Physical activity is as tied to obesity as we expect it to be*

An analysis of American exerise habits and their BMI

Victor Ma

March 16, 2024

1 Introduction

In the context of public health crises, the term 'pandemic' comes with the connotation of infectious diseases sweeping across global populations. Yet, the United States finds itself grappling with a pandemic of a different nature, with similarly far-reaching consequences: obesity. It is no secret that obesity comes with a multitude of direct health impacts- including increased risk of stroke, high blood pressure, type 2 diabetes, and even mental health problems like clinical depression and anxiety ("Health Effects of Overweight and Obesity" 2022). Obesity is recognized as a chronic complex disease defined by excessive fat deposits that can impair health. In particular, individuals with a Body Mass Index (BMI) of 30 or above are considered obese.

America's problem with obesity is long-lasting, with a 30.5% rate of obesity observed in the early 2000s (Centers for Disease Control and Prevention 2021). Marketing companies have taken advantage of this disease through the spreading of false information, in the form of fad diets, devices and products tied to fat loss, and misleading nutritional claims that often prioritize profit over health. In the modern era, fitness knowledge has been democratized with the advent of social media and digital platforms aiding in dissemination of health information (Johnson and Lee 2019). Despite this, obesity in America has only gotten worse- increasing to 41.9% in 2020 as one of the top 10 most obese countries in the world ("Health Effects of Overweight and Obesity" 2022).

Obesity is a multifaceted challenge with ties to lifestyle, socioeconomic factors, and access to health education and resources. In this study I use a logistic regression model in order to predict the odds that an individual is obese based on the amount of physical activity they do

*Code and data are available at: https://github.com/bestmustard/activity-bmi

per week and their level of income. As logistic regression models are used to model binary outcomes, my outcome will be whether an individual has a BMI of 30 or above or not.

The data I am using is from the Center of Disease Control and Prevention (CDC), a government-officiated service organization dedicated to health research in the United States. In particular, this dataset contains information on an adult's diet, physical activity, and weight status from CDC's Behavioral Risk Factor Surveillance System- America's premier system for collecting data about health-related risk behaviours conducted through telephone surveys (Centers for Disease Control and Prevention (CDC), National Center for Chronic Disease Prevention and Health Promotion, Division of Nutrition, Physical Activity, and Obesity 2023). I will be focusing on the variables of activity level and level of income as explanatory variables in the logistic regression model. The estimand will be the probability that an individual is obese based on these factors.

My report is structured into four main sections following the introduction. In the first section, I describe the data utilized for my analysis, presenting the CDC dataset as well as graphs that show the distribution of the explanatory variables. The second section details the logistic regression model, including the rationale for its use and an interpretation of preliminary findings. Next I will analyse the variables' impact on obesity prevalence through the use of graphs and specific numerics from my results. Finally, I discuss the implications of my findings, address potential weaknesses in my study, and suggest directions for future research.

This analysis is conducted using R, using several R packages to facilitate my analysis and presentation. This includes tidyverse for data manipulation and visualization, knitr for report generation, modelsummary for model interpretation, and rstanarm for Bayesian regression modeling (R Core Team 2021; **Dataverse?**; Xie 2021; Arel-Bundock 2021; Goodrich et al. 2022; Kay 2021; Wickham et al. 2021). Some portions including ggplot graphs and the "Discussion" section were written with the help of ChatGPT4 OpenAI (2023).

2 Data

I used a dataset called "Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System" pulled directly from the CDC website (Centers for Disease Control and Prevention (CDC), National Center for Chronic Disease Prevention and Health Promotion, Division of Nutrition, Physical Activity, and Obesity 2023). The .csv file obtained from the website contains 93250 data points with the relevant information of an individual's activity level based on their survey response, BMI, and various demographics. The dataset is owned by the Division of Nutrition, Physical Activity, Obesity (DNPAO), a division under the CDC which focuses directly on preventing chronic diseases by promoting better nutrition practices.

