**The Correlation of Obesity, Fast-food and Total Sugar Consumption**

ECON 400: Econometrics

Final Research Paper

Amy Hong

**I. INTRODUCTION**

Obesity is a major health problem in the United States. According to CDC report, the prevalence of obesity increased from 30.5% to 42.4%, and the prevalence of severe obesity increased from 4.7% to 9.2% from 1999 through 2018. There are several contributing factors for obesity. An overconsumption of food is the commonly known factor. Research published by the World Health Organization found that there is a correlation between increase in body mass index and increase in fast-food consumption. According to the research, fast food makes up about 11% of the average American diet, thus, it may be the big contributing factor to increase the rate of obesity.

For my final paper, I would like to investigate the correlation between obesity and fast food and high sugar consumption. My model will evaluate whether fast food and high sugar consumption have significant effect on body mass index (BMI). Individual with obesity has the high BMI. I expected to see a positive effect of fast food and high sugar consumption on BMI. More and higher consumption would lead to increase in BMI, so as the risk of having obesity.

Because food consumption is not the absolute contributing factor to obesity, the model includes demographic variables, such as gender, age, race, household income and level of education. According to CDC report, there is the association between obesity and socioeconomic status. Thus, including demographic variables to the model would allow to investigate additional contributing factors for obesity.

**II. DATA**

Dataset for the model was obtained from Kaggle, an online community of data scientists and machine learning practitioners. A user from Kaggle extracted a dataset from 2013-2014 National Health and Nutrition Examination Survey (NHANES), <https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey>. This dataset consists of six csv data files: survey on demographic, diet, examination, labs, medications and questionnaire. Among six files, I used demographic, diet and questionnaire data files, which include variables of interest for my model.

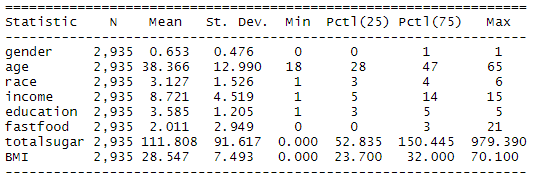
Prior to modeling, three data files were combined as one dataset by using unique identification number, SEQN, as a reference. Then, I cleaned this dataset to only have variables of the interest. The final dataset consists of 8 variables, one dependent variable and seven independent variables. Sample size of the final dataset is 2955 respondents to the NHANES survey.

**III. MODEL and METHOLOGY**

**A. Variables**

1. Independent variable: Body Mass Index (BMI)
2. Dependent variables:
   1. Gender: categorical variable, pregnant females were excluded from the final set
      1. Female as 0
      2. Male as 1
   2. Age: range from 18 to 65
   3. Race: categorical variable
      1. Hispanic as 1
      2. Non-Hispanic White as 3
      3. Non-Hispanic Black as 4
      4. Non-Hispanic Asian as 6
   4. Total household income (income):
      1. $0 to $4999 as 1, and increment of 1 for every $5000 increment until 10
      2. $75000 to $99999 as 14
      3. $100000 and above as 15
   5. Level of education (education)
      1. Less than 9th grade as 1
      2. 9-12th grade, but no high school diploma as 2
      3. High school diploma and GED as 3
      4. Some college or AA degree as 4
      5. College graduate or above as 5
   6. Number of fast foods as meal per week (fastfood): range from 0 per week to 21 per week
   7. Total sugar consumption per week (totalsugar): range up-to 1115.5gm per week

**B. Descriptive Statistics**



The sample size of the dataset is n = 2935, which consists of 1916 males and 1019 females. The mean values for each variables, excluding categorical variables (gender and race), indicates that the sample has the average: 1) age of 38 years old, 2) income of 8.7 (referring income range of $45000-$64999), 3) education of 3.5 (referring at least high school diploma to some colleges), 4) 3 meals as fastfood per week, 5) total sugar consumption of 112gm per week and 6) body mass index of 28.5.

**C. Modeling**

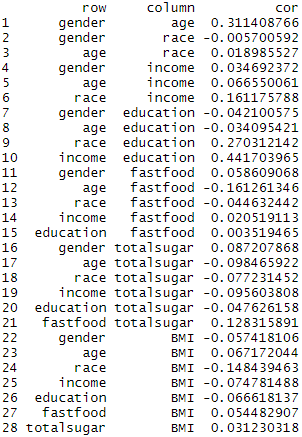
Prior to modeling, a correlation matrix of variables was created to check if two or more variables are highly correlated with one another. Highly correlated variables may affect the outcome of the model. It would be better to adjust highly correlated variables by either removing one variable with the lower correlation or creating a new variable by combining the correlated variables.

