What are the culprits causing obesity? A machine learning approach in variable selection and parameter coefficient inference

Manhong Zhu
Scientist at PHS, zhumanhong@ufl.edu

Andrew Schmitz
Eminent Scholar and Professor, Food and Resource Economics Department, University of Florida, aschmitz@ufl.edu

Troy G. Schmitz
Associate Professor, Morrison School of Agribusiness, Arizona State University, Troy.Schmitz@asu.edu

Selected Paper prepared for presentation at the 2017 Agricultural & Applied Economics Association Annual Meeting, Chicago, Illinois, July 30-August 1

Introduction

The obesity epidemic has become a global health challenge, coupled with a controversy on what to blame for causing this problem. A rich body of literature argues that the major factor causing obesity is over consumption of caloric sweeteners, particularly from sugar-sweetened beverages (SSBs) (see review and meta-analysis studies such as Malik et al., 2013 and Imamura et al., 2015). However, there is also the argument that sugar overload alone does not cause obesity; other factors such as processed foods, food portion size, lack of physical activity, smoking habit, and genetics are also linked to obesity (NIH.org). Without a conclusive answer, government prevention policies, such as SSB taxes, are often criticized by the public.

When assessing the effect of certain variables on obesity, applied researchers typically choose only a small number of control variables allowing simple variable transformation and interaction based on economic intuition. These pre-specifications are intuitively appealing and generally do a good job when the data is low dimensional in that only a few characteristics per observation are available. However, when the data is high dimensional, such as data from the National Health and Nutrition Examination Surveys (NHNES) including dozens of predictors or characteristics for each observation, researchers rarely know which of the many control variables, and what combination of them, actually correspond to the dependent variable (e.g., Body Mass Index). Then, an ad hoc decision of model specification can often lead to incorrect conclusions, especially when the inference of regression coefficients is the objective of interest.

Machine learning (ML) methods, motivated partly by the availability of Big Data (large data sets), have been commonly used by retail firms, health care and internet industries to improve business decisions. In this paper, we propose the application of supervised ML methods to assess the main factors causing obesity in the United States. First, through clustering analysis of the control variables, we determine the significance level of control variables. Then, we apply supervised ML methods to allow the automated learning of the relationship between an arbitrary number of control variables and the dependent variable to determine the impact that significant control variables, or the interaction of these variables, have on obesity. This analysis will also provide insight on designing practical and plausible obesity prevention policies.

Methods and Data

We use two standard ML clustering methods, K-Means and the Birch clustering algorithm to discover the significant control variables. Then, we apply four supervised ML methods from statistics and computer science literature which generally yield considerably more accurate regression estimates than linear or logistic models in applied economics research (e.g., Bajari et al, 2015; and Belloni et al., 2014): Random Forest Regression (RFR), Support Vector Regression (SVR), Kernel Ridge Regression (KRR), and Neural Networks (NN). Using these four models, we conduct a cross-model validation to verify the regression results. All proposed models are available in the open software package R and Python machine learning libraries.

The proposed data set is composed of data from eight waves of NHANES which were publicly released in 2-year cycles from 1999-2000 to 2013-2014. NHANES is a national survey designed to assess the health and nutritional status of civilian noninstitutionalized U.S. population. These surveys collected dietary intake data in two 24-hour recall periods on consumer demographics, self-perceptions of health and nutrition status, and health related behaviors. About 12,000 persons per 2-year cycle were asked to participate in NHANES. Response rates varied by year, but an average of 10,000 participated in data collection. Variables that we are particularly interested in include sugar intake from different sources, fat intake, physical activity, smoking habit, frequency of dining out, which are all included in the survey data.

In more detail, we use the clustering algorithms on the NHANES data set control variables to split the data into several clusters. Then, we determine the significant control variables based on standard deviation for each variable in a data cluster, corresponding to the largest average BMI for instance. Finally, the four regression models are used to conduct parameter inferences of the significant control variables.

Expected Results

Each clustering algorithm is excepted to produce clusters that largely overlap corresponding to a specific level of BMI. We expect to find several control variables that are significant predictors of BMI. For instance, high caloric diets and low activity life styles are expected to be significant. Each of the four ML regression methods is expected to produce similar results on the impact of significant control variables, through which we can compare the impact of different control variables, sugar intake and fat intake for instance, on obesity (or BMI).

To be continued.

References

Imamura, F., O’Connor, L., Ye, Z., Mursu, J., Hayashino, Y., Bhupathiraju, S. N., & Forouhi, N. G. (2015). Consumption of sugar sweetened beverages, artificially sweetened beverages, and fruit juice and incidence of type 2 diabetes: systematic review, meta-analysis, and estimation of population attributable fraction.

Malik, V. S., Pan, A., Willett, W. C., & Hu, F. B. (2013). Sugar-sweetened beverages and weight gain in children and adults: a systematic review and meta-analysis. The American journal of clinical nutrition, 98(4), 1084-1102.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. The American Economic Review, 105(5), 481-485.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. The Journal of Economic Perspectives, 28(2), 29-50.