# Leveraging data mining techniques to understand drivers of obesity

R. Salehnejad[1], R. Allmendinger[1], Y.-W. Chen[1], M. Ali[1], A. Shahgholian[2], P. Yiapanis[3] and M. Mansur[4]

***Abstract*— Substantial research has been carried out to explain the effects of economic variables on obesity, typically considering only a few factors at a time, using parametric linear regression models. Recent studies have made a significant contribution by examining economic factors affecting body weight using the Behavioral Risk Factor Surveillance System data with 27 state-level variables for a period of 20 years (1990-2010). As elsewhere, the authors solely focus on individual effects of potential drivers of obesity than critical interactions among the drivers. We take some steps to extend the literature and gain a deeper understanding of the drivers of obesity. We employ state-of-the-art data mining techniques to uncover critical interactions that may exist among drivers of obesity in a data-driven manner. The state-of-the-art techniques reveal several complex interactions among economic and behavioral factors that contribute to the rise of obesity. Lower levels of obesity, measured by a body mass index (BMI), belong to female individuals who exercise outside work, enjoy higher levels of education and drink less alcohol. The highest level of obesity, in contrast, belongs to those who fail to exercise outside work, smoke regularly, consume more alcohol and come from lower income groups. These and other complementary results suggest that it is the joint complex interactions among various behavioral and economic factors that gives rise to obesity or lowers it; it is not simply the presence or absence of individual factors.**

## I. Introduction

Obesity has reached epidemic proportions globally with at least 2.8 million people dying each year as a result of being overweight or obese [11], worldwide obesity more than doubling since 1980, and almost 40% and 15% of adults aged 18 years and over being currently regarded overweight and obese, respectively [12]. Obesity can also lead to adverse health conditions, such as diabetes, cardiovascular diseases, musculoskeletal disorders and even cancer, as well as emotional issues and negative social experiences caused, for example, by weight discrimination, bullying and a lack of confidence [14], [13].

Formally, obesity is defined as a body mass index (BMI) of at least 30, and overweight as a BMI between 25 and 30. In turn, the BMI is calculated as the body weight (kg) divided by the square of the body height ($m^2$), or $\text{BMI\_}[kg/m^2] = (\text{body weight})/(\text{body height})^2$. This study focuses on adult obesity. The survey in [29] offers a thorough overview of childhood obesity.

In addition to being a public health concern, obesity has also become a public financial concern impacting the economy annually by an estimated $190 billion largely due to human capital, productivity, direct medical and transportation expenses [16], [15]. Luckily, obesity is preventable, and with the ultimate goal of tackling obesity, its impact has urged economists to examine whether obesity is an economic phenomenon involving individuals responses to incentives [17]. This has led to a number of theoretical (see e.g. [18], [19], [20], [27], [21], [26], [17]) and experimental studies (see e.g. [22], [23], [25], [24] in the past two decades aiming to understand the relationship between obesity, costs, and eating and living habits.

Theoretical studies on obesity considered various aspects, such as casting weight as a function of eating and exercise choices made via a utility-maximization process [19], and accounting for time costs of eating alongside monetary costs and how the costs are affected by innovations like microwaves and vacuum packing [18]. The impact of an intertemporal dimension on weight has also been investigated to account for the fact that eating in the present results in future health costs [20], [27], [21], [26], [17] with [17] being the most recent and comprehensive theoretical study conducted.

Findings from empirical studies are somehow contradictory. While several studies find a relationship between weight and personal financial resources [22], [25], the relationship fails to appear in [23], [24]. Several studies document a correlation between obesity and a variety of economic factors, such as costs of eating, state-level prices of grocery food, restaurant prevalence, cigarettes, and alcohol. For example, a negative association between food prices and obesity is found in [32]. A positive relationship between restaurant prevalence and BMI is found in [33], [34]. Higher cigarette prices has been found to be associated with higher obesity in [36].

Although, substantial research has been carried out to explain the effects of economic variables on obesity, typically considering only a few factors at a time, using parametric linear regression models. Courtemanche et al [17] have recently made a significant contribution by examining economic factors affecting body weight using the Behavioral Risk Factor Surveillance System data with 27 state-level variables for a period of 20 years (1990-2010). As elsewhere, the authors solely focus on individual effects of potential drivers of obesity than critical interactions among the drivers. We take some steps to extend the work of Courtemanche et al [17] to a gain a better understanding of the drivers of obesity. We employ state-of-the-art data mining techniques to uncover critical interactions that may exist among drivers of obesity in a data-driven manner.

[1]Alliance Manchester Business School, The University of Manchester
[2]School of Social Sciences, Business & Law, Teesside University
[3]School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University
[4]Institute of Statistical Research and Training, University of Dhaka

978-1-4673-8988-4/17/$31.00 

TABLE I
VARIABLE DEFINITIONS & SOURCES

| Variables | Definitions | Data Source |
|---|---|---|
| BMI | Body mass index | BRFSS |
| **Individual Characteristics** | | |
| Race | Race/ethnicity categories | BRFSS |
| Gender | Gender of respondent | BRFSS |
| Children | Number of children less than 18 years old live in household | BRFSS |
| Marital | The marital status | BRFSS |
| Age | Imputed Age value collapsed above 80 | BRFSS |
| **Individual Economic Variables** | | |
| Income | Annual household income from all sources | BRFSS |
| Employment | Are you currently employed? | BRFSS |
| RENTHOM | Own or Rent Home | BRFSS |
| **Behavioral Variables** | | |
| EXERANY | Physical activities or exercises in Past 30 Days | BRFSS |
| ALCDAY5 | Days in past 30 had alcoholic beverage | BRFSS |
| MENTHLTH | Number of Days Mental Health Not Good | BRFSS |
| **Regional Variables** | | |
| UnemptRate | Proportion of labor force unemployed | BLS |
| MedIncome | Median household income in $10,000s(2010)$ | BLS |
| AvgPricePack | Average price of pack of cigarettes (2010$) | Tax Burden on Tobacco |
| Avg_steak_price | Average price of steak | C2ER |
| Avg_grocery_items | A measure of the relative price level for groceries | C2ER |
| Restaurant | Restaurants per 10,000 residents | QCEW |
| Avg_beer_price | Average price of beer | C2ER |
| Avg_wine_price | Average price of wine | C2ER |
| Avg_pizza_price | Average price of Pizza | C2ER |
| Avg_chips_price | Average price of chips | C2ER |

## II. MODEL INPUT: DATA

This work has utilized data from various sources for the year 2014. These consist of the *Behavioral Risk Factor Surveillance System (BRFSS)*, the *Council For Community and Economic Research (C2ER)*, the *Quarterly Census of Employment and Wages (QCEW)*, the *Bureau of Labor Statistics (BLS)*, and *The Tax Burden on Tobacco* data sources.

