Demographic Analysis of Cholesterol Levels amongst Americans using Bayesian Hierarchical Modeling Techniques

Nathan Rosenfeld, Shalin Adhvaryu, Zizhe Wang

Introduction

In the landscape of public health, the role of cholesterol in cardiovascular health cannot be overstated. Effecting as many as 40% of Americans, elevated cholesterol levels are a known harbinger of increased cardiovascular risk, including heart diseases and strokes[1]. This work sets out to explore the intricate dance of cholesterol levels across different demographic terrains, using data gathered by the National Health and Nutrition Examination Survey (NHANES).

While high cholesterol may an issue for all Americans, medical outcomes differ greatly in the United States. Factors like age, gender, and race seem to play a significant role in how these cholesterol levels manifest and fluctuate. By understanding the nuances of how cholesterol levels interact with economic and societal factors, we can use the results of this study to discover which populations might be predisposed to greater cardiovascular risks than others.

Data and Background

The data for this project is from the National Health and Nutrition Examination Survey (NHANES). NHANES is a thorough survey of the health and nutritional statistics and habits of many selected individuals from 30 locations throughout the United States. Since 1999, the survey has been conducted on a biennial basis. The CDC, who conducts the NHANES survey, uses a sample group which is intended to be representative of the demographic makeup of the US population.

The data for NHANES is gathered by setting up 30 testing centers throughout the country. Individuals selected to participate in the survey answer questionnaire forms, come in for interviews, undergo medical examination and allow laboratory tests to be done. Given the steps taken by the CDC during data collection to make sure that it was a good representative sample of the US population, it doesn't seem necessary to "second-guess" the methods or to account for them in the analysis.

The survey is available at: https://www.cdc.gov/nchs/nhanes/index.htm. For each biennial survey (the most recent being 8/2021-8/2023) there are tables containing: demographic data, dietary data, examination data, laboratory data, questionnaire data, and limited access data. For the purposes of this study, we utilized the 2017-18 dataset, as it is the most recent survey that is completed and unaffected by the 2020 Coronavirus pandemic, which harmed data reliability.

NHANES data sets contain 5,000 individuals per year and are provided in XPT or SAS format. These data files were converted to comma-separated values (csv).

There are multiple different measures of cholesterol, but for this study, Low-Density Lipoprote (LDL) cholesterol measured in concentration of milligrams per deciliter (mg/dL) was analyzed. LDL is the "bad" cholesterol which can cause blockages in blood arteries leading to potentially fatal consequences. An LDL level of 150 mg/dL or more is considered borderline and a level above 200 mg/dL suggest high risk of negative health consequences[2].

As mentioned in the introduction, we are mainly interested in demographics and socioeconomic factors and their relationship to cholesterol levels. All respondents who are below the age of 20 and older than 65 were removed from the sample. This study was focused on the adult population, as it is very rare for children to be diagnosed with high cholesterol barring genetic factors. Those above 65 were removed to avoid selection bias, in which older populations appear to be healthier than young and middle aged since the unhealthy individuals have already passed away. Also, any individuals who report being told by their healthcare provider that they have high cholesterol were also removed since many of these individuals are prescribed medication which artificially lowers their cholesterol levels. Using this data we conducted some preliminary analysis on the distribution of cholesterol levels by various predictive variables.

Figure 1 below is the distribution of cholesterol levels by gender. "F" represents females and "M" represents males. Thre are differences between the distributions as the female population has a less likely chance of having a cholesterol level above 150 mg/dL than males.

