Proceedings of CONF-MPCS 2024 Workshop: Quantum Machine Learning: Bridging Quantum Physics and Computational Simulations

DOI: 10.54254/2753-8818/51/2024CH0143

Correlation between obesity and income: A statistical analysis using linear regression models

Yibing Jiang

Shanghai American School, Shanghai, 200000, China

yibing01pd2025@saschina.org

Abstract. The rate of obesity across the world has seen a steady increase in the past several decades. This disease can lead to significant harm associated with one's body functioning and mental health. The potential causes of obesity incorporate social, environmental, and personal factors. This article attempts to cover the correlation between obesity and income level using a linear regression model. Samples were extracted from the National Health and Nutrition Examination Survey about 545 individuals from 2017 to 2020. The dataset was subdivided into four cohorts, which include non-Hispanic White male, non-Hispanic White female, non-Hispanic Black male, and non-Hispanic Black female, respectively, to avoid variables that might confound with the explanatory variable. The study found out that the correlation for White male, White female, and Black female can be either positive, negative, or non-existent, and positive on the other hand for the Black male cohort. However, the study concluded that the linear regression model is ineffective for the topic of analysis as the metrics indicate. The incompatibility of the model can be attributed to the fact that other variables that can possibly confound with the explanatory variable are neglected and not filtered out.

Keywords: Obesity, BMI, income-to-poverty ratio, linear regression, confounding variables.

1. Introduction

The rate of obesity across the world has seen a steady increase in the past several decades; more adults nowadays are suffering from the symptom [1]. The obesity rate is predicted to reach a vast number in future years: according to Kelly et al., while the number of obese adults was approximately 396 million in 2005, in 2030, it will reach up to about 573 million if adjusted for secular trends and 1.12 billion if not [2]. As obesity has already become one of the most prevalent causes of death in modern days, being on the top fifth of the list, citizens and scholars should start to appreciate the severeness of the health issue and investigate possible ways to resolve the issue [3].

Obesity can lead to significant harm associated with one's body functioning and mental health. It poses several other symptoms and diseases such as multiple chronic diseases, such as diabetes, cancer, cardiovascular diseases, etc. [4]. Obesity can also lead to mental health problems and diseases that takes on a physical form such as musculoskeletal disorders, which all pose non-negligible influences on one's daily life [5]. Problems associated with mental health mainly stem from certain norms specified by the media, the entertainment industry, and the political environment [6]. Up to 2014, the approximated bill spent by an obese individual on obesity-related healthcare is $1901 per year, a very shocking number by a single patient [5].

DOI: 10.54254/2753-8818/51/2024CH0143

While obesity seems to be a common term used to refer to someone who is overweight, the fact is that the two concepts are very different from each other. According to the World Health Organization (WHO), BMI, which is given by body weight divided by height squared, is the typical index that defines whether a person is underweight, normal, overweight, or obese [7]. A BMI score between 15 and 20 defines underweight; 20 to 25 defines "just right", or normal; 25 to 30 categorizes the area for being overweight; and a score above 30 indicates obesity [7].

Besides digging into the potential effects of obesity, the topic of observing and analyzing the causes of obesity has also received abundant attention from scientists. These causes include social, genetic, environmental, and personal factors. An Example regarding personal factors is the imbalance between consuming calories and burning off calories in the long term [8]. Another example in terms of genetic influences is a mutation in the hormone leptin, which regulates the balance between energy intake and expenditure and sometimes works to decrease the former while increasing the latter [9]. However, in this paper, the social aspect is going to be the main focus. The particular social factor that the author is going to discuss in this paper is income.

Mathieu-Bolh claimed the income and obesity among rich countries are closely related, and the relation is fairly complex and subtle, where the correlation can be either positive or negative [10]. Thus, to analyze the true relationship between income and obesity, statistical models are needed. One study that utilized mathematical models was conducted by Kim and Knesebeck, which applied meta-analytic methods along with random-effect models [11]. This paper will attempt to use a linear regression model to discover the relationship between income and obesity and the relative coefficients associated with this model.