After the correlation check, OLS model was applied to dataset. First OLS model included all seven dependent variables. Then, the model was modified based on the significance of each variable coefficient. When the variable coefficient is significant at 95% confidence level, the coefficient was kept for the next round of OLS model evaluation; otherwise, variables were omitted from the model.

BMI = *c* + *b1*gender + *b2*age + *b3*race + *b4*income + *b5*education + *b6*fastfood + *b7*totalsugar

**IV. ANALYSIS and INTERPRETATION**

**A. Correlation Check**

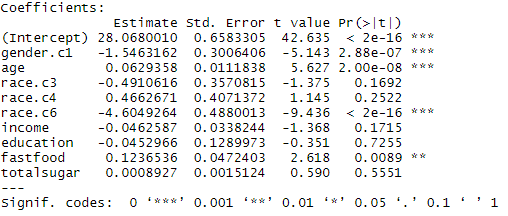
A screenshot of a cell phone

Description automatically generated

Some correlation between income and education was observed but not significantly strong to be concerned; thus, both income and education were included to the initial OLS model. Other than income and education, no apparent correlation between variables were detected.

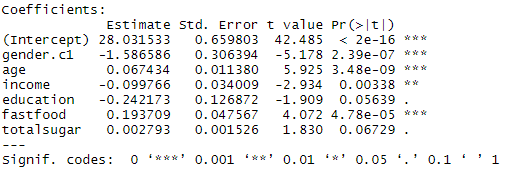
**B. OLS model**

a. First model: BMI = f{gender, age, race, income, education, fastfood, totalsugar}



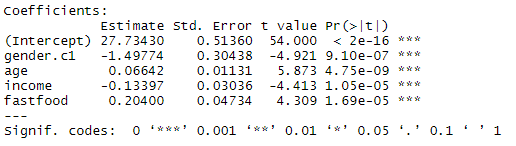
In the first OLS model, three independent variables (gender, age and fastfood) and one partial independent variable (race) are statistically significant at the 95% CI level. Because race variable was only partially significant, I decided to omit race to see if it would make other variable significant to the model. According to the first OLS model, race.c6, which is non-Hispanic Asian group, shows a negative effect on BMI. This indicates that Asian Americans tend to have lower BMI, which also may imply that the prevalence of obesity is relatively low in Asian Americans comparing to others.

b. Second model: BMI = f{gender, age, income, education, fastfood, totalsugar}



In the second OLS model, omitting race variable shows that other variables play more stronger role (or having more significant effect) on BMI. At 95% CI level, four variables (gender, age, income and fastfood) are statistically significant. Although education and total sugar are not statistically significant at 95% CI level, they are still significant at 90% CI level. The second OLS model indicates that gender, income and education have a negative effect on BMI; whereas, age, fastfood and total sugar consumption have a positive effect on BMI. To improve model further, variables that are not statistically significant at 95% CI level were omitted for the final model.

c. Final model: BMI = f{gender, age, income, fastfood}



In the final model, all variables are statistically significant at 95% CI level. The final model shows that gender and income have a negative effect on BMI, and age and fastfood have a positive effect on BMI. Based on model, male and individual with high income tend to have lower BMI than female and individual with low income. BMI increases as age gets older. More fast food consumption contributes to increase BMI.

**BMI** = 27.7343 - 1.4977**gender** + 0.0664**age** – 0.1334**income** + 0.2040**fastfood**

**V. CONCLUSION**

I predicted that high fastfood and total sugar consumption will increase the body mass index. The high body mass index, specifically above 30, is the indicator of obesity. The final OLS model indicates that fast food consumption has a positive effect on BMI. This result matches with the prediction I made in the beginning of the paper. As my prediction, model shows the BMI increase when more fast foods are consumed as meals. Total sugar consumption was not significant in the model at 95% CI level; it omitted from the final model. Thus, the final model does not tell if the high sugar consumption has an effect on increasing BMI. Three demographic variables were found significant in the final model in addition to fastfood variable. Gender and income have the negative effect on BMI. Male tends to have the smaller BMI than female. Higher income individuals tend to have the smaller BMI than lower income individuals. Age has the positive effect on BMI; older individuals tend to have the higher BMI than younger individuals. Although all variables are statistically significant in the final model, the low R-squared value, 0.024, indicates that the model only 2.4% of variability is explained by the model. In the future study, I would include more variables that may affect BMI. Including more variables and conduct modeling process may improve the model, which model would more explain variability.

**Reference**

1. <https://www.cdc.gov/obesity/data/adult.html>
2. <https://www.publichealth.org/public-awareness/obesity/>
3. <https://www.nbc26.com/lifestyle/fast-food-consumption-cdc-report>
4. <http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software>

**Stargazer Output**

Coefficient Test:



OLS models