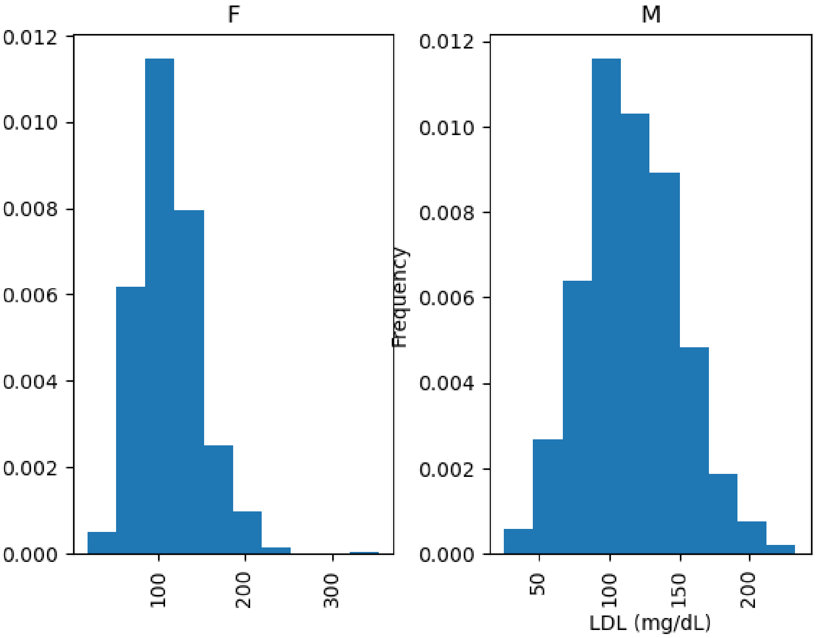

Figure 1. Distribution of LDL Cholesterol by Race

We also investigated the distribution of cholesterol levels by race. This includes the four major racial categories as stated by the US census. Those are: Hispanic (H), Asian (A), African American (AA), and white (W). While it is possible that some individuals can be of both Hispanic and African American, Asian, or white ancestry, those individuals were assigned just a Hispanic race.

In Figure 2 below, we see the different distributions of cholesterol levels. White and African Americans have a more symmetrical distribution around a mean LDL cholesterol of 100 mg/dL, while Hispanic's adults have a high distribution of lower cholesterol level. Asians are strongly distributed with cholesterol levels above 100 mg/dL relative to other racial groups.

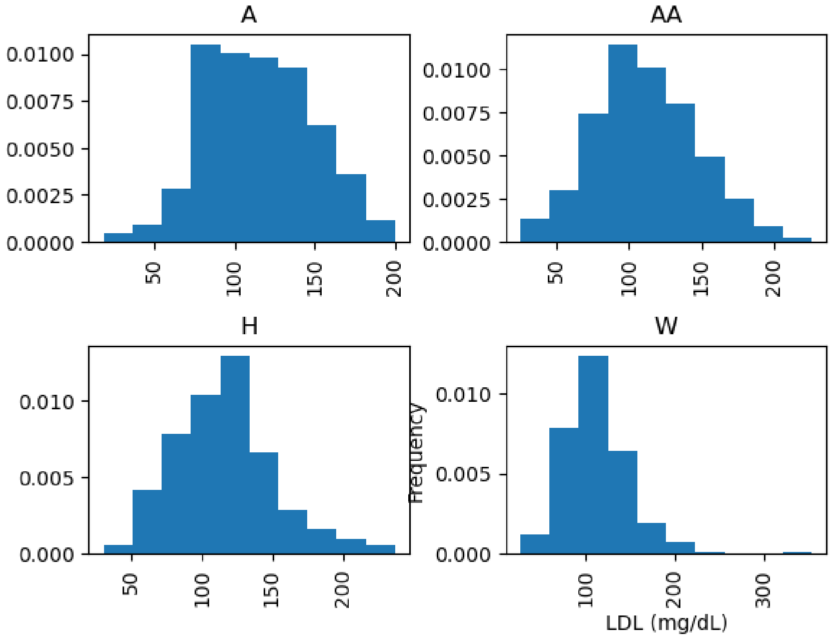

Figure 2. Distribution of LDL Cholesterol by Race

Finally, the distribution of cholesterol between family incomes was examined. Based on the NHANES data, family income was reported as a multiple of the national poverty index. A score less than one indicates a family in poverty, while a score above one means they are above the poverty line. The highest possible score is five. To easily group the data, the floor of the poverty threshold was taken, resulting in six different economic groups. A score between 0-1 was labeled as 0, a score between 1-2 as 1, etc.

From Figure 3, we see that as one's family's financial situation improve, cholesterol levels actually increase. Among the poorest Americans, very few have levels above 125 mg/dL, while those in the highest groups financially, 4 and 5, are the most likely to to have cholesterol levels above 150 mg/dL.

