Evaluating the impact of BMI on Cholesterol

Datasci 203: Lab 2

Rachel Gao, Cole Agard, Charles C. Lucas, Clara Zhu

Contents

1 Introduction

According to the National Institute of Health, as of 2018, nearly one in three adults in the United States are overweight, and more than two in five adults are obese[1]. The increasing prevalence of obesity and its associated health complications has become a public health concern. One of the most significant health risks associated with obesity is elevated cholesterol levels, a leading risk factor for cardiovascular diseases such as heart attacks and strokes.

In this study, we intend to understand the effects of one's weight health on cholesterol. To operationalize weight health, we will use Body Mass Index (BMI), calculated as weight in kilograms divided by height in meters squared, a widely used measurement for screening obesity. In general, a BMI of 18.5 to 25 is considered healthy, while more than 40 is considered severe obesity.

To operationalize cholesterol, we considered the use of total cholesterol and other cholesterol measurements, such as HDL, LDL, and Triglycerides[2], all of which contributes to total cholesterol. However, multiple risk factors can affect HDL and LDL levels, often in opposite directions. Many studies, including The Framingham Study[3], suggest that reviewing HDL and LDL levels separately can provide more valuable information than examining total cholesterol alone. The same study also suggests that cholesterol ratio (total cholesterol divided by HDL) may be an even more informative measure for individuals of certain demographics. Therefore, we believe using cholesterol ratio would be the most appropriate for our study. In general, a cholesterol ratio of below 3.5 is desirable, while greater than 5 is concerning.

In order to answer the question:

> **Does a higher BMI cause a higher cholesterol ratio?**

we analyzed a large dataset of individuals with varying BMI values and cholesterol profiles. Utilizing advanced statistical techniques, such as Large Sample Ordinary Least Squares Regression, we were able to examine the relationship between BMI and cholesterol ratio while accounting for covariates and potential confounding factors. The results of this study will not only contribute to our understanding of the underlying relationship between BMI and cholesterol but also provide crucial information for developing targeted interventions and policies aimed at reducing cholesterol levels and promoting better cardiovascular health.

2 Data and Methodology

We obtained the data from the 2005-2006 National Health and Nutrition Examination Survey (NHANES)[4]. The NHANES conducts annual surveys of individuals across the United States. Each row represents a unique individual identifiable by a unique sequence number. From the dataset, we identified 6360 individuals with valid BMI and cholesterol measurements. We randomly sampled 30% of the data for exploration and model building, while the remaining 70% (4452 individuals) was used to generate all statistics in this study.

Our study focused on the causal effect of BMI on cholesterol ratio. Based on our exploratory plot (Figure 1) and the correlation of 0.34, it is evident that there is a relationship between BMI and cholesterol ratio. Utilizing the tools demonstrated in DATASCI 203 with the consideration of our sample size and the fact

[1] U.S. Department of Health and Human Services. (n.d.). Overweight & Obesity Statistics - Niddk. National Institute of Diabetes and Digestive and Kidney Diseases.

[2] HDL: High-Density Lipoprotein ("good") cholesterol levels. LDL: Low-Density Lipoprotein ("bad") cholesterol levels. Triglycerides: A type of fat that the body uses for energy.

[3] Castelli, W. P., Anderson, K., Wilson, P. W. F., & Levy, D. (2010, July 14). Lipids and risk of coronary heart disease the Framingham Study. Annals of Epidemiology.

[4] United States Department of Health and Human Services. Centers for Disease Control and Prevention. National Center for Health Statistics. National Health and Nutrition Examination Survey (NHANES), 2005-2006. Inter-university Consortium for Political and Social Research [distributor], 2012-02-22. https://doi.org/10.3886/ICPSR25504.v5