2. Methodology

2.1. Data collection

This study utilizes data collected from the National Health and Nutrition Examination Survey (NHANES) from 2017 to 2020 about 15560 surveyed individuals. 545 individuals were chosen as the subjects of this study after eliminating potential outliers and filtering out particular race and age groups since these factors are correlated with BMI [12]. Gender is also split to prevent confounding with the independent variable.

2.2. Variable selection

The subjects' income-to-poverty ratio is selected as the explanatory variable to predict the BMI, measured by one's weight divided by height squared. Because race, age, and gender have been found to correlate with BMI, this study will attempt control for these confounding variables by splitting the sample into four different groups: non-Hispanic white male, non-Hispanic white female, non-Hispanic black male, and non-Hispanic black female. Non-Hispanic white and black individuals were selected as the subjects since they represent the majority of the sample chosen and the US population [13]. Moreover, the study will use young adults aged 20 to 29 from the dataset to avoid further confounding.

2.3. Model introduction

This study will use a simple linear regression model to assess the correlation between BMI and income. The model suggests that the relationship between the independent and the dependent variable is linear. A confidence interval (CI) will be applied to the data in order to infer the population parameter, which is the true relationship between BMI and income for US adults aged 20 to 29 from 2017 to 2020. A confidence interval approximates the range of values of which the population parameter has some chance of falling in between. This chance is determined by the confidence level, which claims the probability, or the level of confidence, that the confidence interval captures the true value of the population parameter. For instance, a 90 percent confidence level means that the probability of capturing the true population parameter within a certain confidence interval is 90 percent. In this study, a 95

DOI: 10.54254/2753-8818/51/2024CH0143

percent confidence level will be used to calculate the associated confidence interval to approximate the true slope that associates income with BMI.

Furthermore, the formula of a confidence interval is given by the point estimate adding or subtracting the sample's margin of error. The margin of error is calculated by multiplying the test statistic by the standard error of the sample statistic. Thus, the mathematical expression for the confidence interval of a linear regression line is given as b ± (t * SE(b)) for 100 (1-α)% confidence level. The t-statistic is used since population standard deviation of the slope is unknown, and it provides an estimation of the z-statistic which models a normal curve. The t-statistic is based on the degree of freedom of the sample, obtained from the sample size minus two.

In addition, this study will also attempt to make use of hypothesis testing along with confidence interval. A hypothesis test comprises two hypotheses: the null hypothesis (Ho) and the alternative hypothesis (Ha). The null hypothesis assumes that a given parameter is equal to a certain value, whereas the alternative hypothesis assumes the parameter is either greater than, less than, or simply unequal to the assigned value. A hypothesis testing is also based on the p-value, which states the probability, when the null hypothesis is true, that the alternative hypothesis will happen due to chance solely. Therefore, using a significance level (α) of 5 percent in this study, a p-value lower than α denotes that the null hypothesis should be rejected, and that there is sufficient evidence to embrace the alternative evidence; on the other hand, a p-value greater than or equal to α indicates that the null hypothesis has been failed to rejected from and no convincing evidence supports the alternative hypothesis. The two hypotheses that this study will use are exhibited as follows:

$$Ho: \beta = 0 \quad (1)$$

$$Ha: \beta \neq 0 \quad (2)$$

where β denotes the population regression line correlating BMI with income-to-poverty ratio. The study uses these two hypotheses for hypothesis testing since a value of 0 may be contained in a computed confidence interval for the samples. The two hypotheses will allow for a clarification of whether the association between the independent and the dependent variable exists. This study hypothesizes that the linear correlation between BMI and income-to-poverty ratio for the four cohorts listed in 2.3 is negative.

3. Results and discussion

3.1. Simple linear regression

Tables 1 to 4 show the results of the linear regression analyses for the cohorts. Results contain the regular coefficients, standard error coefficients, t-statistics, and p-values for the constants and the x-variables (Income-to-poverty ratio). The linear regression model uses the values of the x-variables predict the outcome of the y-variables, which are the BMIs for the four cohorts. Sample sizes for each cohort are also listed in the following table 1.

