DATASCI 203 Lab 2

Charles Lucas
Cole Agard
Rachel Gao
Clara Zhu

April 18th, 2023

Berkeley
UNIVERSITY OF CALIFORNIA

Hi everyone, thank you for joining us today, I'm here with my teammates Cole, Rachel, and Clara to go over the lab 2 findings.

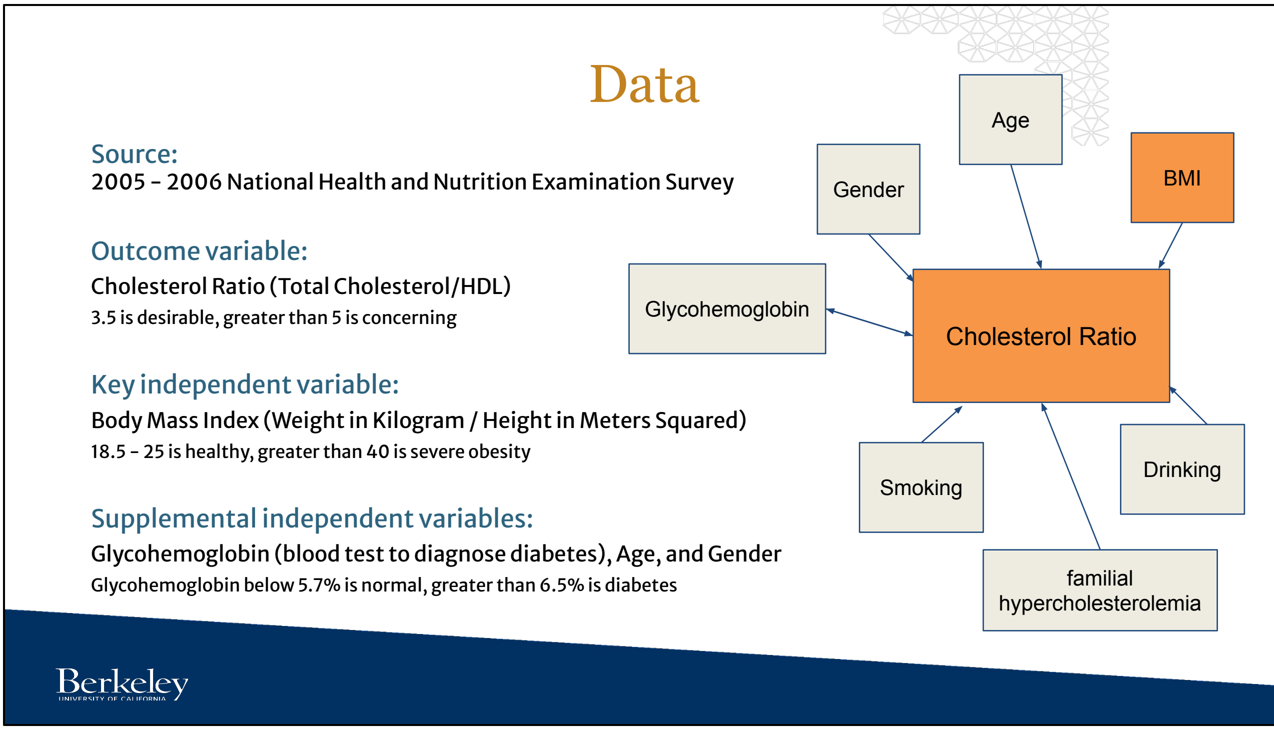
Research Question:

Does a higher BMI cause higher cholesterol?

Various research have suggested that one of the significant health risks associated with obesity is elevated cholesterol levels, a leading risk factor for cardiovascular diseases such as heart attacks and strokes.

This motivated our study to understand the effects of one’s weight health on cholesterol, to operationalize weight health, we used Body Mass Index, and to operationalize cholesterol, we used cholesterol ratio, which studies have suggested to be better than total cholesterol alone. Therefore, we posed the research question “Does a higher BMI cause higher cholesterol?”

