Americans who are wealthier and perform more strength training are less likely to be obese*

An analysis of how exercise habits and wealth correlate with BMI

Victor Ma

April 19, 2024

This study examines the correlation between an individual's weekly physical activity, their income level, and their probability of being obese. The only correlation found between physical activity was with individuals who strength train having lower rates of obesity, and higher income levels tended to trend with lower obesity. The findings from this study can help researchers understand the mechanisms that drive weight loss, and also shows individuals that nutrition is a greater factor than exercise with respect to body composition.

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

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

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 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) 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; Xie 2021; Arel-Bundock 2021; Goodrich et al. 2022; Kay 2021; Wickham et al. 2021). Some portions including ggplot graphs and the "Data", "Summary Interpretation", and "Discussion" sections 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) 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.

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.).

2.1 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 physical 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.

2.2 Variables of Interest

2.2.1 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).

2.2.2 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.

2.2.3 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.

2.3 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 1 and Figure 2 below:

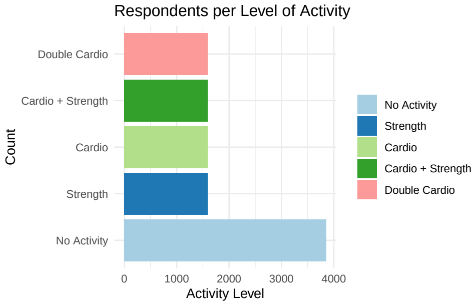

Figure 1: Number of respondents at each activity level

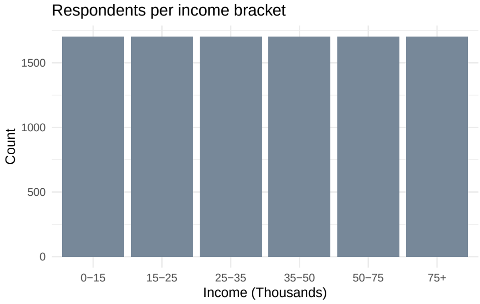

Figure 2: Number of respondents per income bracket

Figure 1 tells us that Americans who do not perform any leisurely exercise are the most well represented, with all other groups having an equal number of points in this data.

In Figure 2, oddly enough every income level was equally represented in this study.

Below are some figures representing the proportion of people who were considered obese based on their activity level and their wealth.

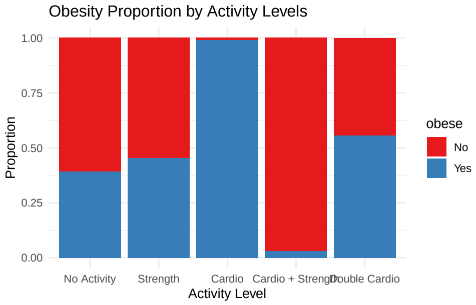

Figure 3: Prevalence of obesity by activity level

Figure 3 shows no correlation between obesity and activity level. The proportion of people with obesity doing moderate cardio is unexpected.

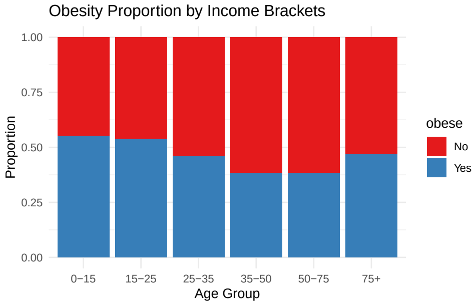

Figure 4: Prevalence of obesity by income level

We can see that in Figure 4, obesity tended to trend down with higher levels of income.

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).

3.1 Model Specification

The logistic regression model is defined as:

$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_{\text{activity},i} + \beta_2 X_{\text{income},i}$$

3.2 Model set-up

- Y_i is the binary indicator of having a BMI of 30 or above (1) versus a BMI below 30 for respondent i.
- $X_{\text{activity},i}$, $X_{\text{income},i}$ are the activity level and income of respondent i, respectively.
- β_0 represents the model intercept, while β_1 and β_2 are coefficients quantifying the effects of activity level and income on the likelihood of being considered obese.

I fit my logistic regression model to the data using ‘stan_glm()’ function from the ‘rstanarm’ package in R Goodrich et al. (2022). This function will automatically determine each of the β coefficients in the model, using a smaller slice sample of 3000 from the data we processed. This function also uses Bayesian logistic regression with the default priors from ‘rstanarm’.