Table 1. Results of linear regression analysis for the sample

Cohort	Predictor	Coef	SE Coef	t-stat	p-value
WM[a] (n = 120)	Constant	28.4824	1.2618	22.5722	0.0000
	Income-to-poverty ratio	-0.2843	0.5426	-0.5240	0.6012
WF[b] (n = 150)	Constant	27.7637	1.3395	20.7273	0.0000
	Income-to-poverty ratio	-0.0505	0.5196	-0.0971	0.9228
BM[c] (n = 122)	Constant	24.5489	0.9920	24.7481	0.0000
	Income-to-poverty ratio	0.9406	0.4278	2.1989	0.0298
BF[d] (n = 153)	Constant	29.7183	1.0413	28.5384	0.0000
	Income-to-poverty ratio	0.3083	0.5414	0.5694	0.5699

Note: [a] Non-Hispanic White male; [b] Non-Hispanic White female; [c] Non-Hispanic Black male; [d] Non-Hispanic Black female.

DOI: 10.54254/2753-8818/51/2024CH0143

3.2. Confidence interval

A 95 percent confidence interval for each cohort can be computed by utilizing the coefficient, standard error coefficient, and the t-statistics from the linear regression analysis. Table 2 shows the upper and the lower bound of the confidence intervals of the slopes relating BMI to income-to-poverty ratio for each of the four cohorts. Results are rounded to four digits after the decimal.

Table 2. Confidence intervals of the regression line relating BMI to income-to-poverty ratio

Cohort	Confidence Interval	
	Lower (95%)	Upper (95%)
WM	-1.3588	0.7901
WF	-1.0792	0.9783
BM	0.0937	1.7875
BF	-0.7614	1.3780

3.3. Hypothesis testing

This section will use hypothesis testing to validate (or invalidate) the association between BMI and income-to-poverty ratio for non-Hispanic white male, non-Hispanic white female, and non-Hispanic black female. Table 3 shows the p-value of the regression line for each of the cohorts and the interpretations based on the significance level. Results are rounded to four digits after decimal.

Table 3. Hypothesis testing of the cohorts

Cohort	*p*	Interpretation (using $\alpha = 0.05$)
WM	0.6012	Fail to reject Ho; no sufficient evidence for Ha
WF	0.9228	Fail to reject Ho; no sufficient evidence for Ha
BM	0.0298	Reject Ho; sufficient evidence for Ha
BF	0.5699	Fail to reject Ho; no sufficient evidence for Ha

3.4. Discussion

From the indexes above, those of the confidence intervals exhibit the fact that the correlation between BMI and income-to-poverty ratio for the cohorts of non-Hispanic White male, non-Hispanic White female, and non-Hispanic Black female are either positive, negative, or non-existent. The hypothesis testing procedure is applied along with the confidence interval and interrelates with it (note that confidence level = 100 (1 - α) %). The results corroborate this theory as for each of the three cohorts mentioned, there is no sufficient evidence suggesting that the population slopes are non-zero in the case where the confidence level is 95 percent. On the other hand, the non-Hispanic Black male cohort indicates that its confidence interval includes values only above zero, and the hypothesis testing also finds sufficient evidence for the population slope to be non-zero.

Different from the hypotheses posited in the earlier sections, which stated that the relationships ought to be negative, the true association for the non-Hispanic White male, female, and Black Female cohorts may also not exist or be positive. The reason for the assumption was based on the fact that poor households in the US have lower access to healthy and fresh food than rich households; poverty-dense areas are sometimes known as “food deserts” because of this, and it is estimated that 43 percent of households below the poverty threshold in the US are unable to suffice their own food supplies [14]. Therefore, individuals who are in poverty usually resort to cheap food sources that are energy-abundant but possess fewer nutritional values to suffice their daily needs. In addition, poverty may also contribute to sedentariness, which subsequently incurs obesity: people living in poverty-dense areas may experience higher rates of violence and have lower access to physical equipment, thereby spending less time outdoors and thus becoming obese [14]. Nevertheless, one important phenomenon to note is that poor people have higher involvement rate in labor-intensive occupations than rich people because these jobs do not require high levels of education. On the other hand, individuals above the poverty threshold

DOI: 10.54254/2753-8818/51/2024CH0143

involve more in sedentary jobs, where they spend most of their time in front of computers. The correlation between sedentariness and income, therefore, is indefinite.