The BRFSS dataset is based on telephone surveys of randomly selected residents of all fifty U.S. states and contains information about their health-related risk behaviors, chronic health conditions, and use of preventive services. In addition to a number of health related variables such as BMI, amount of physical activities undertaken, alcohol consumption, status of mental health, we also collect individual-level demographic information, such as age, gender, race and ethnicity, marital status, number of children in the household, and information regarding respondents' socio-economic status such as annual household income, employment status and home ownership. Figure 1 provides the distribution of the BMI across the patients in the dataset considered in this study.

The source of our price data is the Cost of Living Index (formerly known as the ACCRA Cost of Living Index) published by C2ER. We include a composite measure of the relative price levels for groceries and prices of food and beverage items widely consumed in the U.S., namely steak, pizza, chips, wine and beer by averaging over regional prices in each state.

The number of restaurants (per 10,000 residents), representing both fast food and full table service, is collected using data from QCEW.

State-level economic data such as unemployment rate and median household income come from BLS. Finally, cigarette prices (per pack) in different states come from *The Tax Burden on Tobacco*. These prices include both state and federal cigarette excise taxes. A full list of variables with descriptions and sources is presented in Table I. The sample includes 464664 observations. Figure 1 describes the distribution of the BMI measure in the sample. 271694 observations belong to female respondents and the remaining 192970 observations belong to male respondents. Average BMI in the two sub-samples are 27.64 (females) and 28.23 (males) respectively. The highest average BMI (29.43) belongs to Native Hawaiian or other Pacific Islanders and the lowest mean (24.82) belongs to the Asian race.

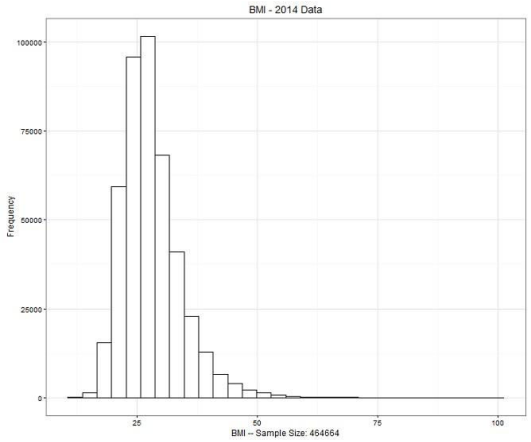


Fig. 1. BMI distribution across the respondents in the dataset considered in this study.

## III. LASSO AND REGRESSION TREE TECHNIQUES

Recent advances in statistics and data mining have led to the introduction of regression techniques that outperform traditional OLS (ordinary least squares) estimation in avoiding

potential overfitting and identifying variables that yield best out-of-sample prediction error. A class of such techniques is known as regularized regression estimators or, more briefly, LASSO (least absolute shrinkage and selection operator) [35]. The method fits a model containing *p* predictors using a technique that regularizes or *shrinks* coefficient estimates of predictively insignificant variables towards zero. LASSO proceeds by fitting a model similar to the OLS or logistic regression, with the difference that it adds a penalty term that shrinks coefficient estimates to zero and thus excluding them from estimation. Formally, the LASSO coefficients, $\hat{\beta}_{\lambda}^{L}$, minimizes the following quantity:

$$\arg_{\beta}\min \sum_{i=1}^{N}\left[y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_i\right]^2 \text{ s.t} \sum_{j=1}^{p}|\beta_{ij}|\leq t \quad (1)$$

or similarly expressed in the form of Lagrangian:

$$\arg_{\beta}\max \frac{1}{2}\sum_{i=1}^{N}\left[y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_i\right]+\lambda\sum_{j=1}^{p}|\beta_{ij}| \quad (2)$$

The term $\lambda\sum_{j=1}^{p}|\beta_{ij}|$ is the LASSO penalty term, also referred to as $L_1$ penalty term. The penalty term is the $L_1$ norm of the coefficient vector $\beta$, $||\beta||$. LASSO is fitted using cyclical coordinate descent algorithm which successively optimizes the objective function in (1) over each parameter with others fixed, and cycles repeatedly until convergence occur. The parameter $\lambda$ serves as the tuning parameter where $\lambda = 0$ corresponds to the full OLS or ordinary logistic model and the strength of the regularization increases as $\lambda \to \infty$. The optimal value of $\lambda$ is chosen using cross validation procedures. Because of the $L_1$ penalty term, LASSO shrinks coefficient estimates that are exactly zero thereby allowing us to filter predictors keeping only the important ones.

We employ LASSO to identify predictively significant variables and narrow down the list of predictors in our sample. The aim behind LASSO is to derive a set of coefficient estimates that minimizes out-of-sample prediction error. In a second step, we perform statistical significance testing for non-zero LASSO estimates of predictively important variables obtained in the first step (using the implementation of [31]).

Not only the presence of a factor may be relevant in causing obesity but interactions among possible drivers of obesity may also be equally critical. Interactions among critical predictors often affect the chance of obesity. Granting this, we are in need of a technique to identify how various drivers of obesity interact with each other to determine the chance of obesity. Traditional parametric regression methods do not offer a way to determine relevant interactions among explanatory variables (predictors) in a data-driven manner. We resort to a class of non-parametric techniques, known in the machine learning literature as regression trees, that serve the purpose well. The tree mechanism involves recursively partitioning the predictor space into a number of small regions based on simple rules and using the mean or median of the realized values of observations (e.g., obesity level) belonging to a region as the predicted value for a new observation that falls in that particular region. All the splitting decision rules, order of important predictors and their interactions are summarized in a visually intuitive way. The segmentation patterns showing up in the tree help identify potential interactions among explanatory variables, shedding light on how various drivers interact to lower or raise the chance of obesity. We employ the regression tree technique to learn about possible interactions among the potential obesity drivers that survive the LASSO predictor selection stage.

Regression tree techniques are often criticized for biased selection of variables which have many possible splits and missing values. In this paper, we opt to use a conditional inference framework proposed in [30]. The technique rectifies the problem of selection bias by choosing predictors for splitting based on a series of tests identifying statistically significant associations between the responses and predictors.

As any non–parametric estimator, regression trees are subject to over-fitting. To achieve an optimal trade-off between the bias and variance (over-fitting), we rely on the out of sample predictive accuracy of the tree models, estimated using cross validation. We select the regression tree model with the lowest prediction error to identify patterns among the variables that best fit the data.