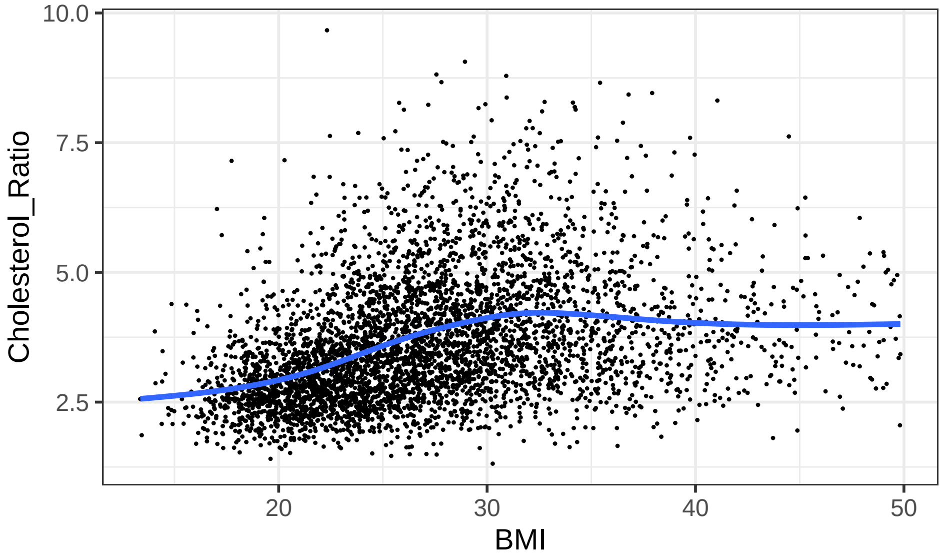

Figure 1: Cholesterol Ratio as a Function of BMI

that there is no strong indication of non-linear relationship based on Figure 1, we developed Large Sample Ordinary Least Squares Regression models of the form:

$$\widehat{Cholesterol}\ Ratio = \beta_0 + \beta_1 \times BMI + Z\gamma$$

where β_1 represents the increase in cholesterol ratio for each unit increase in BMI, Z is a row vector for additional covariates, and γ is a column vector for coefficients.

When considering which covariates to include in our models, we referred to known risk factors for high cholesterol as published by Centers for Disease Control and Prevention[5]. Some known risk factors include type 2 diabetes, age, gender, smoking habits, and drinking habits, in addition to obesity. Obesity is operationalized using BMI and we included glycohemoglobin (%)[6], age, and gender as covariates. We considered incorporating drinking habits as another covariate, but the sample data on drinking habits was self-reported and incomplete. Moreover, the correlation between available drinking habits and cholesterol ratio was low. Consequently, we decided not to include drinking habits as covariate in our models. We also considered the effect of smoking habits on the cholesterol ratio. Of the 6360 individuals in our dataset, only 1248 (20%) individuals confirmed they have smoked more than 100 cigarettes in their life and provided data on their smoking habits. Due to the self-reported nature of this variable, we analyzed the effects of smoking separately from the other covariates as a separate dataset.

The demographics (age and gender) and each physical examination results (cholesterol, BMI, and glycohemoglobin) are stored in separate csv files with each individual identifiable by a unique sequence number. To operationalize cholesterol ratio, we used total cholesterol divide by HDL cholesterol. In cleaning the data, we merged the separate csv files together by the individual sequence number, removed any individuals without valid demographics and examination information, and excluded 39 individuals with BMI above 50, 3 with a cholesterol ratio above 10, and 9 with glycohemoglobin above 12%, as these levels are rare and not representative of the population. Similar to the main dataset, we also split the dataset with smoking habits into explore and confirm, and excluded individuals not representative of the population.

[5] Centers for Disease Control and Prevention. (2023, March 20). Know your risk for high cholesterol. Centers for Disease Control and Prevention. Retrieved April 14, 2023, from https://www.cdc.gov/cholesterol/risk_factors.htm#:~:text=Your%20behaviors%20and%20lifestyle%20choices,can%20lead%20to%20high%20cholesterol.

[6] Also known as Hemoglobin A1c (HbA1c) Test, a blood test used to diagnose diabetes. In general, glycohemoglobin level below 5.7% is normal, while more than 6.5% is diabetes.

3 Results

Table 1 summarizes the results of the four regression models on the main dataset. The key coefficient on BMI was highly statistically significant in all models. In our default model (model 1), each unit of increase in BMI increases the cholesterol ratio by 0.064. Each additional covariate we considered (glycohemoglobin, age, and gender) was also highly statistically significant in all models and contributed to improving the explanatory power of the models. Robust standard errors are presented in parenthesis in the models to account for any possible heteroskedastic errors.

To better understand the results, consider a 40-year-old female with a normal glycohemoglobin of 5.5% and a healthy BMI of 20. According to model 4, her cholesterol ratio would be 3, which is highly desirable. To put this into context, a total cholesterol of 150 mg/dL with HDL of 50 mg/dL would yield a cholesterol ratio of 3. However, if her BMI were to increase to 35, her cholesterol ratio would increase to 3.9. By age 60, assuming she had maintained the same BMI of 35 and glycohemoglobin of 5.5%, her cholesterol ratio would increase to 4. A male individual of the same age with the same BMI and glycohemoglobin would have a cholesterol ratio of 4.5, which would be borderline concerning. And if he is diabetic, his cholesterol ratio would be even higher.