The major limitation of this study is that only the gender, age, and race are extracted from the sampled individuals and no further characteristics were considered. Apart from the extracted features, the subjects' occupation, residential area, relative daily habits and customs, etc., should also be collected and filtered for. In reality, social, economic, environmental, and mental factors can all influence the rate of obesity. This study fails to consider some of these potential factors and therefore introduces confounding variables that had weakened the association between the explanatory and the response variables. Further relevant studies should aim to reduce (and hopefully, to eliminate) the chance of confounding variables. By doing so, it will be more valid to construct an association between the two chosen variables.

Furthermore, the relevant error metrics imply that a model different from the linear regression one would be better at discovering the correlation between income and BMI in the US. Table 4 shows the error metrics. Since an absolute value that is close to 1 for both multiple regression and R-squared indicates a more proper fit of the regression, the statistics listed below suggest a poor fit of the model. Additionally, the Q-Q plot of the residuals should be about normally distributed for the linear regression model to be valid. However, figures 1,2,3, 4 indicate that none of the residual distributions for the four cohorts are about normal. Therefore, a statistical model besides linear regression is more appropriate to be used to analyze the relationship between BMI and income level.

Table 4. Multiple regression and R-squared statistics for the four cohorts

Cohort	Multiple regression	R-squared
WM	0.0482	0.0023
WF	0.0089	0.0001
BM	0.1968	0.0387
BF	0.0463	0.0021

From the statistics above, it can be concluded that only the non-Hispanic Black male cohort shows a somewhat weak relationship between BMI and income-to-poverty ratio. The other three cohorts, however, do not fit a linear regression model.

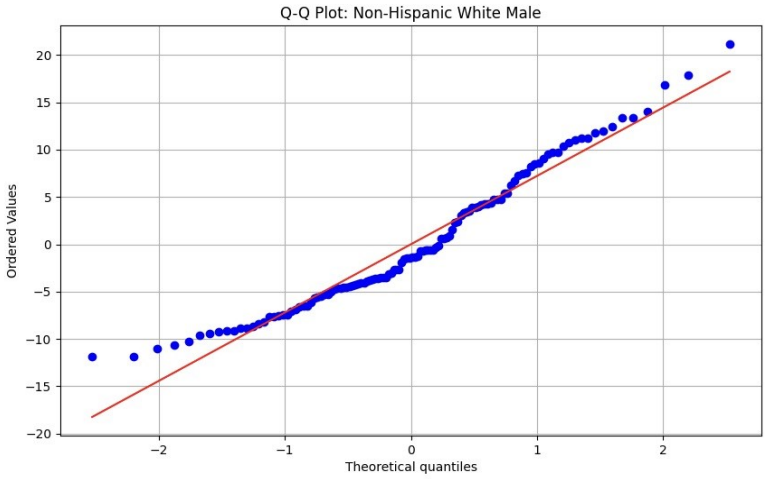

Figure 1. Q-Q plot of non-Hispanic White male cohort.

