber**

**DR1TSUGR - Total sugars**

**DR1TCARB - Carbohydrate**

**DR1TKCAL - Energy**

**DR1TPROT - Protein**

**Analytic Modeling Techniques**

**Use Lasso to select variables from the Dietary Interview. In order to consider the effect from the insurance, we build the logistic regression model adding interactive terms between insurance and variables in the total intakes dataset. Use the cross validation method. Then plot the ROC curve to display the performance of the model. Give a conclusion on whether the eating habits have the same effect on these two groups of people.**

**Model with insurance interaction term:**

**Where: insurance is 0 or 1. X are the variables that we choose in the total nutrient intake dataset.**

**Consider formula like this:**

1. **B1 \* X + B2 \* insurance \* X**
2. **B1 \* Insurance \* X + B2 \* (1 - insurance) \* X**

**So, if we choose to build model (1), we will use (B1 + B2) as the margin effect on people with insurance, and use B1 as the effect on people without insurance;**

**if we choose to build model (2), we will use B1 as the margin effect on people with insurance, and use B2 as the effect on people without insurance;**

**In our project, we would like to build a model like (2).**

**Software/Programming to Be Use**

**Python, R, and Stata**