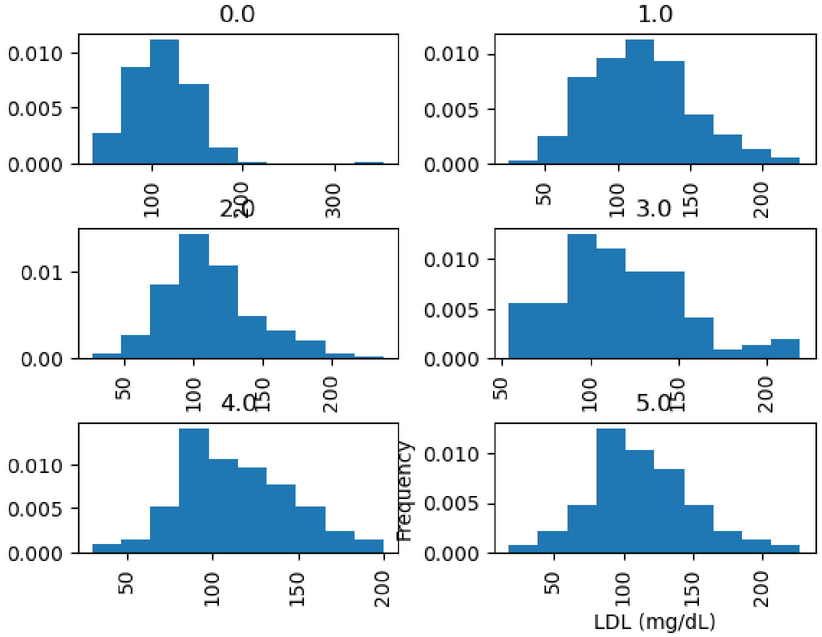

Figure 3. Distribution of LDL Cholesterol by Economic Status

Methodology

Statistical Approach: Univariate Hierarchical Model

Given the hierarchical nature of the NHANES data, with individuals nested within demographic groups, we employ a univariate hierarchical Bayesian model. This model is particularly adept at handling variations within and between groups, thus offering a more nuanced analysis of LDL cholesterol levels in relation to age, gender, and race.

Model Specification

The univariate hierarchical model is specified as follows:

- **Level 1 (Individual-Level Model)**: For each individual i in group j,

  $$LDL_{ij} \sim \mathcal{N}(\theta_j, \sigma_j^2)$$

  where LDL_{ij} is the LDL cholesterol level for the i-th individual in the j-th group, θ_j is the group-specific mean, and σ_j^2 is the group-specific variance.

- **Level 2 (Group-Level Model)**: The group-specific parameters θ_j and σ_j^2 are modeled as:

  $$\theta_j \sim \mathcal{N}(\mu, \tau^2)$$

  $$\sigma_j^2 \sim \text{Inverse-Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$$

  where μ is the overall mean, τ^2 is the between-group variance, and ν_0, σ_0^2 are the shape and scale parameters for the Inverse-Gamma distribution of the group-specific variances.

Prior Setting of the Hyperparameters

The hyperparameters are set with the following priors:

$$\mu \sim \mathcal{N}(\mu_0, \gamma_0^2)$$

$$\tau^2 \sim \text{Inverse-Gamma}(\eta_0/2, \eta_0\tau_0^2/2)$$

$$\nu_0 \sim \text{Gamma}(\alpha_0, \beta_0)$$

$$\sigma_0^2 \sim \text{Inverse-Gamma}(\kappa_0/2, \kappa_0\theta_0^2/2)$$

where μ_0 and γ_0^2 are the mean and variance for the overall mean μ; η_0 and τ_0^2 are the shape and scale for the between-group variance τ^2; α_0 and β_0 are the parameters for the Gamma distribution of ν_0; κ_0 and θ_0^2 are the parameters for the Inverse-Gamma distribution of σ_0^2.

Posterior Distributions

The posterior distributions are derived from the likelihood and the prior distributions. They are computed as follows:

- The posterior of θ_j and σ_j^2 are updated based on the individual-level LDL cholesterol observations and the prior distributions.
- The posterior of μ is updated based on the group-specific means θ_j and its prior.
- The posterior of τ^2 is updated based on the group-specific means θ_j and its prior.
- The posteriors of ν_0 and σ_0^2 are updated based on the group-specific variances σ_j^2 and their priors.