DOI: 10.54254/2753-8818/51/2024CH0143

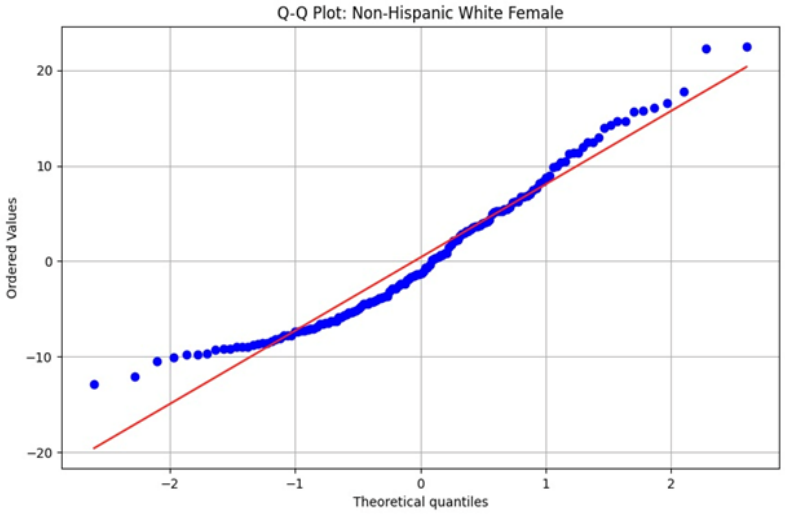

Figure 2. Q-Q plot of non-Hispanic White female cohort.

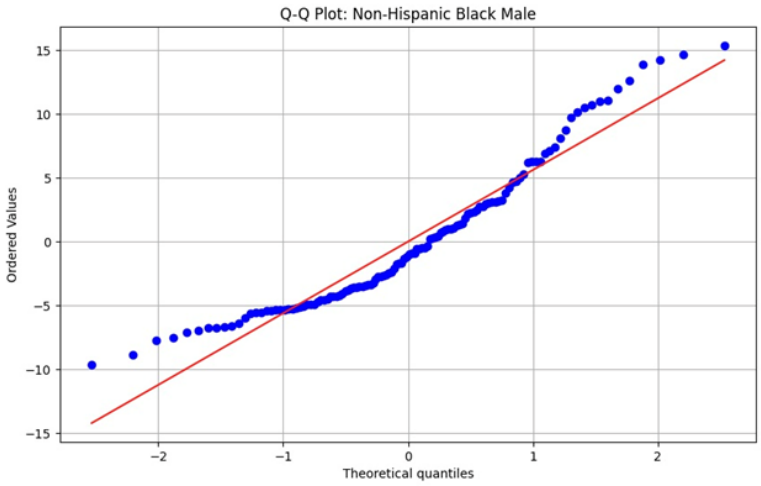

Figure 3. Q-Q plot of non-Hispanic Black male cohort.

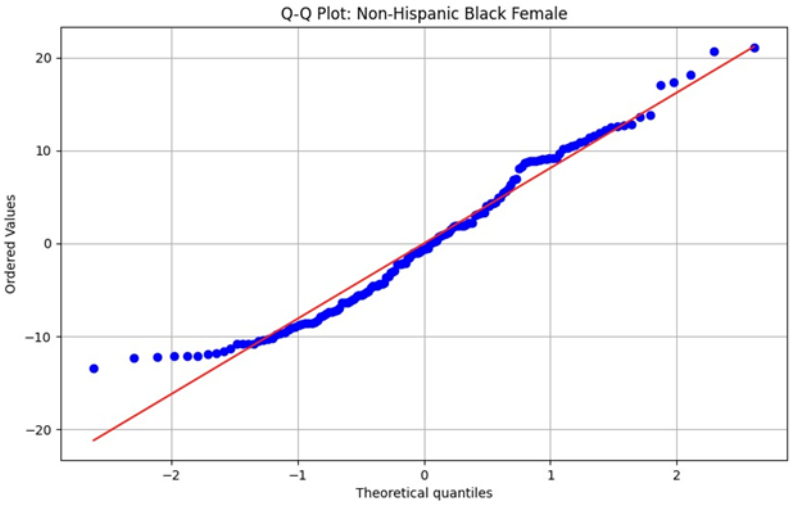

Figure 4. Q-Q plot of non-Hispanic Black female cohort.

DOI: 10.54254/2753-8818/51/2024CH0143

In order for the residuals to be approximately normally distributed, the Q-Q plots ought to demonstrate that the points are close to the linear line. However, the residuals on the plots from figures 1, 2, 3, 4 all exhibit a heavy left tail and a light right tail, indicating right-skewed distributions. Therefore, the residuals are not normally distributed and the accuracy of the linear model is thus undermined.