It should be noted that there are several control parameters that determine the complexity of regression trees, including the number of tree layers and, in the case of the unbiased tree estimator, the significance level for selecting optimal cuts. We started with the default choices built in the unbiased regression tree estimator as our benchmarks and next examined the predictive accuracy of alternative trees arising from changing the number of tree layers or setting the significance level at 1%, 5% and 10% using 10-fold cross validation. Changing the control parameters does not give rise to noticeably different trees with better prediction accuracy scores.

To our knowledge, this is one of the very few studies in health studies that exploit regression trees with an aim to find drivers of obesity.

## IV. Empirical Analysis

We start by applying the Lasso technique to the data to identify factors that appear as predictively most significant and help understand drivers of obesity. We next use the predictively significant variables to construct a number of unbiased regression trees to study interactions among the predictors. We end the analysis by reporting the results of several robustness checks.

### *A. LASSO Analysis*

Table II reports the Lasso logistic regression results. The dependent variable is an indicator variable that takes value one if the BMI exceeds 30 and otherwise takes zero. Individuals with a BMI exceeding 30 is considered as overweight. Column (1) reports Lasso logistic estimates for the full set of variables entering our study. Columns (2) through (5) report

TABLE II

TABLE SHOWING LASSO LOGISTIC MODEL RESULTS. COLUMN (1) PRESENTS THE SIMPLE LASSO LOGISTIC RESULTS, SHOWING PREDICTIVELY SIGNIFICANT VARIABLES. THE DOTS INDICATE PREDICTIVELY INSIGNIFICANT VARIABLES. COLUMNS (2) THROUGH (5) REPORT LASSO LOGISTIC COEFFICIENT ESTIMATES FOR SUBSETS OF THE PREDICTIVELY SIGNIFICANT VARIABLES. THE DEPENDENT VARIABLE TAKES VALUE ONE IF THE BMI EXCEEDS 30 AND OTHERWISE TAKES ZERO. ALL RESULTS ARE CALCULATED USING R PACKAGES GLMNET AND HDM. STATISTICAL SIGNIFICANCE NOTATION: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

| | Lasso Model | Model I | Model II | Model III | Model IV |
|---|---|---|---|---|---|
| Intercept | −1.17 | | | | |
| Race | 0.024 | 0.019 (0.002)$^{***}$ | 0.019 (0.002)$^{***}$ | 0.015 (0.002)$^{***}$ | 0.028 (0.002)$^{***}$ |
| Gender | −0.109 | −0.044 (0.007)$^{***}$ | −0.028 (0.007)$^{***}$ | −0.061 (0.007)$^{***}$ | −0.069 (0.007)$^{***}$ |
| Children | . | . | . | . | . |
| Marital | −0.026 | −0.010 (0.002)$^{***}$ | −0.016 (0.002)$^{***}$ | −0.027 (0.002)$^{***}$ | −0.022 (0.002)$^{***}$ |
| Age | −0.001 | 0.000 (0.000) | 0.000 (0.000) | −0.001 (0.000) | −0.001 (0.000)$^{***}$ |
| Income | −0.005 | | −0.004 (0.000)$^{***}$ | −0.004 (0.000)$^{***}$ | −0.004 (0.000)$^{***}$ |
| Employment | . | | . | . | |
| RENTHOM | 0.022 | | 0.050 (0.004)$^{***}$ | 0.032 (0.005)$^{***}$ | 0.035 (0.005)$^{***}$ |
| EXERANY | 0.281 | | | 0.314 (0.008)$^{***}$ | 0.309 (0.008)$^{***}$ |
| ALCDAY5 | . | | | . | . |
| MENTHLTH | −0.003 | | | −0.003 (0.000)$^{***}$ | −0.002 (0.000)$^{***}$ |
| UnemptRate | . | | | | . |
| MedIncome | . | | | | . |
| AvgPricePack | 0.003 | | | | 0.004 (0.000)$^{***}$ |
| Avg_steak_price | 0.065 | | | | 0.062 (0.005)$^{***}$ |
| Avg_grocery_items | −0.008 | | | | −0.008 (0.001)$^{***}$ |
| Restaurant | . | | | | . |
| Avg_beer_price | −0.047 | | | | −0.081 (0.009)$^{***}$ |
| Avg_wine_price | . | | | | . |
| Avg_pizza_price | −0.024 | | | | 0.000 (0.008) |
| Avg_chips_price | 0.258 | | | | 0.203 (0.016)$^{***}$ |

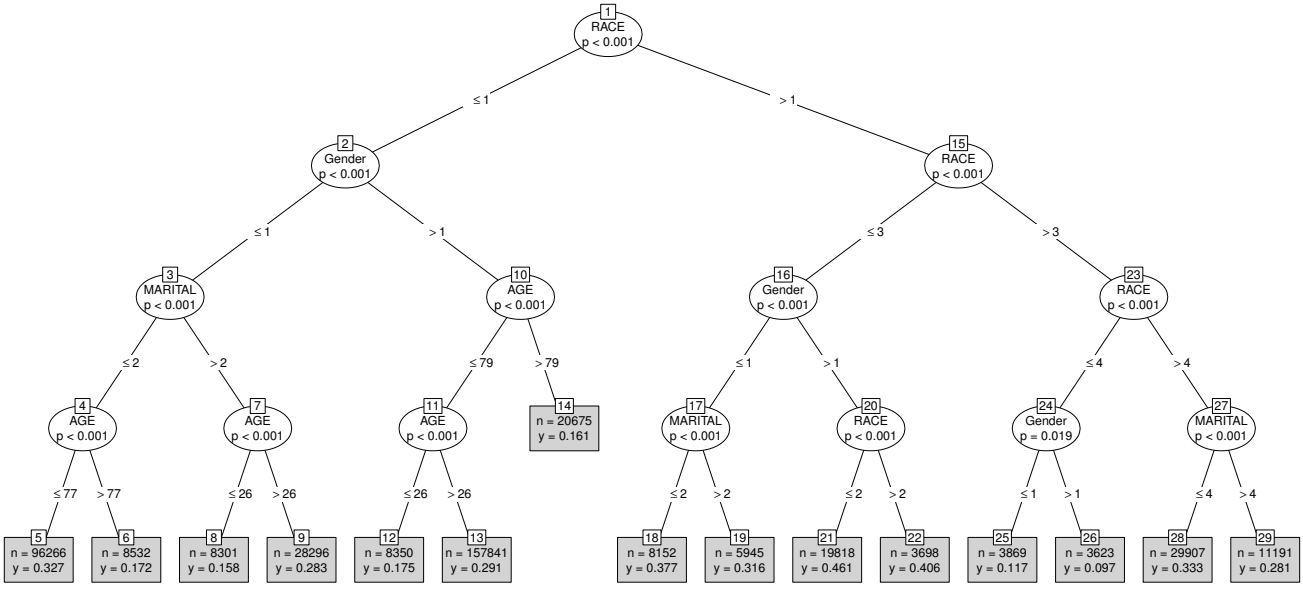


Fig. 2. Plot showing the unbiased tree applied to the predictively significant LASSO variables in column (2) of Table 2, where the dependent variable is an indicator variable that takes value one if the BMI exceeds 30 and otherwise takes zero. The variables appearing higher in the tree are predictively most significant. The final nodes state the frequency of obesity for the segment of the sample within which the observations fall. Not all the LASSO variables appear in the tree. Categorical variable Race takes 9 values, with 1 standing for white only, 2 for black only, 3 for American native or Alaskan native only, 4 for Asian only and so forth. Gender takes 1 for male respondents and 2 for female respondents. Marital takes value 1 for married respondent, 2 for divorced respondent, 3 for widowed, 4 for separated and 5 for never married.

the logistic regression estimates for four models where we introduce several categories of independent variables in steps. The introduction of the variables in steps enables us to better understand stylized facts present in the sample.