The result emphasizes the significance of maintaining a healthy BMI in order to lower cholesterol ratio, especially as one ages. It also suggests that those with type 2 diabetes should pay special attention to their cholesterol ratio, and that males should be more mindful of their cholesterol ratio than females.

Table 1: Estimated Regressions

	Outcome Variable: Cholesterol Ratio			
	(1)	(2)	(3)	(4)
BMI	0.064***	0.058***	0.055***	0.058***
	(0.003)	(0.003)	(0.003)	(0.003)
Glycohemoglobin		0.199***	0.151***	0.131***
		(0.031)	(0.034)	(0.033)
Age			0.005***	0.005***
			(0.001)	(0.001)
GenderMale				0.510***
				(0.034)
Constant	1.908***	1.006***	1.144***	0.924***
	(0.071)	(0.160)	(0.166)	(0.164)
Observations	4,401	4,401	4,401	4,401
Adjusted R^2	0.113	0.128	0.134	0.178
Residual Std. Error	1.144 (df = 4399)	1.135 (df = 4398)	1.131 (df = 4397)	1.102 (df = 4396)

Note: HC_3 robust standard errors in parentheses.

We also generated a similar regression model by including the number of cigarettes one smoked in a year as an additional covariate. The results show that smoking, BMI, glycohemoglobin, and gender are statistically significant on cholesterol ratio. Among those who have smoked more than 100 cigarettes in their lifetime, holding all other factors constant, an additional 100 cigarettes smoked per year could lead to a 0.04 unit increase in cholesterol ratio. However, we caution that the results from this model may not be generalizable to the population due to limitations in our sample, which excluded individuals who did not report their smoking history or who have never smoked. We, therefore, excluded the model results from Table 1. Further research with more representative samples is needed to draw definitive conclusions on the impact of smoking on cholesterol ratio.

4 Limitations

To produce consistent estimates using Large Sample Ordinary Least Squares Regression models, two assumptions must be met: 1) independent and identically distributed (iid) samples and 2) unique Best Linear Predictor (BLP) exists.

In evaluating the iid assumption, we noted that when collecting data, individuals were sampled from counties across the United States, which could introduce geographic clustering. Additionally, individuals may have referred friends and relatives to complete the survey, violating the independence assumption. However, we recognize that the technicians deliberately oversampled individuals of certain ages and demographics to have a more accurate representation of the United States population, which may reduce some of the iid violations previously mentioned.

In evaluating the unique BLP exists assumption, we analyzed the distribution of all variables and found no evidence of heavy-tailed distributions. We also noted that there is no perfect collinearity within the independent variables as no independent variables were automatically dropped, and there is no evidence of multicollinearity based on the variance inflation factor from our model.

Despite our efforts to include as many covariates as we could, several omitted variables could bias our estimates. For example, people with familial hypercholesterolemia, a genetic disorder, are known to have higher cholesterol ratios than those without. This could result in a positive omitted variable bias, driving the results away from zero and making our estimates overconfident. Other omitted variables, such as drinking habits, can be analyzed similarly. We also considered the possibility of diet and physical activity as omitted variables, but as an individual's diet and physical activity directly contributes to their BMI, the effect of diet and physical activity is already reflected by BMI. Therefore, we do not believe diet and physical activity are omitted variables in our analysis.

We also acknowledge the possibility of reverse causality, where an individual's cholesterol ratio could affect their glycohemoglobin levels. Although the relationship between high cholesterol and diabetes is under debate, we recognize that the positive reverse causality bias could result in overconfident estimates. We do not believe BMI is an outcome variable of any covariates in our model.

5 Conclusion

Our study used a Large Sample Ordinary Least Squares Regression to evaluate the relationship between BMI and cholesterol ratio. Holding all covariates constant, our model predicted a 0.055 to 0.064 unit increase in cholesterol ratio for every unit increase in BMI. We also analyzed the effects of glycohemoglobin, age, and gender on cholesterol ratio, finding that higher glycohemoglobin levels and older age lead to increased cholesterol ratios, and males have higher cholesterol ratios than females.

It's important to note that our adjusted R-squared was below 20%, indicating there may be other factors, such as genetics and family history, with stronger effects on cholesterol ratio. Unfortunately, our sample did not contain reliable information on these variables, which limited our ability to include them in our analysis. Future researchers may consider exploring these factors in more detail.

Overall, we hope that our research provides individuals and physicians with a better understanding of the factors contributing to high cholesterol, which may help develop targeted interventions accurately. Furthermore, the impact of weight loss interventions on cholesterol levels is an area of significant interest, as it can provide insights into practical strategies for improving cardiovascular health among individuals with high BMI.