Estimating the Impact of Eating Habits on Obesity

Chris DeMaio, Hope Huang, Zukang Yang

2023-04-22

Introduction

Obesity is a significant public health concern. It is associated with an elevated risk of multiple adverse health outcomes, including heart disease, stroke, type 2 diabetes, and certain cancers. Obesity also has sizable economic and social implications. Studies show that medical costs for people with obesity in the U.S. tend to be 30% to 40% higher than those for people without obesity[1]. Poor dietary habits constitute a primary contributor to obesity and a body of research consistently reveals that diets high in calories, saturated fat, and added sugars are strongly linked to heightened obesity risk. Consequently, gaining a deeper comprehension of the connection between eating habits and obesity can empower individuals to make more informed decisions regarding their dietary and lifestyle choices.

This study investigates the influence of some eating habits on individual's obesity levels while accounting for demographic characteristics such as gender, age, and family history of obesity. Leveraging a set of regression models based on data about respondents' dietary habits and demographic information, we estimate the changes in BMI(body mass index) and ultimately gather insights about the dietary habits that significantly influence the likelihood of obesity.

Data and Methodology

Our study is based on a dataset comprising 2,111 instances, obtained from a research study[2] aimed at estimating obesity levels in individuals from Mexico, Peru, and Colombia. The dataset is publicly available through the UCI Machine Learning Repository. Each row of the data represents an individual from the countries mentioned earlier, with ages between 14 and 61, and contains information about their eating habits, physical conditions, and demographics. The data was collected in 2019 using a web-based survey platform. We performed all exploration and model building on a 30% subsample of the data. The remaining 70%, totaling 1,479 rows, were used to generate the statistics in this report.

We consider several key eating habits as independent variables, including frequent consumption of high-caloric food, number of main meals, frequency of eating between meals, and water consumption. Our selection of these variables is based on observed statistically significant correlations with obesity during the exploratory process. Additionally, to operationalize obesity, we utilize the commonly used BMI metric which is calculated by dividing weight by height squared. It is a widely accepted measure of an individual's fitness. Due to its simple and cost-effective nature, BMI is commonly employed by medical professionals to assess obesity levels.

For example, one of the main variables we are examining is if individual's who frequently consume high-caloric food influences their BMI. We can see from figure 1 that in both male and females, the regression lines for those who consume high-caloric food are higher than those who do not. We also observe that BMI and Age do not have a straightforward relationship. Though, we do notice that females have a sizable cluster of data between 40 and 50 BMI and 20 and 25 years old and males do not. This indicates there could be a difference in average BMI between males and females.

[1] Public Health Considerations Regarding Obesity. StatPearls, 2022, https://www.ncbi.nlm.nih.gov/books/NBK572122/.

[2] Fabio Mendoza Palechor, Alexis de la Hoz Manotas, Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico, 2019, https://doi.org/10.1016/j.dib.2019.104344.

Our exploratory analysis on the subsample indicates that several readily available variables associated with physical conditions, such as smoking and frequency of physical activities, do not exhibit statistically significant correlations, nor do they influence the direction and magnitude of the variables of interest. Although such variables are potential confounding factors that require careful consideration in our models, our results suggest that their exclusion would be appropriate, possibly due to data quality issues. As such, we removed them from our analysis.

Our primary interest is exploring the relationship between eating habits and obesity levels. To this end, we developed large-sample linear regression models incorporating dietary and demographic variables. Specifically, we created a base model that included only eating habits, an extended model that incorporated demographic characteristics, and a full model that added a new variable, $R \cdot (drinking\ alcohol)$, to the analysis. We chose to include $R \cdot (drinking\ alcohol)$ despite not initially considering it a significant contributor to obesity, as we lacked sufficient evidence to exclude it from our analysis. Therefore, we fitted regressions of the form,

$$\widehat{BMI} = \beta_0 + \beta_1 \cdot R \cdot (high\ calorie\ consumption) + \beta_2 \cdot (number\ of\ meals) + \\ \beta_3 \cdot R \cdot (frequency\ of\ food\ between\ meals) + \beta_4 \cdot R \cdot (water\ consumption\ above\ 2L) + \mathbf{Z}\gamma$$

where R is an indicator for the categorical eating habit variables, β_1, β_2, β_3, and β_1 represent the respective effect of the eating habit variables on BMI. $\mathbf{Z}$ is a row vector of additional covariates, and γ is a column vector of coefficients. By fitting these 3 models, we measure how the estimated coefficients and robust standard errors to the key variables change across the models. This helps us assess how these variables are affected by endogeneity bias.