Lasso provides two options for selecting predictively significant variables. The first corresponds to the regularization parameter that yields minimum mean squared out-of-sample prediction error, calculated using cross validation. The sec-

ond corresponds to the regularization parameter that it is one standard deviation away from the regularization parameter minimizing the out-of-sample mean squared error, selecting a smaller number of predictors. We aim to understand factors that drive obesity. Since the cardinality of the variable set is not large, we opt for the first choice that yields a richer model with minimum mean squared error.

Lasso drops out predictively insignificant variables, shown by dots in Table II. Several variables appear as predictively insignificant. They are the number of children in the household (Children), whether the respondent is employed for wages (Employment), the frequency of alcoholic beverage during the past 30 days (ALCDAY5), the proportion of labor force unemployed in the region (UnemptRate), the regional median income in 2014 (MedIncome), the number of restaurants (Restaurant) and regional average wine prices (avgewine). We take the remaining predictively significant variables as potential drivers of obesity in our sample.

We use the output of the Lasso exercise to run a series of logistic regression models, reported in Table II. Column (2) reports the coefficient estimates for the predictively significant individual variables race-ethnicity (Race), gender (Gender), marital status (Marital) and age (Age). Race positively relates with obesity, gender (being female) and marital status (being married) negatively relate to obesity and age fails to be significant. Column (3) adds several individual economic variables to the variables in model 1. Of these variables, income level (Income) negatively relates to obesity whereas whether the respondent owns or rents her home (RENTHOM) relates positively to obesity - all significant at $0.01\%$ level. The frequency of obesity among those who rent their home are more likely to be higher. It is the real income status of the participant that significantly relates to the occurrence of obesity (also reflected in her home ownership status), not whether the person's employment status. Column (4) adds predictively significant behavioral variables that report whether the participant takes regular exercise (EXERANY) and mental conditions (MENTHLTH). The variables appear as significant at $0.01\%$ or below. Physical exercise negatively correlates with the incidence of obesity whereas beverage consumption positively correlates with obesity (the variable takes one for participants who report physical exercise in the past 30 days and higher values for those who do not report it). Similarly, mental conditions such as depression, stress and emotional challenges correlate positively with obesity (the variable takes lower values for participants who report mental health challenges and higher values who do not).

The last column adds predictively significant variables that relate to regional economic conditions or measure prices of typical goods that may contribute to obesity including regional average price of a cigarettes packet, average prices of steak, beer, wine, pizza and chips. Average price of beer negatively relates to obesity. Average prices of cigarettes, chips and steak positively relate to obesity, which may *prima facie* appear unintuitive. Inspecting the data, it transpires that the positive correlation between cigarettes prices and obesity occurs in the sub-sample of participants with comparatively lower incomes. Among those with comparatively higher incomes, there is a negative correlation between cigarettes prices and obesity. Low income in a region may drive up smoking and obesity, leading to the observed positive association. Similarly, the positive association between average prices of chips and steaks and obesity is highly stronger in the subset of individuals with comparatively low incomes. In the high income segment of the data, either these variables fail to appear as statistically significant or their economic significance is substantially weaker.

### *B. Unbiased Regression Tree Analysis*

It is plausible to conjecture that various drivers of obesity crucially interact with each other to cause obesity. Parametric regression techniques are not suitable for capturing possible interactions among the predictors in a data-driven manner. We borrow the unbiased regression tree technique from the machine learning literature to further shed light on stylized facts present in the data and capture possible interactions among the predictors. We apply the unbiased regression tree estimator to the predictively significant variables in Table II, starting with column (2), where we only include individual characteristics. Applying the tree estimator to the variables yields the tree model in Figure 2. Predictors that appear higher in the tree (i.e. earlier splits) or appear multiple times are predictively more significant than variables that occur in the lower layers in the tree. Variables that fail to appear in the tree are predictively insignificant. Race appears at the root node suggesting ethnicity as the most significant factor in predicting obesity. The variable segments the data into sub-samples of White (the left-hand side branch) and non-White (the right-hand side branch). On the left-hand branch, gender appears as the second most predictively significant variable - the variable further segments the sub-sample into white male (left hand branch) and white female (right hand side branch). The frequency of obesity among white male respondents is higher $(0.298)$ than the frequency of obesity among white female respondents $(0.271)$ Marital status and age emerge as the third most predictively significant variables. The lowest frequency of obesity belongs to Asians on the right-hand side branch. The frequency of obesity among Asian male respondents is $(0.117)$ whereas among the female Asian is $(0.097)$. The highest frequency of obesity occurs among black female respondents $(0.461)$. It is, for example, the interaction of ethnicity and gender that may raise or lower obesity rather than each factor individually.

The tree in Figure 3 corresponds to the predictively significant variables in column (3) where we add individual economic variables. Individual economic variables forces out variable Race from the root of the tree. Income appears as predictively most significant variable and is negatively correlated with obesity. The frequency of obesity among those with an income exceeding \$75,000 is $0.251$ while among those with an income level below \$75,000 is $0.432$. Gender, marital status and race appear in the second and third layers of the tree as predictively significant. The highest

