Unveiling Total Cholesterol Levels: Insights into key factors

Yingzhen Wang, Tong Liu, Gabriel Alwan, Miranda Qinglan Ouyang, Jackie Lindstrom

Abstract

High-density lipoprotein cholesterol (HDL-C), often referred to as “good” cholesterol, is responsible for the transport of lipids. Higher levels of HDL cholesterol has been shown to be associated with better health outcomes, such as lower risk of cardiovascular disease(2). Therefore, it is important to investigate factors that may affect levels of HDL cholesterol. This paper discusses the construction of a predictive model for predicting an individual’s total HDL cholesterol in mmol/L using data from The National Health and Nutrition Examination Survey (NHANES). The significant variables retained in the final model after using AIC backwards stepwise elimination are gender, race, depression status, average systolic blood pressure, smoking status, age grouped into ten-year categories, and the log transformed days of alcohol consumption over the past year, along with the interaction between age and depression, and the interaction between age and blood pressure. While the model provides insights into predicting HDL- cholesterol levels, accuracy is limited by a low multiple R^2 of 0.1215, so only 12.15% of the variation in HDL cholesterol is explained by the model. Strengths and limitations of this predictive model are further addressed in the paper.

Background

High-density lipoprotein cholesterol (HDL-C), often referred to as “good” cholesterol, is responsible for the transport of lipids. It has been established through research that HDL-cholesterol has anti-atherosclerotic effects, meaning that it prevents the accumulation of cholesterol on artery walls(2). This happens through a combination of HDL-cholesterol capturing and removing cholesterol, antioxidative properties, and anti-inflammatory properties(5). Therefore, maintaining adequate HDL-cholesterol levels is essential to health.

More specifically, high HDL-cholesterol has been associated with numerous health outcomes, including lowering the risk of cardiovascular disease (5). In addition, lower HDL-cholesterol levels may increase the risk of Type 2 Diabetes(4). It is valuable to create a model for predicting HDL-cholesterol levels, as this could be an important indicator of a person’s cardiovascular disease risk, Type 2 Diabetes risk, and overall health.

There are numerous established associations between different health related and demographic variables and HDL-cholesterol. For example, it has been shown that the male individuals are associated with lower levels of HDL-cholesterol than female individuals, and older women had lower HDL-cholesterol levels(1). Additionally, BMI has been shown to be negatively correlated with HDL-cholesterol (7). Cigarette smoking and lack of physical activity is also associated with higher HDL-cholesterol levels and dyslipidemia, a term used to characterize either the presence of excessive LDL cholesterol levels or low HDL-cholesterol levels(3). The research into these variables informed our decision to investigate variables such as BMI, gender, smoking, and physical activity in our preliminary model.

Materials

Study Population:

The study population for this analysis comes from The National Health and Nutrition Examination Survey, abbreviated NHANES. This is an ongoing study conducted by the Center for Disease Control starting in 1959 (6). The study population is people living in the United States, specifically the civilian noninstitutionalized population. Data is collected through a household survey. Many variables are collected on health and nutrition.

Variables of Interest

We are mostly interested in lifestyle factors that have a possible relationship with HDL-cholesterol. We are ultimately interested to see if a healthy lifestyle is linked to desirable HDL-cholesterol levels. Specifically, we study the variables of smoking status, alcohol consumption, and physical activity. However, we also include other non-lifestyle factors (age, race, gender, depression, blood pressure, and BMI) in order to control the model. The literature provided above aims to find a relationship between HDL-cholesterol and heart disease(2). They mention how a healthy diet was linked to high HDL-cholesterol. This prompts us to investigate other decisions people make outside of eating healthy. These decisions are presumably linked to one's cardiovascular health. We want activities that are both good for your heart (physical activity) and bad (alcohol consumption, smoking).

Predictor variables include:

- **Demographics:** Gender (categorical) and Race.
- **Behavioral Factors:** Smoking history (binary variable) and number of days alcohol has been consumed over the past year
- **Health Indicators:** Systolic blood pressure (BPSysAve), Body Mass Index (BMI), and depression status (categorized as None, Several, or Most days of feeling depressed).
- **Physical Activity:** Number of physically active days reported by the individual.
- **Age Categories:** Participants were categorized into distinct age groups (20–30, 30–40, etc.) for further analysis.