2.1 Source Reliability

The Centers for Disease Control and Prevention (CDC) is recognized as a reliable source for obesity data for several reasons. The CDC's Division of Nutrition, Physical Activity, and Obesity conducts comprehensive surveillance and research to understand and address obesity, focusing on policy and environmental strategies to promote healthy eating and active living. The organization collects data from the largest scale health survey systems in the United States, including both the Behavioral Risk Factor Surveillance System (BRFSS) and the National Health and Nutrition Examination Survey (NHANES) ("Data & Statistics | Overweight & Obesity | CDC," n.d.; "Overweight & Obesity | CDC," n.d.).

The BRFSS conducts more than 400,000 interviews annually across the United States, offering state-level estimates of obesity prevalence among other health parameters. The large sample size it offers provides a better understanding of the population when compared to smaller samples ("4 Comparison of Data Sources Used to Assess Obesity Prevalence and Trends," n.d.).

2.2 Limitations

The CDC is a reputable government-associated organization but the data did not come without inherent limitations. As with any telephone survey, respondents are susceptible to lying which would represent false data points. In addition, even if respondents believe they are telling the truth, it is possible that they do not have an accurate measurement of, for example, their activity level. There is no information available about the methods used to validate this information on the CDC website.

The data used in this paper does not represent the full dataset, as data points had to be removed for any missing responses. The previously more robust dataset with 93250 data points was reduced to 10218 in this process. The categories available for both the respondent variables do not allow for some details, as the markers for physical activity were very specific. Many individuals may not adhere directly to the possible responses in the survey, and it is

impossible to account for individuals who perform more physiccal activity than the provided options. The income levels also do not provide a broad perspective, with the maximum level being $75,000+. There is no evidence or rationale provided regarding why these options were chosen.

2.3 Variables of Interest

2.3.1 Activity Level

Physical activity is a determinant of energy expenditure and is fundamentally linked to obesity and weight management. Regular physical activity can significantly reduce the risk of becoming obese by increasing the number of calories the body uses for energy (Hill and Peters 2003). Conversely, sedentary lifestyles are closely associated with obesity due to low energy expenditure. Research has consistently shown that low physical activity levels are predictive of obesity development over time. Incorporating various forms of exercise, including strength training and aerobic activities, can aid in maintaining a healthy weight and preventing obesity (Sallis and Glanz 2012).

2.3.2 Income

Income level is a social determinant of health that influences obesity rates. Higher income levels often correlate with better access to healthy foods, recreational facilities, and health services, which can contribute to lower obesity rates (Pickett and Wilkinson 2005). Conversely, lower income levels are associated with limited access to healthy food options, reliance on cheaper, calorie-dense processed foods, and reduced opportunities for physical activity (Drewnowski and Specter 2010). This economic disparity creates environments conducive to obesity development, particularly in communities where affordable healthy options are scarce. Research indicates that socioeconomic status, including income, plays a substantial role in the prevalence and distribution of obesity within populations.

2.3.3 Other Variables

While I believed nutrition information would have been a suitable variable, the options provided in the dataset were only if the individual had “No Fruits” or “No Vegetables” in their regular diet. I did not think these two options were enough to make relevant conclusions.

2.4 Data Preparation and Cleaning

The data was first downloaded as a .csv file directly from the CDC website and then saved in parquet format using the arrow package for efficient storage and access (Richardson et al. 2024). The cleaning process involved filtering for the columns for the relevant variables which were "Topic" ("Physical Activity - Behavior", "Obesity / Weight Status", "Fruits and Vegetables - Behavior"), "Question" (specific responses under the topic), BMI, and Income.

The respondent variable of activity level was transformed for simplicity. Initially, the possible values included: "Percent of adults who engage in no leisure-time physical activity", "Percent of adults who engage in muscle-strengthening activities on 2 or more days a week", "Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)"... and so on. I removed the specifics and labelled them "No Activity", "Strength", "Cardio", "Cardio + Strength", "Double Cardio", as the markers of strength or cardio training remained the same throughout (strength meant 2 days of strength training, cardio meant 150 minutes of moderate intensity or 75 minutes of "vigorous-intensity"). "Double Cardio" was named as such since it was defined as double the minutes of cardio as "Cardio".

