**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**

**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 Lasso penalty in the model to select variables from the Dietary Interview.**

**Use the cross-validation method to choose the best penalty parameter of our model.**

**Because we have samples without diabetes weight more than samples with one, we choose to use the AUC value as our model performance measurement. 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**

**So, we will use (B1 + B2) as the margin effect coefficients on people with insurance, and use B1 coefficient as the effect on people without insurance;**

**Software/Programming to Be Use**

**Python, R, and Stata**

**Outline of our project**

1. **Data preprocessing part:**
2. **Select variables from each of the dataset mentioned above and combine all the dataset together.**
3. **First to remove observations with missing values;** 
   1. **transform age variable into factor format with 5 levels;**
   2. **transform income variable into factor format with 13 levels;**
   3. **standardize all the continuous intakes variables;**
   4. **add a variable called “survey day” denoting the observations from which dataset.**
   5. **Add interaction variables (intakes variables multiplied by insurance) and remove insurance variable.**
4. **Data visualization part: (only analyze day1 intakes data): (haven’t done yet)**

**Using plot or table to see the average values (using sample weights and give a 95%CI) of microelements (nutrients, vitamins) at each level of age gender and insurance status for those people who have diabetes problem;**

**Using plot or table to see the same measurements for people who don’t have diabetes problem;**

**Using plot or table to see the differences between the above two values.**

**Here, Microelements are zinc, iron, sodium, calcium; Nutrients are fat sugar carb protein; Vitamins are VA, VC, VE**

**After showing 9 tables or plots, we can give a general conclusion of the difference in eating habits of people with or without diabetes at each level.**

1. **Model establish part:**

**We decide to split 20% of our data as test data, and then we use the other 80% for training and cross-validation procedure (10 folds).**

**First to split our data; and then use cross validation method to choose the lambda which has the best AUC value in cross-validation dataset; then use this lambda to build our model with training set data; then predict the test set diabetes status and obtain AUC value from test set to see our model performance.**

**We then do the above procedure serval times to avoid overfitting problems by randomly choose different test data.**

**Assume we have done it N times, then we will get n different optimal lambda values and n AUC values for N models’ performances.**

**Then we use the mean value of those lambdas to build our final model for whole data and use the mean value of those AUC value as our estimation of the final model performance.**

1. **Margin effect part & model interpretation part: (haven’ t done yet)**

**We need to show the margin effect of people’ s eating habits on diabetes status with and without insurance. And conclude whether there are significant differences.**

**Besides, we also need to interpret our model for solving our main problem.**

1. **Other things can be improved: (haven’ t done yet)**
   * 1. **Parallel coding: for the cross-validation part and the randomly choose test data N times part we can use parallel skills to improve the efficiency of our program.**
     2. **Because we have approximately 15000 observations and only one tenth of then are having diabetes problem, that means our data is quite not balance. Besides using the AUC to measure model performance, to better handle this data, we can generate some new observations with diabetes problem using resampling skills like bootstrap to rebuild our model.**