Outcome

The primary outcome variable was total HDL- cholesterol (TotChol), measured as a continuous variable and measured in mmol/L. We use statistical methods on the data in order to get a better understanding of the underlying population. In theory, an unbiased population can help us accurately come to conclusions about the outcome, which in this case is HDL-cholesterol. Using a sample that is both large in size and heterogeneous can help us come to accurate statistical conclusions on causation. Additionally, proper sampling techniques must be used (e.g. random sampling techniques).

Methods

1. Data Preparation and Cleaning

The initial dataset was drawn from NHANES. To ensure clarity and consistency, several preprocessing steps were undertaken:

a. Data Filtering and Cleaning

We restricted the sample to individuals aged 20 years and above to ensure the suitability of certain variables (e.g., smoking status).

We removed observations and variables with excessive missing data and kept only those essential for our analysis: TotChol as the outcome and selected predictors including Gender, Race1, Depressed, BPSysAve, Smoke100, Age_Category, PhysActiveDays, BMI, and AlcoholYear.

Variables like Depressed and Smoke100 were converted into categorical factors to facilitate their use in the regression model.

b. Variable Recoding and Transformation

To capture potentially nonlinear relationships with age, the Age variable was grouped into categories (Age_Category) such as 20–30, 30–40, 40–50, 50–60, and 60–80 years, with age 60-80 as the reference group. This categorization allowed us to examine whether the impact of other factors on TotChol differs by age group.

Additionally, variables exhibiting skewed distributions (e.g., BMI and AlcoholYear) were log-transformed. BMI was replaced by BMI_log, and AlcoholYear was replaced by AlcoholYear_log. These transformations aimed to achieve more symmetric distributions and ensure that regression assumptions were better met.

2. Preliminary Exploratory Data Analysis

Before modeling, we performed an exploratory analysis to understand the data structure and relationships between variables:

a. Basic Distributions and Visualizations:

Histograms and bar plots for Age, Age_Group, and Age_Category were created to visualize the distribution of participants across different age brackets. Likewise, histograms of continuous variables (TotChol, BPSysAve, PhysActiveDays, BMI, AlcoholYear) were examined to identify normality, skewness, or outliers.

b. Categorical Variables and TotChol:

Mean bar plots were generated to compare TotChol levels across different categories of Gender, Race1, Depressed, Smoke100, and Age_Category. This helped us preliminarily identify which categorical factors might influence TotChol levels and in what direction.

c. Correlation Analysis:

To prevent multicollinearity issues, we explored the correlation matrix of continuous predictors (BPSysAve, BPDiaAve, BMI, AlcoholYear, PhysActiveDays). Visualization using correlation heatmaps revealed a high correlation between BPSysAve and BPDiaAve. To avoid redundant predictors, we retained only BPSysAve for subsequent modeling.

3. Model Building and Diagnostics

a. Initial Model Construction:

We began by fitting a comprehensive linear model including all main effects of interest.

TotChol served as the response variable. Predictors included Gender, Race1, Depressed, BPSysAve, Smoke100, Age_Category, PhysActiveDays, AlcoholYear_log, and BMI_log.

b. Model Diagnostics and Assumption Checks:

Multicollinearity: Variance Inflation Factors (VIFs) were calculated to ensure that no severe multicollinearity was present.

Normality and Linearity of Residuals: Residual histograms, Q-Q plots, and added-variable plots were examined to confirm that residuals were approximately normally distributed and that relationships were adequately linear.

Independence and Homoscedasticity: The Durbin-Watson test helped assess residual independence, while residual vs. fitted value plots checked for heteroscedasticity. Overall, no significant patterns or violations were observed.

Outliers and Leverage Points: Influence diagnostics were conducted to identify any unusual observations that might disproportionately affect the model fit.

4. Model Refinement and Feature Selection

After confirming the initial model's viability, we refined the model through:

a. Backward Stepwise Elimination:

Stepwise regression was used to remove non-significant predictors, yielding a more parsimonious model that still captured the essential features influencing TotChol.

b. Interaction Terms:

Theoretical and empirical considerations suggested the potential importance of interactions. For example, adding interactions between Age_Category and Depressed, as well as between Age_Category and BPSysAve, allowed us to see if the effect of depression status or blood pressure varied across different age groups. Incorporating these interaction terms improved the model's explanatory power and relevance.