The exact formulation of these posterior distributions typically requires the use of computational methods, such as Markov Chain Monte Carlo (MCMC), due to the complexity of the hierarchical model.

Computational Implementation

We implement the MCMC procedure using a computational tool like Stan or PyMC3, which are well-suited for Bayesian hierarchical models. These tools generate samples from the posterior distributions, allowing us to estimate the model parameters and make probabilistic statements about the LDL cholesterol levels.

Results

Result Values

The following lists the 0.025, 0.5, and 0.975 percentiles for the observed θ_j and σ_j for each group:

Group	θ_j			σ_j		
	0.025	**0.5**	**0.975**	**0.025**	**0.5**	**0.975**
Asian.0	106.4295	112.4757	119.8164	23.568	35.754	64.3456
Black.0	103.9585	107.2877	110.6724	11.0144	13.1470	15.9115
Latino.0	107.3670	111.0793	114.6548	12.6763	15.2424	18.9809
White.0	110.2717	114.1162	118.5513	14.0634	17.0141	21.3861
Asian.1	104.7927	110.9290	116.3658	18.0038	23.5552	32.5250
Black.1	105.5335	109.1111	112.4952	11.8201	14.2061	17.3683
Latino.1	112.1460	115.1571	118.3810	11.8556	13.9211	16.6856
White.1	109.4097	112.3074	115.2327	10.7903	12.8125	15.4937
Asian.2	104.4504	110.7658	116.4096	19.5825	26.3385	37.6769
Black.2	106.0600	110.6711	115.1122	13.2440	16.7932	21.8911
Latino.2	106.2047	111.1257	115.8212	15.3122	19.3237	24.9728
White.2	105.3557	109.9130	114.1974	12.7475	15.9892	20.6349
Asian.3	107.3123	112.8802	119.0734	18.3148	23.9847	33.9152
Black.3	107.9721	113.5803	121.0481	17.9901	24.9727	37.8477
Latino.3	107.8429	113.8454	122.8339	23.5696	33.2511	54.0582
White.3	109.7084	114.3808	120.0253	15.0281	19.0333	25.0769
Asian.4	105.5578	111.4338	116.9787	18.2195	23.6886	32.5001
Black.4	103.4968	110.6330	116.5889	20.0896	28.6913	45.2477
Latino.4	105.5025	111.1963	117.0050	16.7515	22.7649	32.7841
White.4	107.4937	112.9367	119.1977	16.741	22.4111	32.0356
Asian.5	108.4894	112.5886	116.9484	13.2769	16.5773	21.4738
Black.5	104.4487	111.0255	117.1643	18.3295	25.6883	40.0795
Latino.5	108.1233	113.5886	120.1535	18.3475	24.1629	33.3853
White.5	106.6866	110.5182	114.1620	12.2125	14.9502	18.8950

We also have the following inter-quartile range for the general parameters μ, τ^2, σ_0^2, and ν_0:

	0.25	**0.5**	**0.75**
μ	111.1331	111.8313	112.5062
τ^2	5.8658	9.0925	13.8096
σ_0^2	289.3648	321.4377	356.6872
ν_0	3	4	5

Model Diagnostics

The bar plots in Figure 4 show that the model reached stationarity over the 5000 iterations which were run. As the IQR is basically unchanged for each group of 500 iterations we can see that the algorithm reached stationary values for the four shown parameters.