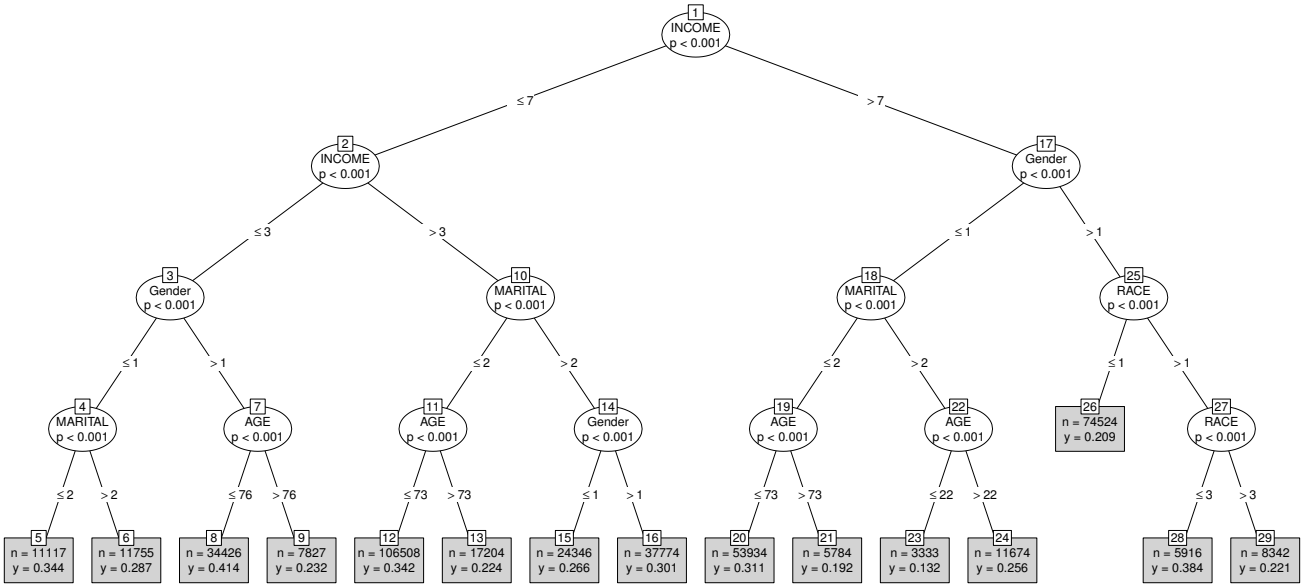


Fig. 3. Plot showing the unbiased tree applied to the predictively significant LASSO variables in column (3) of Table 2, where the dependent variable is an indicator variable that takes value one if the BMI exceeds 30 and otherwise takes zero. The variables appearing higher in the tree are predictively most significant. The final nodes state the frequency of obesity for the segment of the sample within which the observations fall. Not all the LASSO variables appear in the tree. The description of variables age, gender, marital and race are as before. Variable Income takes 10 values. The variable takes value 1 when the respondent income is less than $10,000, 2 when it lies between $10,000 and $15,000, 3 when income lies between $15,000 and $20,000, 8 for incomes of $75,000 or above. The last two values of the variable represent cases where respondents replied Not sure or refused answering the question.

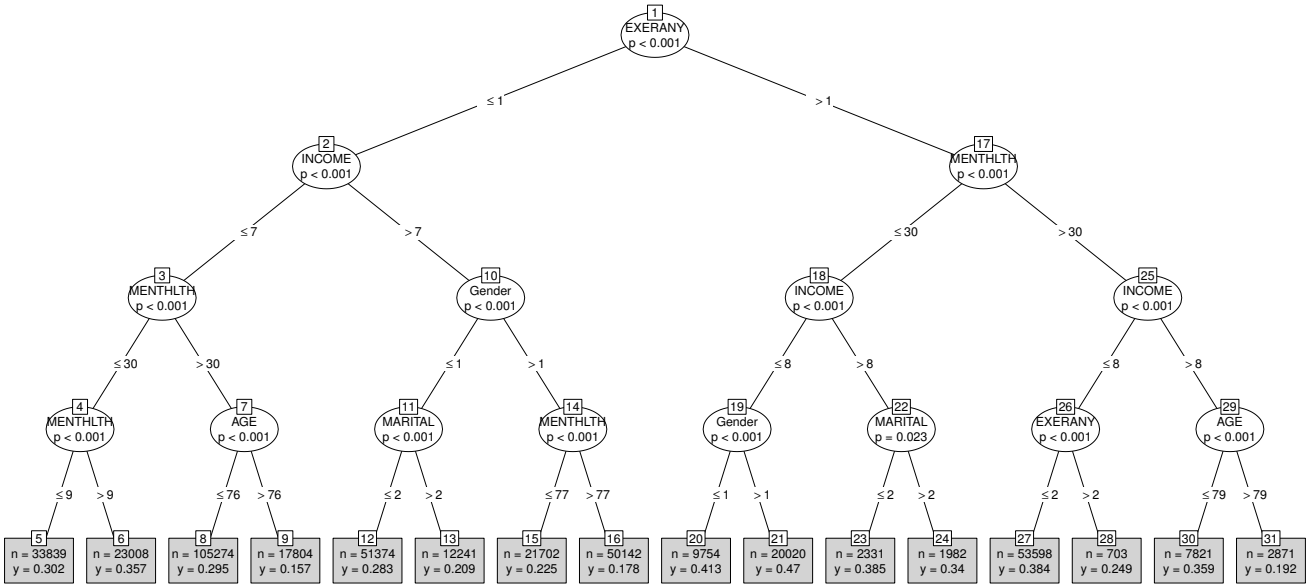


Fig. 4. Plot showing the unbiased tree applied to the predictively significant LASSO variables in column (4) of Table 2, where the dependent variable is an indicator variable that takes value one if the BMI exceeds 30 and otherwise takes zero. The variables appearing higher in the tree are predictively most significant. The final nodes state the frequency of obesity for the segment of the sample within which the observations fall. Not all the LASSO variables appear in the tree. There are two new variables in the tree: Exercise (EXERANY) and mental health (MENTHLTH). Exercise takes four values: 1 for those who report regular physical activity in the past 30 days, 2 for those who report no physical activity and 3 and above for those who state “Not sure” or refuse to give any response. Values between 1 and 30 for MENTHLTH measure the number of days in the past 30 days during which the respondent did not experience good mental health and values above 30 represent cases where the respondent states “None”, “Not Sure” or refused to provide any response.

frequency of obesity $(0.414)$ belongs to female respondents with an annual income of \$20,000 or less, who are not older than 76 years.

The tree in Figure 4 adds behavioral variables exercise (EXERANY) and mental health (MENTHLTH), corresponding to the third model in column (4). Exercise segments the data into sub-sample of those who report regular physical exercise in the past 30 days (left-hand side branch) and those who report no exercise (right-hand side branch). The frequency of obesity among those who report exercise is $0.264$ whereas it is $0.349$ among those who report no exercise, suggesting that regular exercise is correlated with lower obesity. The highest frequency of obesity $(0.47)$ belongs to female respondents with lower income, who do not report any exercise in the past 30 days and suffer from stress, depression or other emotional problems. The lowest frequency of obesity occurs among respondents who report exercise, do not suffer from mental health, their income is below \$75, 000 and are more than 76 years old. Mental health is positively correlated with obesity.