Following these refinements, the final model demonstrated improved statistical robustness and offered better insights into subgroup effects and factor combinations influencing TotChol.

5. Model Validation and Predictive Performance

To evaluate the model's predictive capability, we split the data into training and testing sets:

a. Data Partitioning:

Approximately half of the data was used to train the model, while the remaining half served as a test set to validate predictive accuracy, thereby mitigating overfitting risks.

b. Performance Metrics:

Using the trained model to predict TotChol in the test set, we calculated the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics quantified how closely our model's predictions approximated actual values.

c. Comparing Predicted vs. Actual Values:

By plotting the predicted TotChol against the actual TotChol values, we visually assessed the model's accuracy. Predictions aligning closely with the reference line (y = x) indicated a strong predictive performance.

Statistical Software

All analyses were conducted in R, version 4.3.1, with packages for visualization, regression diagnostics, and model selection.

This methodological framework described above ensured a robust examination of the associations between total cholesterol and the selected predictors while addressing potential confounding and interaction effects.

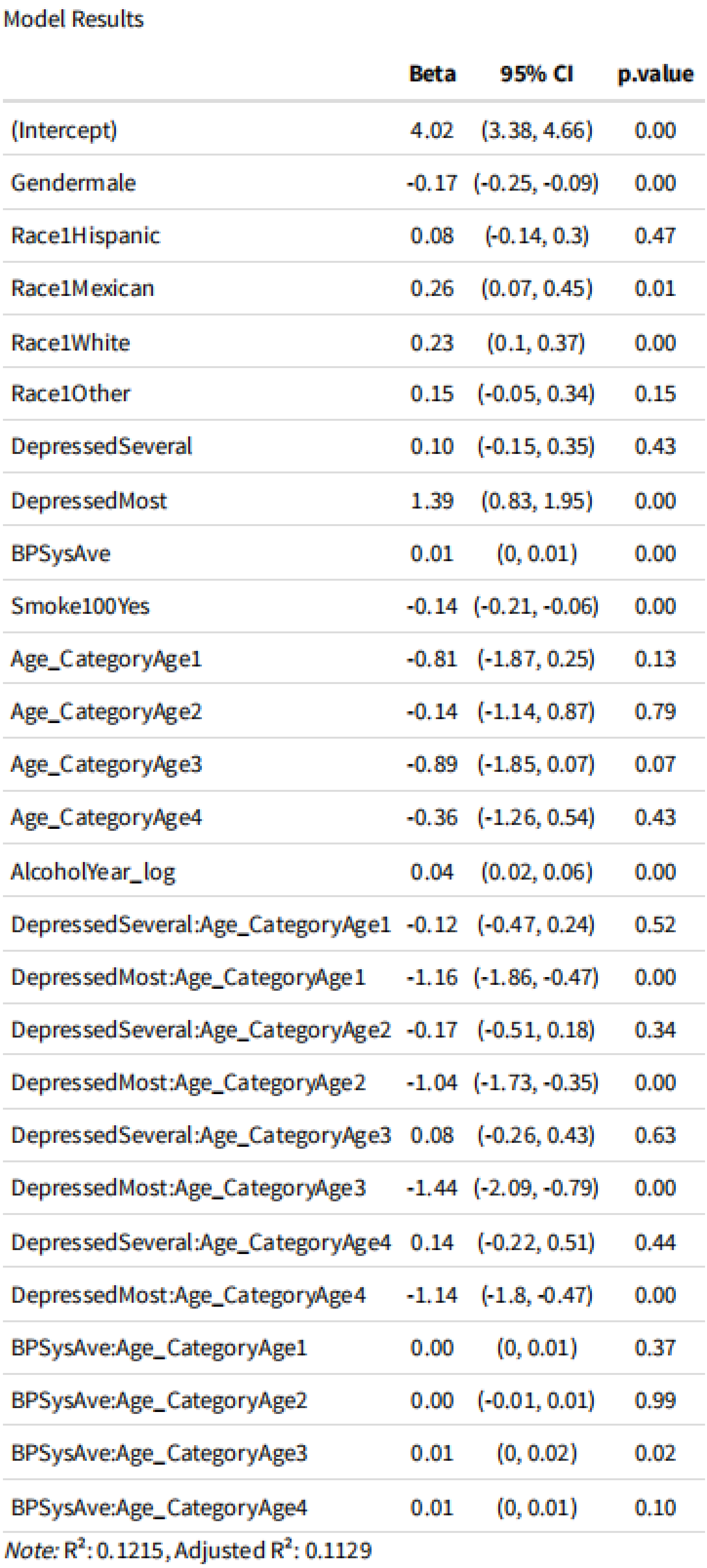
Results

The final model displayed below in Equation (1) and shown in Table 1 shows that we include the main effects of gender, race, depression, blood pressure, smoking, age, alcohol, the interaction between age and depression, and the interaction between age and blood pressure.