3.2.1 Model Justification

We can see the model summary in Table 1:

3.2.1.1 Summary Interpretation

Each coefficient in a logistic regression model quantifies the change in the log odds of the outcome per unit change in the predictor. In this context, with the probability of being obese as the outcome:

(Intercept) (0.06): This is the log-odds of being obese when all other variables are zero. Since it’s close to zero, it suggests that when no activity is accounted for and income is at its baseline (likely the lowest income category), the likelihood of being obese is near the 50% mark on the probability scale.

activityStrength (0.18): Engaging in strength training activities is associated with a 0.18 increase in the log-odds of being obese compared to no physical activity. The positive coefficient suggests a slight association with higher obesity odds, which might be unexpected.

activityCardio (5.37): Engaging in cardio activities is associated with a substantial increase (5.37) in the log-odds of being obese. This result is counterintuitive as cardio exercises typically correlate with weight loss or lower obesity rates.

activityCardio + Strength (-2.90): Engaging in both cardio and strength activities is associated with a decrease in the log-odds of being obese. This suggests that combined exercise types might be effective at reducing the likelihood of obesity.

activityDouble Cardio (0.75): Performing a double amount of cardio is associated with an increase in the log-odds of being obese. Like the cardio variable, this positive association is counter to typical expectations.

For the income variables (reference being likely the lowest income category):

income15-25 (0.10): Being in the $15,000 to $25,000 income bracket is associated with a slight increase in the log-odds of being obese.

income25-35 (-0.72): This income bracket is associated with a decrease in the log-odds of being obese, suggesting that a higher income level correlates with lower obesity odds.

income35-50 (-0.98): Similarly, this indicates a greater decrease in the log-odds of being obese, reinforcing the trend that higher income brackets correlate with lower obesity odds.

Table 1: Explanatory model of obesity based on activity and income level

	Obesity Likelihood
(Intercept)	0.06
	(0.12)
activityStrength	0.18
	(0.11)
activityCardio	5.37
	(0.49)
activityCardio + Strength	−2.90
	(0.26)
activityDouble Cardio	0.75
	(0.12)
income15-25	0.10
	(0.16)
income25-35	−0.72
	(0.15)
income35-50	−0.98
	(0.16)
income50-75	−1.06
	(0.16)
income75+	−0.40
	(0.16)
Num.Obs.	3000
R2	0.340
Log.Lik.	−1435.801
ELPD	−1445.9
ELPD s.e.	25.8
LOOIC	2891.9
LOOIC s.e.	51.6
WAIC	2891.9
RMSE	0.40

income50-75 (-1.06): This bracket sees an even further decrease in the log-odds of obesity.

income75+ (-0.40): This suggests a smaller decrease in the log-odds of being obese for the highest income bracket compared to some of the lower brackets.

It is important to note that these are all associative relationships and don't imply causation.

In order to better interpret the results of the model, I can create a coefficient plot to visually see the effect sizes of the predictor variables on the likelihood of an individual being obese.

Figure 5 maps each predictor variable on the y-axis to an effect size and confidence interval on the x-axis. The effect size is the change in log-odds of being obese for a one-unit increase in the predictor variable, which is essentially how much impact each variable has an effect on being obese

The confidence intervals tell us the range within we can be confident that the true effect lies. Smaller confidence intervals means there is a higher level of precision in the estimate of the effect size. I am not as interested in the intervals that cross zero because that means that there is data to support each side (obese and not obese) and so they are less statistically significant.

```
[1] "term"      "estimate"  "std.error"
```