Figure 5 applies the regression tree estimator to the full set of predictively significant variables in column (5), where we include variables that either capture regional economic conditions or the price of goods that may contribute to the rise of obesity. As in the previous model, exercise appears in the initial node as predictively most significant, and variables income and mental health status appear in the second layer as predictively significant. The additional regional economic variables fails to appear in the tree significantly. Only the average price of cigarettes appears in the lowest layer as predictively significant. The tree reveals important interactions among potential drivers of obesity. The interaction of high income (above \$75, 000) and regular physical exercise leads to lower frequency of obesity $0.228$ compared to $0.291$ for those who report exercise but enjoy a lower income. Or, the frequency of obesity among those report no exercise but suffer form mental problems $(0.441)$ is greater than the frequency of obesity among those who do not report physical exercise or any mental problems $(0.371)$. Beyond identifying individual drivers of obesity, any effort at understanding causes of obesity will require paying attention to complex interactions that may exist among factors affecting obesity. The non-parametric regression tree technique points to possible interactions among predictively significant drivers of obesity by segmenting the data in a data-driven manner.

In a next step to complete the analysis, we build on the four basic models in Table (1) by adding to each model the first and second order interactions found in the corresponding regression tree, where first–order interactions refer to interactions of the variable appearing in the initial node of the regression tree and those appearing in the second layer. Similarly, second–order interactions refer to interactions between variables appearing in the initial node, the second layer and third layer of the unbiased tree. We transform variables Gender, Marital Status, Race, Mental Health and Exercise into factor variables that take values one and zero. Table (3) presents the results. The majority of interactions present in the regression trees appear as statistically significant. Starting with Model I, the coefficients for interaction terms between Race and Gender and Race and Marital status are significant at the $1\%$ significance level. The coefficient of dummy variable Race is $-0.473$, meaning that being a non-Hispanic white reduces the odds of obesity by $38\%$. The coefficient of Gender is $-0.239$; being male reduces the odds of obesity by $21\%$. The coefficient of the interaction term between Race and Gender is $(0.380)$. The odds of obesity among non-Hispanic white males is $46\%$ greater relative to the non-Hispanic white female group. Similarly the coefficient of the interaction term between Race and Marital Status is $(0.071)$. The odds of obesity among married non-Hispanic white is $7\%$ greater relative to the non-married, non-Hispanic white. A similar interpretation applies to other statistically significant interactions. The interaction terms between gender and marital status in the second column, and gender and exercise in the third and forth columns are all statistically significant at $1\%$, with a positive sign. The interactions between mental health and exercise in columns (3) and (4) are significant at $5\%$ or below, with a negative sign. An empirically adequate understanding of drivers of obesity cannot overlook complex interactions that may exist among drives of obesity. The unbiased regression tree estimator, as we have seen, points to statistically significant interactions in an entirely data-driven manner. This is particularly important in contexts where we deal with a large number of predictors and the theory is weak to point to key interactions. Both AIC and BIC, goodness of fit measures, suggest selecting Model (4) that yields lower values for both indices.

## V. Conclusion

The empirical literature on the drivers of obesity is yet to pay adequate attention to critical interactions that may exist among factors causing obesity. The statistical significance of potential variables is studied using parametric models that fail to reveal significant interactions systematically. This research takes initial steps towards understanding interactions among drivers of obesity. We find, for example, that the interaction of high income (above \$75, 000) and regular physical exercise leads to lower frequency of obesity $0.228$ compared to $0.291$ for those who report exercise but enjoy a lower income. Or, the frequency of obesity among those who do not report physical exercise, experience mental health challenges and have comparatively lower income is higher than other groups in the society $(0.452)$. We also built traditional regression models to estimate the statistical significance of some of the interaction terms. Our approach is not entirely free of limitations some of which include:

- *Variable selection:* Designed for predictor selection, Lasso might remove individual variables that might otherwise be predictively significant when they are interacted with other variables. In that case, the subsequent regression tree cannot pull back the removed variables. The limitation is not unique to Lasso. Other predictor selection techniques, e.g., stepwise regression, suffer from the same limitation. Yet, unlike traditional

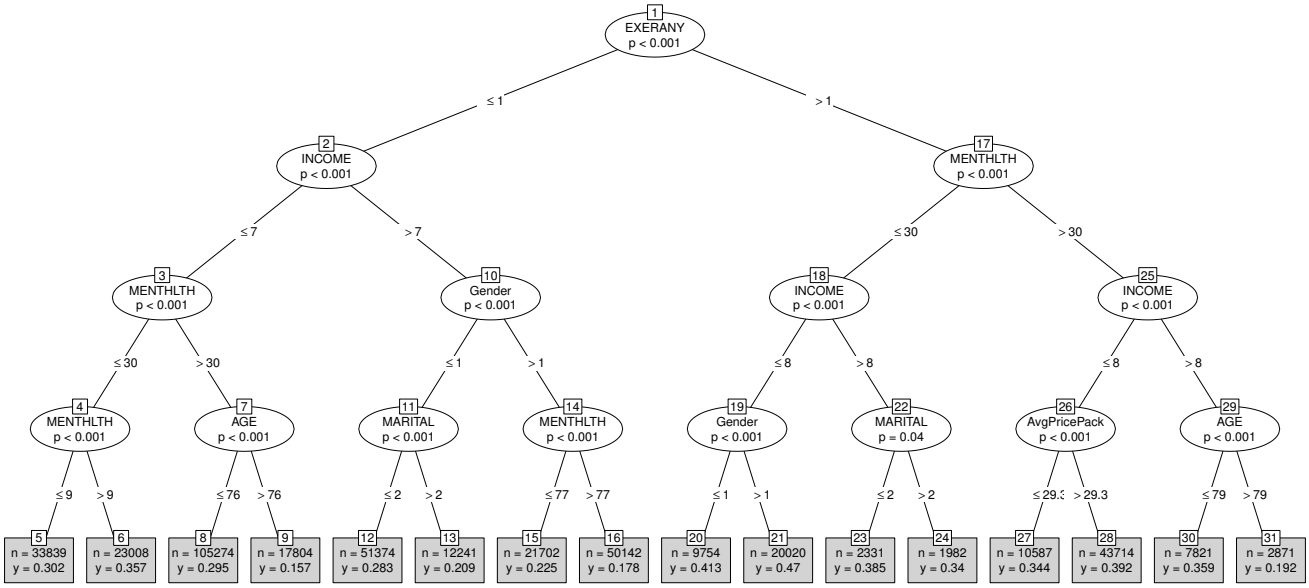


Fig. 5. Plot showing the unbiased tree applied to the predictively significant LASSO variables in column (5) of Table 2, where the dependent variable is an indicator variable that takes value one if the BMI exceeds 30 and otherwise takes zero. The variables appearing higher in the tree are predictively most significant. The final nodes state the frequency of obesity for the segment of the sample within which the observations fall. Not all the LASSO variables appear in the tree. The only new variable in the tree is Average Price Pack (AvgPricePack) that gives the average price of a pack of cigarette in 2000 prices.