I will now turn this over to Cole to go over the dataset we used.



- Our dataset was drawn from the National Health and Nutrition Examination Survey, which is conducted annually to monitor the health and nutritional status of adults and children across the United States. The survey is unique in that it combines interviews and physical examinations. Cholesterol was one of the central aspects of this survey conducted during physical examinations.

- The dataset that we utilized includes survey results from 6000 individuals from across the United States.

- Each row in the dataset represents one individuals response to the survey.

- In our study, primary and supplemental variables were selected based on known risk factors for high cholesterol as published by the Centers for Disease Control and Prevention.
    - LDL (low-density lipoprotein) cholesterol, sometimes

    - called “bad” cholesterol, makes up most of your body's cholesterol.
    - HDL (high-density lipoprotein) cholesterol, sometimes called “good” cholesterol, absorbs cholesterol in the blood and carries it back to the liver.
    - Our outcome variable cholesterol ratio is calculated by total cholesterol divided by HDL, with a value less than 3.5 being desirable. We chose the cholesterol ratio instead of the raw cholesterol measurements because studies in medical literature state it is a better wholistic measure of cholesterol’s negative health implications.
    - Our key independent variable BMI is calculated by weight in kg divided by height in meters squared, generally a BMI between 18.5 and 25 is considered healthy.
    - Our supplemental variables included glycohemoglobin (a marker for diabetes), age, and gender.
    - By accounting for these variables, the study aimed to isolate the impact of BMI on cholesterol ratio, providing a clearer understanding of the relationship between these two variables.

- **We excluded drinking habits, smoking, and familial hypercholesterolemia from our studies.**
    - We noted that the drinking habits are self-reported and incomplete in our dataset and the correlation with cholesterol ratio is low so we did not include as a supplemental variable.
    - Familial hypercholesterolemia is a genetic disorder that has strong relationship with high cholesterol, however our dataset did not contain information on this variable.
    - We generated an additional model with smoking habits, but noted that smoking habits are also self-reported and the dataset only included individuals who have smoked more than 100 cigarettes in their life so the model has

- its limitations. We will go over the limitations and biases due to omitted variables and reverse causality later in the presentation.

- **Now I will pass it of to Rachel to discuss our model and results achieved by our study.**

Model & Results

Increase in cholesterol ratio for each unit increase in BMI

Constant

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

Row vector for additional covariates

Column vector for coefficients

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.

To answer the research question with our data, we developed Large Sample Ordinary Least Squares Regression models of the form as shown on the left side of the slide where beta_1 represents the increase in cholesterol ratio for each unit increase in BMI. We evaluated the iid and unique BLP exists assumptions and noted they can be reasonably assumed to be met.

The results from the 4 regression models can be seen on the right side of the slide. Robust standard errors are presented in parenthesis in the models to account for any possible heteroskedastic errors.

We added one additional covariate for each model. Our key independent variable, BMI, was statistically significant in all 4 models, as was each additional covariate, and all covariates contributed to the explanatory power of the model as can be seen in the increase in adjusted R^2. In our primary model 1, each unit of increase in BMI results in a point zero 6 unit increase in cholesterol ratio.

To better understand the results, let’s 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 with HDL of 50 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, assume 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 health profile 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.

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 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 point zero 4 unit increase in cholesterol ratio. However, the results from that model are not 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 this table.

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.

I will now pass this to Clara who will go over the limitations and conclusions.

Limitation & Conclusion

Omitted Variables Bias:
Familial hypercholesterolemia, drinking habits, etc.

Reverse Causality Bias:
Individual's cholesterol ratio could affect their glycohemoglobin levels

Inspiration for future research:
The adjusted R-squared was below 20%, indicating there may be other factors, such as genetics and family history, with stronger effects on the cholesterol ratio.

Omitted variables, discuss the direction of bias. Discuss why we don't consider diet and exercise habits as omitted variables (already included in BMI)

Reverse causality, discuss the direction of bias.

We do not believe BMI is an outcome variable of any covariates in our model.

Also discuss the conclusion: 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.