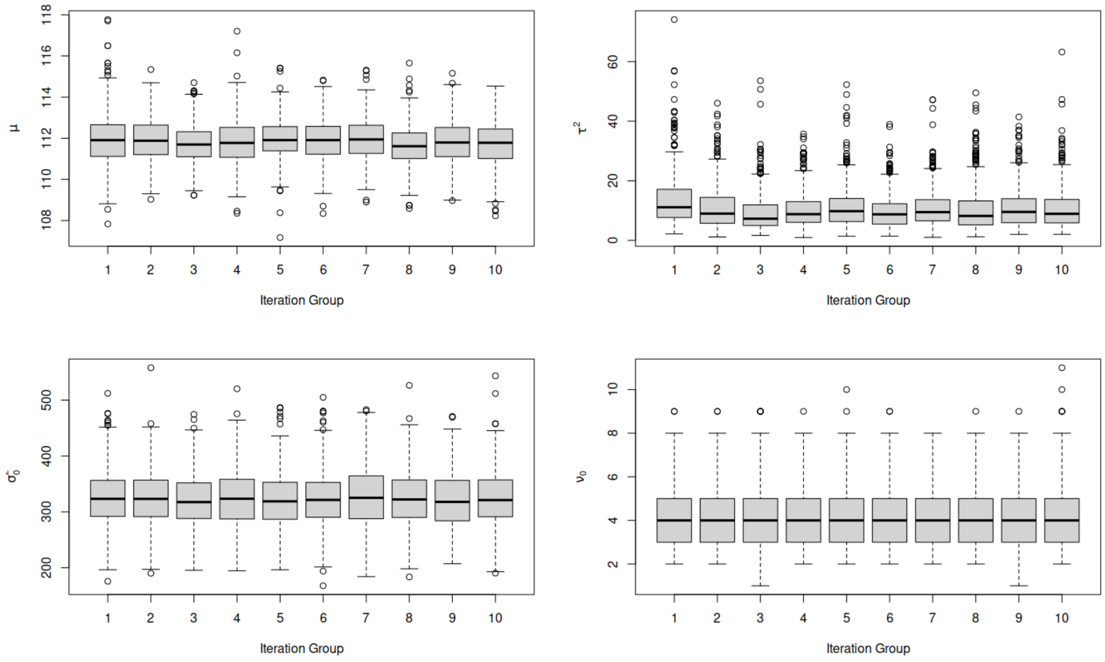

Figure 4. Stationarity Plots

.

Advantage of Bayesian Hierarchical Model

The hierarchical approach that we used has a few advantages over the standard frequentist method when working with our data.

For one thing, several of the groups we are studying have a rather small sample size. The approach that we used allows us to use the information from the other groups to add to our knowledge of the distribution of these groups' distributions. For instance, Figure 5 shows the posterior distribution of θ_j for the group of Asian women with income between 4-5 times the poverty line. The blue line shows a distribution of random values following a normal distribution such that the mean is the sample mean of this group and the standard deviation is that of the group divided by the square root of the sample size. In other words, it is the sampling distribution of a distribution with the parameters equal to the sample parameters.

The red line shows the observed values of θ in our model. The dots in the plot show the observed values. So, this model gives us a far more narrow range for the mean of the distribution for this group. The uncertainty is lessened because the observed sample mean is very close to the mean of all women from all groups.