TABLE III

LOGISTIC REGRESSION MODELS WITH INTERACTION TERMS. MODELS (I) THROUGH (IV) CORRESPOND TO THE LASSO MODELS IN TABLE 2. EACH MODEL ADDS THE FIRST ORDER AND SECOND ORDER INTERACTIONS PRESENT FOUND IN THE CORRESPONDING UNBIASED TREE TO THE VARIABLES TO EACH MODEL IN TABLE 2. COLUMN (1), FOR EXAMPLE, ADDS THE INTERACTIONS BETWEEN RACE AND GENDER AND BETWEEN RACE AND MARITAL STATUS TO THE PREDICTIVELY SIGNIFICANT VARIABLES IN COLUMN (1) IN TABLE 2. WE TRANSFORM VARIABLES GENDER, MARITAL STATUS, RACE, MENTAL HEALTH AND EXERCISE INTO FACTOR VARIABLES. GENDER TAKES VALUE ONE FOR MALE; MARITAL TAKES VALUE ONE FOR MARRIED; RACE TAKES VALUE ONE FOR NON-HISPANIC WHITE; MENTAL HEALTH TAKES ONE FOR REPORTING MENTAL HEALTH; AND EXERCISE TAKES VALUE ONE FOR REPORTING EXERCISE AND ZERO OTHERWISE. STATISTICAL SIGNIFICANCE NOTATION: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

| | Model I | Model II | Model III | Model IV |
|---|---|---|---|---|
| (Intercept) | $-0.701$ $(0.014)^{***}$ | $-0.765$ $(0.016)^{***}$ | $-0.469$ $(0.019)^{***}$ | $-0.118$ $(0.088)$ |
| Race | $-0.473$ $(0.012)^{***}$ | $-0.267$ $(0.008)^{***}$ | $-0.240$ $(0.008)^{***}$ | $-0.298$ $(0.009)^{***}$ |
| Gender | $-0.239$ $(0.014)^{***}$ | $-0.180$ $(0.012)^{***}$ | $-0.134$ $(0.015)^{***}$ | $-0.133$ $(0.015)^{***}$ |
| Marital Status | $-0.048$ $(0.014)^{***}$ | $-0.121$ $(0.010)^{***}$ | $0.079$ $(0.007)^{***}$ | $0.079$ $(0.007)^{***}$ |
| Age | $0.035$ $(0.002)^{***}$ | $0.037$ $(0.002)^{***}$ | $0.033$ $(0.002)^{***}$ | $0.035$ $(0.002)^{***}$ |
| Income | | $-0.006$ $(0.000)^{***}$ | $-0.005$ $(0.000)^{***}$ | $-0.005$ $(0.000)^{***}$ |
| Rent Home | | $0.043$ $(0.004)^{***}$ | $0.024$ $(0.004)^{***}$ | $0.028$ $(0.004)^{***}$ |
| Mental Health | | | $0.284$ $(0.013)^{***}$ | $0.281$ $(0.014)^{***}$ |
| Exercise | | | $-0.644$ $(0.013)^{***}$ | $-0.645$ $(0.013)^{***}$ |
| Avg_Cigarette_Price | | | | $-0.049$ $(0.003)^{***}$ |
| Avg_Steak_Price | | | | $0.065$ $(0.006)^{***}$ |
| Avg_Grocery_Price | | | | $-0.182$ $(0.017)^{***}$ |
| Avg_Beer_Price | | | | $-0.070$ $(0.008)^{***}$ |
| Avg_Pizza_Price | | | | $-0.068$ $(0.005)^{***}$ |
| Avg_Chips_Price | | | | $0.220$ $(0.016)^{***}$ |
| Race:Gender | $0.380$ $(0.016)^{***}$ | | | |
| Race:Marital Status | $0.071$ $(0.016)^{***}$ | | | |
| Gender:Income | | $0.003$ $(0.000)^{***}$ | $0.003$ $(0.000)^{***}$ | $0.003$ $(0.000)^{***}$ |
| Marital Status:Income | | $0.000$ $(0.000)$ | | |
| Gender:Marital Status | | $0.320$ $(0.016)^{***}$ | | |
| Gender:Marital Status:income | | $-0.000$ $(0.001)$ | | |
| Income:Exercise | | | $-0.000$ $(0.000)$ | $-0.000$ $(0.000)$ |
| Mental Health:Exercise | | | $-0.042$ $(0.016)^{**}$ | $-0.038$ $(0.016)^{*}$ |
| Gender:Exercise | | | $0.235$ $(0.018)^{***}$ | $0.236$ $(0.018)^{***}$ |
| Gender:Income:Exercise | | | $-0.000$ $(0.001)$ | $-0.000$ $(0.001)$ |
| AIC | 524728.913 | 520005.231 | 512577.157 | 501987.340 |
| BIC | 524805.791 | 520125.965 | 512730.798 | 502206.453 |
| Log Likelihood | $-262357.456$ | $-259991.616$ | $-256274.579$ | $-250973.670$ |
| Deviance | 524714.913 | 519983.231 | 512549.157 | 501947.340 |
| Num. obs. | 434814 | 431850 | 431198 | 423210 |

predictor selection techniques, Lasso is not subject to overfitting. With factor variables, one can include all combinations of the dummy variables and their interactions with non-factor variables in a Lasso regression to take an important step in uncovering predictively significant interactions. In such a setting, the regression tree technique complements Lasso. Further, mixing heterogeneous distributions can create spurious correlations or independences. It is theoretically possible that Lasso remove predictors that are genuinely related to the dependent variable or keep variables that are spuriously related. Such possibilities suggest that there are no entirely data-driven techniques for explanatory variable selection. One needs to support statistical analysis with background subject-matter information. The emerging data mining causal inference techniques [37] offer a promising avenue to extend the results. The methods also pave the way for analyzing wide samples that include a large number of predictors, enabling us to make progress in better controlling for potential confounders.

- *Model robustness:* A further question to be addressed in future is how robust the models obtained are for small perturbations in the data. This question is of importance because the data considered was obtained from surveys and hence need to be considered with caution.
- *Data completeness:* Capturing all the relevant data to explain drivers of obesity is arguably a very difficult (if not impossible) task. An obvious factor not captured in the survey data is the history of obesity and other health illness episodes in the family of a survey participant. It is known that obesity can be correlated to the family health history, thus including this factor in the dataset could affect the results obtained.