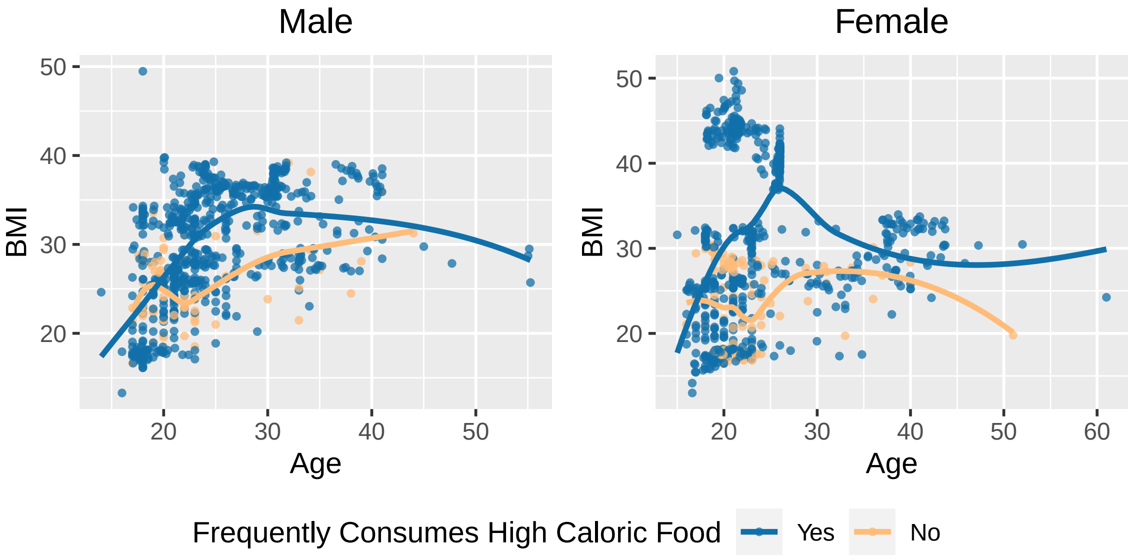

Figure 1: BMI vs. Age by Gender

Results

Table 1 shows the results of our analysis. Overall, across 3 models, some variables exhibit significance with relatively consistent estimated coefficients. This suggests that the inclusion of new variables does not significantly alter the influence of existing variables in the models on the dependent variable, or at most, affects them to a minor extent. However, some other variables, such as $R \cdot (food\ between\ meals - frequent)$, present noticeable changes in their estimated coefficients. We suspect that the uncertainty in estimation is influenced by some omitted variables.

Specifically, The results show that across all models, the key coefficient on $R \cdot (high\ calorie\ consumption)$ is highly statistically significant, with point estimates ranging from 2.52 to 4.64. To put this into perspective,

Table 1: Estimated Regressions

	Output Variable: BMI (Body Mass Index)		
	(1)	(2)	(3)
$R \cdot (high\ calorie\ consumption)$	4.64^{***}	3.19^{***}	2.52^{***}
	(0.40)	(0.41)	(0.42)
number of main meals	0.44^{*}	0.50^{**}	0.27
	(0.20)	(0.19)	(0.19)
$R \cdot (food\ between\ meals - sometimes)$	6.07^{***}	1.75	2.25^{*}
	(0.54)	(0.95)	(0.88)
$R \cdot (food\ between\ meals - frequent)$	-2.45^{***}	-4.34^{***}	-3.47^{***}
	(0.65)	(1.00)	(0.93)
$R \cdot (water\ consumption\ above\ 2L)$	-2.43^{***}	-1.80^{***}	-1.52^{***}
	(0.43)	(0.40)	(0.39)
$R \cdot (male)$		-1.87^{***}	-1.64^{***}
		(0.35)	(0.33)
age		0.17^{***}	0.17^{***}
		(0.03)	(0.03)
$R \cdot (with\ family\ obesity\ history)$		7.22^{***}	7.63^{***}
		(0.41)	(0.41)
$R \cdot (drinking\ alcohol)$			3.27^{***}
			(0.34)
Constant	21.57^{***}	16.94^{***}	14.80^{***}
	(0.76)	(1.11)	(1.05)
Observations	1,479	1,479	1,479
Adjusted R^2	0.22	0.36	0.40
Residual Std. Error	7.05 (df = 1473)	6.39 (df = 1470)	6.22 (df = 1469)
F Statistic	85.36^{***} (df = 5; 1473)	105.51^{***} (df = 8; 1470)	108.30^{***} (df = 9; 1469)