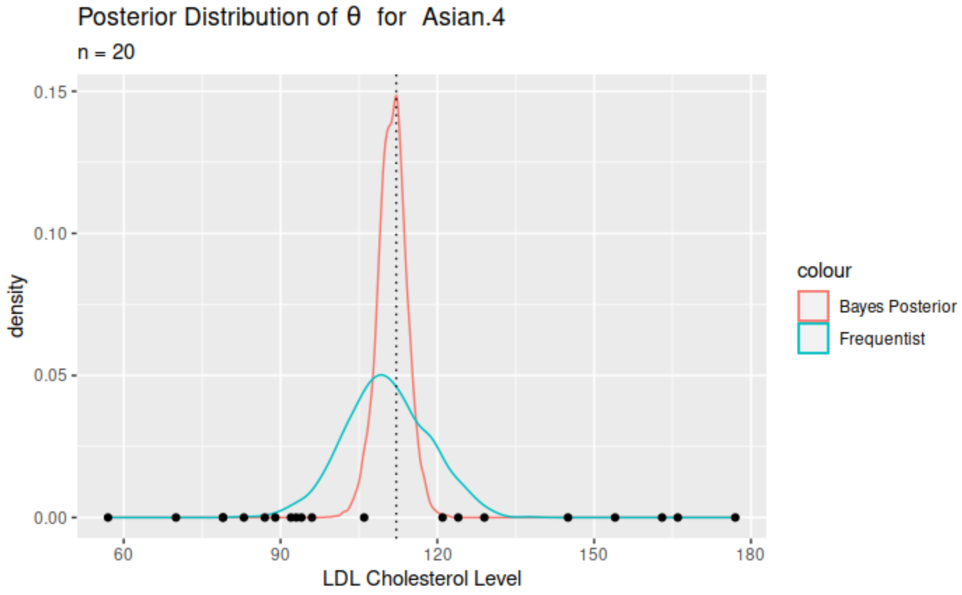

Figure 5

Another advantage to this approach is to make the model more robust to outliers. Figure 6 shows the group of Latino women with incomes from 3-4 times the poverty line. As shown, the very small group has a sample mean which is made much larger due to three very large observations. By incorporating the information from all other groups, the effect of these observations is much lessened, resulting in a group distribution which seems much more credible.

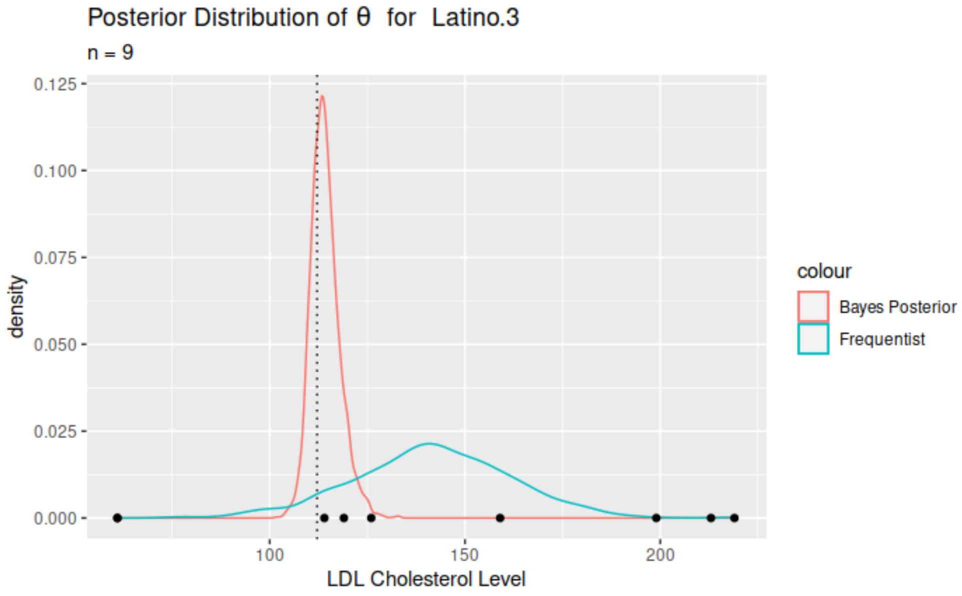

Figure 6

.

Lastly, this model makes it easier to argue more convincingly that there are differences by group. For example, if we were to use a two-sided t-test to compare `Black.0` vs. `Latino.1`, we get a test statistic of -1.792 with an associated p-value of 0.07579. Whereas, by looking at the table of values, we see that the ranges of the middle 95 percentiles of their respective θ_j's are completely disjoint. So, the Bayesian hierarchical model provides a stronger argument for the groups being different in the distributions of their parameters.

Analysis of Hierarchical Model for LDL Levels

Overview

The results presented focus on a hierarchical model examining LDL cholesterol levels across various groups. The analysis is structured around two key components: (1) group-specific parameters (θ_j and σ_j) and (2) general parameters (μ, τ^2, σ_0^2, and ν_0).

Group-Specific Parameters

1. **LDL Level (θ_j):**

    - The median values (0.5 percentile) of θ_j across groups range from approximately 107.29 (Black.0) to 115.16 (Latino.1), indicating a variance in the central tendency of LDL levels among different groups.
    - Groups such as Asian.0, Latino.1, and White.0 exhibit higher upper percentiles (0.975), suggesting a possible skew towards higher LDL levels in these populations.
    - The spread between the 0.025 and 0.975 percentiles varies, with the Asian and White groups generally showing wider intervals, implying greater heterogeneity in LDL levels within these groups.

2. **Standard Deviation of LDL Level (σ_j):**

    - The median standard deviations range notably, with the highest observed in Asian.0 (35.754) and the lowest in White.1 (12.8125), reflecting diverse variability in LDL measurements within groups.
    - The broader intervals in σ_j for groups like Asian.0 and Latino.3 suggest more variability in LDL levels compared to groups like Black.0 and White.1.