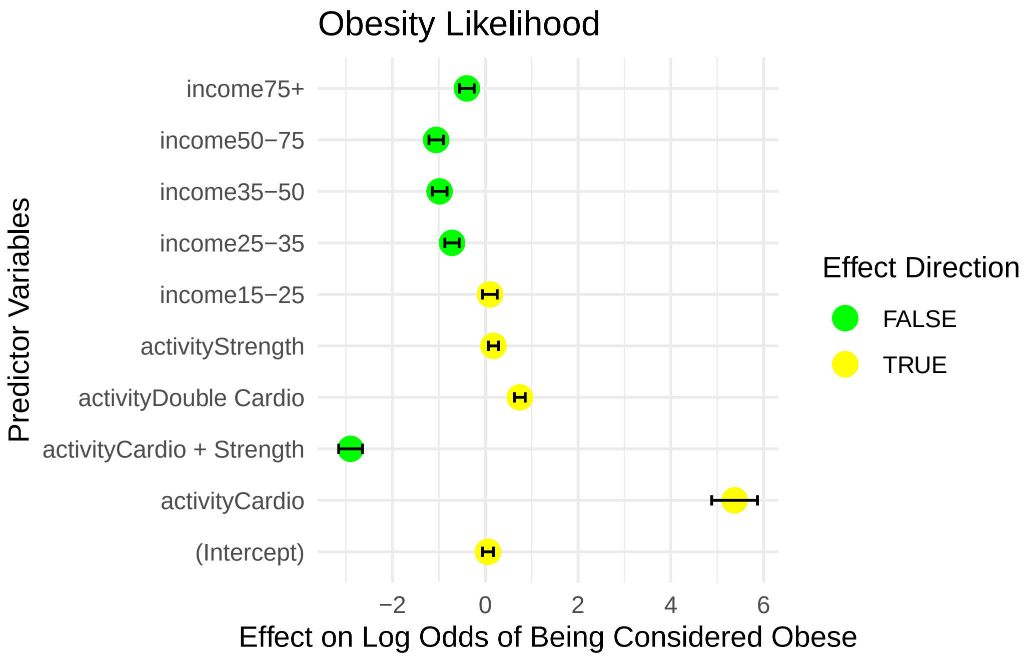

Figure 5: Coefficient plot of demographics

The conclusions I can draw from Figure 5 align with what I expected given the dataset. We can see that generally, higher incomes trend with a lower log-likelihood of being obese. The conclusions drawn from the activity levels also are in line with what we saw earlier on Figure 1, though they are not at all what I expected as I believed exercise to have a negative correlation with obesity. I did not expect that the (Intercept) no

4 Results

Figures Figure 6 and Figure 7, are recreations of the earlier graphs we saw showcasing the proportion of obesity by groups within categories, where the bar is the data from CDC while the point is the prediction generated by the model.

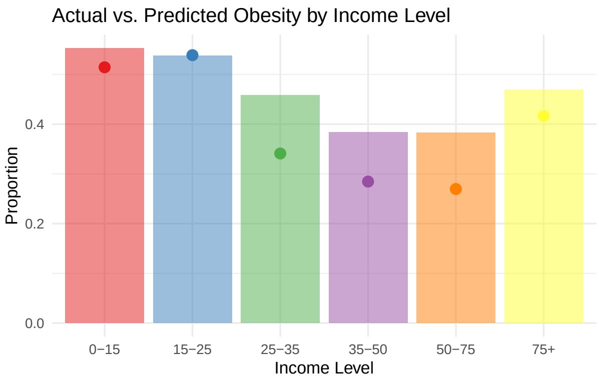

Figure 6: Model prediction for obesity by income level vs. CDC data

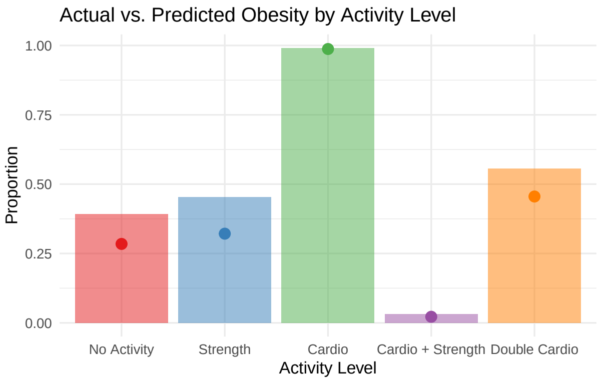

Figure 7: Model prediction for obesity by activity level vs. CDC data

We can see that the model prediction resembled the CDC data by trend, however for most of the categories within both Figure 6 and Figure 7, the prediction for obesity was lower than the data suggested. Moderate cardio having such a high proportion of obesity as unexpected.

5 Discussion

In this paper, we used real-world statistics on Americans' activity levels, income, and whether or not they were considered obese to find a correlation between these behavioural and socioeconomic factors and obesity. What we found was that contrary to popular belief, our model did not predict lower obesity levels for individuals who exercised more. In particular, there was a large proportion of individuals who did moderate cardio and were obese, but there was correlation with lower obesity and strength training. Generally, higher income levels were also associated with lower rates of obesity.

While this data may not be indicative of reality, the lack of correlation between exercise and obesity may reveal truths about societal beliefs in terms of how to get fit. Like myself, many Americans grew up watching TV where commercial ad breaks would feature fad diets, fast 10 minute work outs, and buzz phrases like "do this for 30 minutes a week to lose X lbs of fat!" Fitness may be viewed with a dogmatic stigma due to media portrayals which push the idea of a 'hardcore' lifestyle being required for a fit body.

Due to this, it is possible that individuals believe in exercise being a greater factor for fat loss than nutrition. With the highest category of 300 minutes of moderate cardio used in CDC's survey, an adult weighing 160 lbs can expect to burn approximately 1825 calories in a week (Mayo Clinic Staff 2021). As a single pound of body fat contains roughly 3500 calories, this would seem like half a pound of fat loss per week. However, this does not account for the possibility that an individual consumes more food with added physical activity, and 1825 calories spread out over a week is obly 260 calories per day. That equates to less than a small fries at McDonalds, or 2 or 3 eggs with oil.