4. Conclusion

The study has taken its sample from the National Health and Nutrition Examination Survey about over 14000 individuals from 2017 to 2020. The sample was filtered based on gender, race, and age groups. More specifically, the study included only respondents aged 20 to 29 and classified them based on race and gender; the groups were: non-Hispanic White male, non-Hispanic White female, non-Hispanic Black male, and non-Hispanic Black female, respectively.

The study used a linear regression model to assess the correlation between BMI and income-to-poverty ratio for the filtered sample. It has been found out that apart from the non-Hispanic Black male cohort, which showed that the population slope relating BMI to income level is positive using a 95 percent confidence level, the population slope of the rest of the three cohorts can be either positive, negative, or zero (i.e., a linear relationship between BMI and income does not exist). Furthermore, a hypothesis testing with a significance level of 5 percent was utilized along with the confidence interval to corroborate that no sufficient evidence suggests the population regression line for these cohorts to be non-zero.

The study finds its major limitations to be that there are confounding variables affecting the establishment of validity of this experiment. Besides gender, race, and age groups, other variables that influence BMI include one's residential area, occupation, daily habits and customs, etc. These factors should also be taken into account to avoid confounding with the independent variable. Additionally, the linear regression model performed in this study is not so effective as indicated by the error metrics and relative indexes. However, it is unsure of whether the inefficacy of this model is due to the nature of the incompatibility of the model with the topic of analysis or because of the failure of removing confounding variables. Therefore, while future relevant studies should try out different models, they should also delve further into the linear regression model while ensuring that confounding is minimized to reassess its validity.

References

[1] Pi-Sunyer X 2009 The medical risks of obesity. Postgraduate Medicine, 121(6), 21-33.
[2] Kelly T, et al. 2008 Global burden of obesity in 2005 and projections to 2030. International Journal of Obesity, 32(9), 1431-1437.
[3] Safaei M, Sundararajan E A, Driss M, Boulila W and Shapi A 2021 A systematic literature review on obesity: Understanding the causes & Consequences of obesity and reviewing various machine learning approaches used to predict obesity. Computers in Biology and Medicine, 136.
[4] Hu L, et al. (2017). Prevalence of overweight, obesity, abdominal obesity and obesity-related risk factors in southern China. PLOS ONE, 12(9).
[5] Chooi Y C, Ding C and Magkos F 2019 The epidemiology of obesity. Metabolism, 92, 6-10.
[6] Westbury S, Oyebode O, Rens T V and Barber T M 2023 Obesity Stigma: Causes, Consequences, and Potential Solutions. Current Obesity Reports, 12, 10-23.
[7] Nuttall F Q 2015 Body Mass Index Obesity, BMI, and Health: A Critical Review. Nutrition Today, 50(3), 117-128.
[8] Lin X and Li H 2021 Obesity: Epidemiology, Pathophysiology, and Therapeutics. Frontiers in Endocrinology, 12.
[9] Mahmoud R, Kimonis V and Butler M G 2022 Genetics of Obesity in Humans: A Clinical Review. International Journal of Molecular Science, 23(19).
[10] Mathieu-Bolh N 2022 The elusive link between income and obesity. Journal of Economic Surveys, 36(4), 935-968.

DOI: 10.54254/2753-8818/51/2024CH0143

[11] Kim T J and Knesebeck O D 2017 Income and obesity: what is the direction of the relationship? A systematic review and meta-analysis. BMJ Open, 8(1).
[12] Bruce M A, et al. 2007 One size fits all? Race, gender and body mass index among U.S. adults. Journal of the National Medical Association, 99(10), 1152-1158.
[13] Colby S L and Ortman J M 2015 Projections of the Size and Composition of the U.S. Population: 2014 to 2060. US Cencus Bureau.
[14] Levine J A 2011 Poverty and obesity in the U.S. Diabetes, 60(11), 2667-2668.