General Parameters

1. **Overall Mean LDL Level (μ):**

    - The interquartile range (IQR) for μ is fairly narrow (111.1331 to 112.5062), indicating a relatively stable average LDL level across the entire sample.

2. **Between-Group Variance in Mean LDL Level (τ^2):**

    - The IQR for τ^2 (5.8658 to 13.8096) suggests some variability in mean LDL levels between groups, but not excessively high, implying that while groups differ, these differences are not extremely pronounced.

3. **Baseline Variance in LDL Levels (σ_0^2):**

    - The IQR for σ_0^2 (289.3648 to 356.6872) indicates a considerable level of inherent variability in LDL levels across the entire population.

4. **Degrees of Freedom for the Inverse-Wishart Prior (ν_0):**

    - The values of ν_0 (3 to 5) impact the prior distribution for σ_j, which in turn influences the estimation of group-specific variances.

Implications

- **Group Differences**: The variability in both θ_j and σ_j across different groups underscores the importance of considering demographic factors (like ethnicity) in assessing cardiovascular risk factors such as LDL cholesterol levels.

- **Modeling Considerations**: The spread in the general parameters, particularly in τ^2 and σ_0^2, highlights the complexity of modeling LDL levels. The hierarchical model accounts for both individual and group-level variations, which is crucial for accurate and comprehensive analysis.

- **Clinical Relevance**: Understanding the distribution and variability of LDL levels across different groups can aid in tailoring public health interventions and personalized treatment strategies.

Limitations

Our study, while comprehensive in its methodology, encounters several notable limitations. Firstly, the reliance on the NHANES dataset for data sourcing limits the ability to extrpolate our findings beyond the demographic characteristics specific to the United States population. This specialization may not accurately represent global populations.

Additionally, the complexity of the hierarchical models, particularly in the Bayesian context, presents computational challenges. This complexity necessitates a careful consideration of convergence issues in Markov Chain Monte Carlo (MCMC) methods, which are pivotal for the robustness of our findings.

Another limitation is the cross-sectional nature of the NHANES data. This restricts our ability to draw causal inferences or to observe trends over time, as cross-sectional data provides only a snapshot view at one point in time.

Furthermore, our model, while inclusive of key demographic variables, does not account for other potential confounders such as lifestyle, dietary habits, and genetic factors. The exclusion of these variables could influence the outcomes and interpretations of our study.

Lastly, the statistical assumptions inherent in our hierarchical model, including the normality of distributions and the choice of priors, could potentially impact the results. These assumptions are necessary for the model framework but may introduce biases or inaccuracies.

Future Work

Considering these limitations, future research should take several directions to enhance the understanding and applicability of the study. It would be beneficial to incorporate data from diverse geographical and cultural contexts. Such inclusion would help enhance the global applicability and relevance of the findings, providing a more comprehensive perspective.

Conducting longitudinal analysis is also crucial for future research. Such analysis would offer deeper insights into the causal relationships and temporal trends in LDL cholesterol levels, moving beyond the limitations of cross-sectional data.

Expanding the range of variables included in the analysis is another important avenue. Including factors such as lifestyle choices, socioeconomic status, and genetic information could offer a more holistic and nuanced understanding of the dynamics influencing LDL cholesterol levels.

Exploring advanced computational methods, particularly in the realm of Bayesian inference, could improve the efficiency and scalability of the analysis. This would address some of the computational challenges inherent in hierarchical modeling.

Finally, conducting rigorous validation and comparison with alternative statistical models would greatly strengthen the robustness and reliability of the findings. This step is essential for confirming the validity of the conclusions drawn and for establishing a firm basis for future research and application.

References

1. Rauf, D. (n.d.). Over 40 percent of Americans with high cholesterol unaware they have it. EverydayHealth.com. https://www.everydayhealth.com/high-cholesterol/more-than-four-in-ten-adults-dont-know-they-have-high-cholesterol/
2. Hopkins Medicine, J. (2020, December 4). Lipid panel. Johns Hopkins Medicine. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/lipid-panel