"NA" responses were then filtered out in order to make sure each data point contained all the variables used.

The cleaned dataset was then saved in both CSV and parquet formats.

The distributions for each explanatory variable are illustrated in Figure **??** and Figure **??** below:

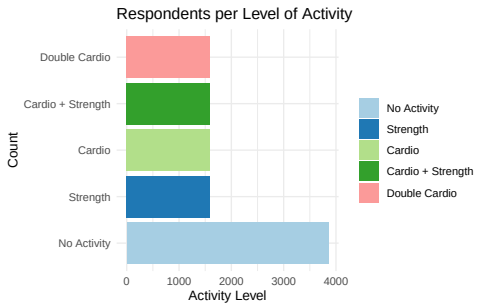

Figure 1: Number of respondents at each activity level

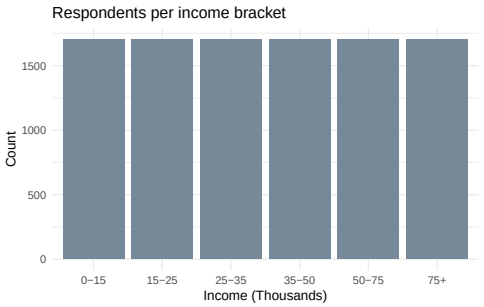

Figure 2: Number of respondents per income bracket

Figure **??** tells us that 4-year degree graduates are the most well represented in the polls, followed by high school graduates, people with some college, and then 2-year degree graduates. People who have not graduated high school are hardly represented at all.

In Figure **??** we see 45-64 being the most dominant age group, with much less representation in the 18-29 range. Also, about 20% more females participated in the study than males.

Below are some figures representing who people voted for based on their age and gender, followed by education level and gender.

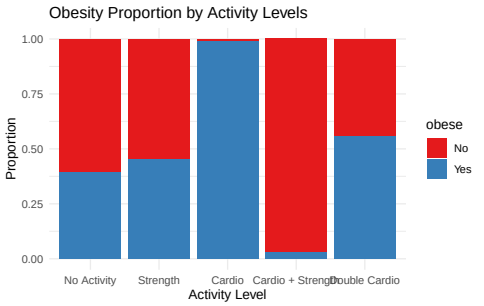

Figure 3: Prevalence of obesity by activity level

Figure **??** shows a negative correlation between votes for Joe Biden and older age groups.

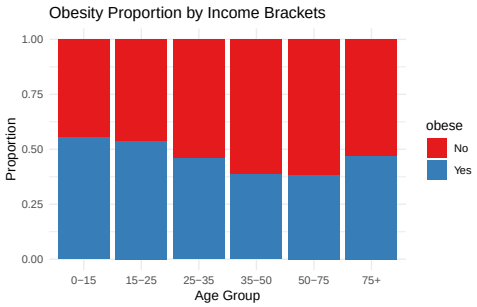

Figure 4: Prevalence of obesity by income level

We can see that in **?@fig-gendervotes**, more women voted for Joe Biden than men.

3 Model

Logistic regression is a model used when the outcome or dependent variable is binary, which fits this scenario perfectly as I am modelling the binary outcome of below 30 BMI or a BMI of 30 and above.

The regression model will calculate the log odds of the probability that a person has a BMI of 30 or above, and then map it to a probability between 0 and 1 through the logistic function.

The standard logistic function $\sigma(t)$ for a real-valued input t is defined as:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

The graph of the logistic function is an S-shaped curve known as a sigmoid curve. It approaches 1 as t goes to positive infinity and approaches 0 as t goes to negative infinity.

In logistic regression, the input t is the linear combination of predictors including the intercept, which can be represented as $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$. The logistic function then translates this into a probability that the dependent variable is 1 (has a BMI above 30).

In this situation, I will be using the predictors activity level and income and then applying the logistic function to get the probability $P(Y_i = 1)$ that a respondent i is considered obese.

This model is particularly strong at handling categorical dependent variables, which each of my explanatory variables fall under (Jr., Lemeshow, and Sturdivant 2013).