Note: HC_1 robust standard errors in parentheses.

our findings indicate that individuals with high-calorie consumption habits, given similar demographic characteristics and other eating habits will, on average, have an increase in BMI of 2.52, as suggested by model 3.

Moreover, the analysis also highlights that the key coefficient on $R \cdot (\textit{daily water consumption above } 2L)$ is statistically significant across all models, with point estimates ranging from -2.43 to -1.52. This implies that respondents who drank over 2L water a day had a reduction in BMI by 1.52, as model 3 suggests.

However, our findings also indicate that the coefficients on the two indicator variables derived from $frequency\ of\ food\ between\ meals$ did not present an interpretable pattern. Specifically, the coefficients on $R \cdot (food\ between\ meals - sometimes)$ are not all significant, with model 2 showing insignificance after including demographic variables, but significance in model 3 after including $R \cdot (drinking\ alcohol)$. Additionally, all models' coefficients on $R \cdot (food\ between\ meals - frequent)$ are statistically significant. However, their negative values make it difficult to interpret the effect of this variable in conjunction with that of $R \cdot (food\ between\ meals - sometimes)$. For instance, respondents who frequently ate between meals had less BMI, while those who sometimes ate between meals tended to have much higher BMI. We require further investigation to understand the reason behind this counter-intuitive finding.

Interestingly, $number\ of\ main\ meals$ became insignificant in model 3 after introducing $drinking\ alcohol$. Our study did not initially consider alcohol consumption as a contributing factor to obesity level, but our analysis shows that $drinking\ alcohol$ is statistically significant and increases the adjusted R^2 by a considerable margin. However, the interaction between $number\ of\ main\ meals$ and $drinking\ alcohol$ remains a mystery and requires further investigation.

Limitations

Ensuring consistent regression estimates requires that observations are independently and identically distributed (iid). Nevertheless, our data was collected using a web platform with a survey where anonymous users from Mexico, Peru, and Colombia responded to each question, which may result in geographical clustering. As such, the data might not originate from the same distribution. Additionally, as the data did not label the country to which each user belonged, we could not account for the mixed effect between the country of origin and eating habits. Moreover, it is worth noting that a significant portion, up to 77% of the data utilized in our data were generated synthetically by the original study through the application of the SMOTE[3] technique on the initial 23% of the data. The synthetic data further violates the iid assumption for large sample regression as they depend on the initial data from which they were generated.

The accuracy of our study may be restricted by our dependent variable, because the BMI fails to account for various pertinent attributes, such as differences in bone density, muscle mass, and sex. For instance, an individual who identifies as male and possesses above-average muscle mass, but below-average height, may register a BMI score exceeding 30, despite maintaining a healthy physique.

As far as structural limitations, several omitted variables may bias the estimates. An example variable that may interact with the key variables in the true model is stress level. We expect a positive correlation between stress level and the key variables if high stress influences one to consume more high-calorie food. Since stress level likely positively affects BMI, there is a positive omitted variable bias on the key variables. Therefore, the omitted variable bias is away from zero. This makes the hypothesis test overconfident.

Conclusion

Based on our findings, we should encourage individuals to avoid frequently consuming high-caloric food, drink more than 2L of water a day, and avoid drinking alcohol. However, this is only a single study and it only serves to provide some guidance for future research. The aim of this study is to examine which eating habits have significant influence on an individual's BMI so people with higher BMI's know where they should focus their efforts to becoming healthier.

[3] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer SMOTE: synthetic minority over-sampling technique, 2002, https://www.jair.org/index.php/jair/article/view/10302.