## ACKNOWLEDGMENT

The authors thank the Alliance Manchester Big Data Forum for facilitating this research.

## REFERENCES

[1] R. E. Bank and R. K. Smith, *General sparse elimination requires no permanent integer storage*, SIAM J. Sci. Stat. Comput., 8 (1987), pp. 574–584.

[2] S. C. Eisenstat, M. C. Gursky, M. Schultz, and A. Sherman, *Algorithms and data structures for sparse symmetric gaussian elimination*, SIAM J. Sci. Stat. Comput., 2 (1982), pp. 225–237.

[3] A. George and J. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice Hall, Englewood Cliffs, NJ, 1981.

[4] K. H. Law and S. J. Fenves, *A node addition model for symbolic factorization*, ACM TOMS, 12 (1986), pp. 37–50.

[5] J. W. H. Liu, *A compact row storage scheme for cholesky factors using elimination trees*, ACM TOMS, 12 (1986), pp. 127–148.

[6] J. W. H. Liu, *The role of elimination trees in sparse factorization*, Tech. Report CS-87-12, Department of Computer Science, York University, Ontario, Canada, 1987.

[7] D. J. Rose, *A Graph-Theoretic Study of the Numerical Solution of Sparse Positive Definite Systems of Linear Equations*, Graph Theory and Computing (1973), pp. 183–217.

[8] D. J. Rose, R. E. Tarjan, and G. S. Lueker, *Algorithmic aspects of vertex elimination on graphs*, SIAM J. Comput., 5 (1976), pp. 226–283.

[9] D. J. Rose and G. F. Whitten, *A recursive analysis of disection strategies*, in Sparse Matrix Computations (1976), pp. 59–84.

[10] R. Schrieber, *A new implementation of sparse gaussian elimination*, ACM TOMS, 8 (1982), pp. 256–276.

[11] World Health Organisation, *10 facts on obesity*, http://www.who.int/features/factfiles/obesity/en/, (accessed October 5, 2016).

[12] World Health Organisation, *Obesity and overweight*, http://www.who.int/mediacentre/factsheets/fs311/en/, (accessed October 5, 2016).

[13] R. Puhl and C. Heuer, *The Stigma of Obesity: A Review and Update*, Obesity, 17 (2009), pp. 941–964.

[14] J. Sallade, *A comparison of the psychological adjustment of obese vs. nonobese children*, Journal of Psychosomatic Research, 17 (1973), pp. 89–96.

[15] R. A. Hammond and R. Levine, *The economic impact of obesity in the United States*, Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, 3 (2010), pp. 285–95.

[16] E. Finkelstein, I. Fiebelkorn, and G. Wang, *National Medical Spending Attributable to Overweight and Obesity: How Much, and Who's Paying?*, Health Affairs, 2003, pp. W219-W226.

[17] C. J. Courtemanche, J. C. Pinkston, C. J. Ruhm, and G. L. Wehby, *Can Changing Economic Factors Explain the Rise in Obesity?*, Southern Economic Journal, 82 (2016), pp. 1266–1310.

[18] D. Cutler, E. Glaeser, and J. Shapiro, *Why Have Americans Become More Obese?*, Journal of Economic Perspectives, 17 (2003), pp. 93–118.

[19] T. Philipson and R. Posner, *The Long-Run Growth in Obesity as a Function of Technological Change*, Perspectives in Biology and Medicine, 46 (2003), pp. S87–S107.

[20] J. Komlos, *Obesity and the Rate of Time Preference: Is there a Connection?*, Journal of Biosocial Sciences, 36 (2004), pp. 209–219.

[21] C. Ruhm, *Understanding Overeating and Obesity*, Journal of Health Economics, 31 (2012), pp. 781–796.

[22] D. Lakdawalla and T. Philipson, *The Growth of Obesity and Technological Change: A Theoretical and Empirical Investigation*, National Bureau of Economic Research Working Paper #8965, (2002).

[23] M. Lindahl, *Estimating the Effect of Income on Health and Mortality Using Lottery Prizes as an Exogenous Source of Variation in Income*, Journal of Human Resources, 40 (2005), pp. 144–168.

[24] J. Cawley, J. Moran, and K. Simon, *The Impact of Income on the Weight of Elderly Americans*, Health Economics, 19 (2010), pp. 979–993.

[25] M. Schmeiser, *Expanding Wallets and Waistlines: The Impact of Family Income on the BMI of Women and Men Eligible for the Earned Income Tax Credit*, Health Economics, 18 (2009), pp. 1277–1294.

[26] E. Finkelstein, O. Khavjou, H. Thompson, G. Trogdon, L. Pan, B. Sherry, and W. Dietz, *Obesity and Severe Obesity Forecasts through 2030*, American Journal of Preventive Medicine, 42 (2012), pp. 563–570.

[27] C. Baum and S. Chou, *The Socio-Economic Causes of Obesity*, National Bureau of Economic Research Working Paper #17423, 2011.

[28] D. Miljkovic, W. Nganje and H. de Chastenet, *Economic factors affecting the increase in obesity in the United States: Differential response to price*, Food Policy, 33 (2008) 48-60.

[29] P. Anderson and K. Butcher, *Childhood Obesity: Trends and Potential Causes*, The Future of Children, 16 (2006), 19-45.

[30] T. Hothorn, K. Hornik, and A. Zeileis, *Unbiased recursive partitioning: A conditional inference framework*, Journal of Computational and Graphical statistics, 15 (2006) 651–674.

[31] V. Chernozhukov, C. Hansen, M. Spindler, *High-Dimensional Metrics in R*, (2016), available at https://arxiv.org/abs/1603.01700

[32] C.J. Courtemanche, G. Heutel, and P. McAlvanah, *Impatience, Incentives, and Obesity*,The Economic Journal, 125(2015) 1-31.

[33] R. Dunn, *Obesity and the Availability of Fast-Food: An Analysis by Gender, Race/Ethnicity and Residential Location*,American Journal of Agricultural Economics, 92(2010) 1149–1164.

[34] J. Currie, S., DellaVigna, E. Moretti, and V. Pathania, *The Effect of Fast Food Restaurants on Obesity and Weight Gain*, American Economic Journal: Economic Policy, 2(2010) 32–63.

[35] R. Tibshirani, *Regression shrinkage and selection via the lasso.*, Journal of the Royal Statistical Society B, 58( (1996), 267–88.

[36] C. Baum, *The Effects of Cigarette Costs on BMI and Obesity*, Health Economics, 18 (2009) 3–19.

[37] S. Athey and G. Imbens, *Recursive partitioning for heterogeneous causal effects*, Proceedings of the National Academy of Sciences, 113 (2016), 7353–7360.