It is impossible to predict how an individual's body composition will change without knowing their nutritional information, specifically their caloric intake. There are various reasons why higher incomes might trend with lower obesity rates. People with higher incomes also tend to have achieved higher levels of education, which may be tied to access of information. They also have more access to help from professionals, better training facilities, and more food options. The study "Nutrition quality of food purchases varies by household income: the SHoPPER study" published in BMC Public Health highlights that lower-income households tend to purchase foods of lower nutritional quality compared to higher-income households. This is due to financial constraints limiting the purchase of healthier options like fruits and vegetables, leading to a higher purchase of less healthful foods such as frozen desserts or fast food. The study emphasizes that food purchasing patterns significantly mediate income differences in dietary intake quality (French et al. 2019).

6 Weaknesses and Next Steps

As outlined previously, the data points were self reported which could lead to false data due to dishonesty or lack of care in measurement. The dataset was also filtered to remove incomplete datapoints, resulting in almost 90% of the initial dataset being removed. The discrete categories used for both physical activity and income may oversimplify the spectrum of exercise habits which overlooks the nuances of individual physical activity patterns. The income cap at $75,000 also fails to show variations in prevalence of obesity in higher income brackets.

Future research should aim to incorporate objective measures of physical activity, perhaps through wearable technology, to diminish self-reporting biases. A broader dataset, possibly integrating direct measures of physical activity and detailed nutritional intake, would enable a more comprehensive analysis. Investigating the impact of higher income brackets beyond $75,000 and accounting for regional cost-of-living differences could refine understanding of the income-obesity relationship. Additionally, longitudinal data could shed light on the temporal dynamics between exercise habits and BMI changes over time.

Further, integrating geospatial analyses to assess environmental factors, such as access to recreational spaces and healthy food outlets, could add further context to obesity determinants. Given the multifaceted nature of obesity, interdisciplinary studies combining data from healthcare, urban planning, and social sciences could offer more holistic insights.

References

Arel-Bundock, Vincent. 2021. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://CRAN.R-project.org/package=modelsummary.

Centers for Disease Control and Prevention. 2021. "Adult Obesity Facts."

Centers for Disease Control and Prevention (CDC). 2023. "Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System." Atlanta, Georgia: https://chronicdata.cdc.gov/Nutrition-Physical-Activity-and-Obesity/Nutrition-Physical-Activity-and-Obesity-Behavioral/hn4x-zwk7; Centers for Disease Control and Prevention.

"Data & Statistics | Overweight & Obesity | CDC." n.d. Centers for Disease Control; Prevention; https://www.cdc.gov/obesity/data/index.html.

Drewnowski, Adam, and S. E. Specter. 2010. "Obesity and the Food Environment: Dietary Energy Density and Diet Costs." *American Journal of Preventive Medicine* 27: 154–62.

French, Simone A., Christy C. Tangney, Melissa M. Crane, Yamin Wang, and Bradley M. Appelhans. 2019. "Nutrition Quality of Food Purchases Varies by Household Income: The SHoPPER Study." *BMC Public Health* 19 (231).

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

"Health Effects of Overweight and Obesity." 2022. Centers for Disease Control; Prevention; https://www.cdc.gov/healthyweight/effects/index.html.

Hill, James O., and John C. Peters. 2003. "Energy Balance and Obesity." *Circulation* 104: 51–52.

Johnson, Mark K., and Angela R. Lee. 2019. "The Role of Digital Media in Shaping Health Behaviors: Opportunities and Challenges." *Health Communication Today* 24 (5): 456–67.

Jr., David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. New York: John Wiley & Sons.

Kay, Matthew. 2021. *Tidybayes: Tidy Data and Geoms for Bayesian Models*. https://CRAN.R-project.org/package=tidybayes.

Mayo Clinic Staff. 2021. "Exercise for Weight Loss: Calories Burned in 1 Hour." https://www.mayoclinic.org/healthy-lifestyle/weight-loss/in-depth/exercise/art-20050999.

OpenAI. 2023. "ChatGPT: Optimizing Language Models for Dialogue." https://openai.com/.

Pickett, Kate E., and Richard G. Wilkinson. 2005. "The Social Determinants of Health: The Solid Facts." *International Journal of Epidemiology* 34: 1245.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Sallis, James F., and Karen Glanz. 2012. "Role of Physical Activity in the Prevention of Obesity in Children." *International Journal of Obesity* 14: 34–38.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.