$$\begin{aligned}
&\hat{Y}_i = \beta_0 + \beta_1 \mathit{Gender}_i + \beta_2 \mathit{RaceHispanic}_i + \beta_3 \mathit{RaceMexican}_i + \beta_4 \mathit{RaceWhite}_i \ \beta_5 \mathit{RaceOther}_i + \\
&\beta_6 \mathit{DepressedSeveral}_i + \beta_7 \mathit{DepressedMost}_i + \beta_8 \mathit{BPSysAve}_i + \beta_9 \mathit{Smoke100Yes}_i + \\
&\beta_{10} \mathit{Age\_CategoryAge1}_i + \beta_{11} \mathit{Age\_CategoryAge1}_i + \beta_{12} \mathit{Age\_CategoryAge1}_i + \\
&\beta_{13} \mathit{Age\_CategoryAge1} + \beta_{14} \mathit{AlcoholYear\_log} + \beta_{15} \mathit{DepressedSeveral}_i * \mathit{Age\_CategoryAge1}_i + \\
&\beta_{16} \mathit{DepressedMost}_i * \mathit{Age\_CategoryAge1}_i + \beta_{17} \mathit{DepressedSeveral}_i * \mathit{Age\_CategoryAge2}_i + \\
&\beta_{18} \mathit{DepressedMost}_i * \mathit{Age\_CategoryAge2}_i + \beta_{19} \mathit{DepressedSeveral}_i * \mathit{Age\_CategoryAge3}_i + \\
&\beta_{20} \mathit{DepressedSeveral}_i * \mathit{Age\_CategoryAge3}_i + \beta_{21} \mathit{DepressedSeveral}_i * \mathit{Age\_CategoryAge4}_i + \\
&\beta_{22} \mathit{DepressedSeveral}_i * \mathit{Age\_CategoryAge4}_i + \beta_{23} \mathit{BPSysAve}_i * \mathit{Age\_CategoryAge1}_i + \\
&\beta_{24} \mathit{BPSysAve}_i * \mathit{Age\_CategoryAge2}_i + \beta_{25} \mathit{BPSysAve}_i * \mathit{Age\_CategoryAge3}_i + \\
&\beta_{26} \mathit{BPSysAve}_i * \mathit{Age\_CategoryAge4}_i + \epsilon_i
\end{aligned} \tag{1}$$

The interaction terms were included in the model in order to better explain any age linked patterns in depression and blood pressure. This result is expected as we would not expect the effects of all covariates to be consistent across different age groups. The inclusion of interaction terms gives us a more effective and sophisticated model. It allows us to get a better understanding of underlying relationships between covariates. Noticeably, our original model was close to being correct as the backward selection procedure only eliminated BMI and physical activity variables. Below we have the R-output of the summary of our final model.

Table 1. Final model results post analysis.



Model Results

	Beta	95% CI	p.value
(Intercept)	4.02	(3.38, 4.66)	0.00
Gendermale	-0.17	(-0.25, -0.09)	0.00
Race1Hispanic	0.08	(-0.14, 0.3)	0.47
Race1Mexican	0.26	(0.07, 0.45)	0.01
Race1White	0.23	(0.1, 0.37)	0.00
Race1Other	0.15	(-0.05, 0.34)	0.15
DepressedSeveral	0.10	(-0.15, 0.35)	0.43
DepressedMost	1.39	(0.83, 1.95)	0.00
BPSysAve	0.01	(0, 0.01)	0.00
Smoke100Yes	-0.14	(-0.21, -0.06)	0.00
Age_CategoryAge1	-0.81	(-1.87, 0.25)	0.13
Age_CategoryAge2	-0.14	(-1.14, 0.87)	0.79
Age_CategoryAge3	-0.89	(-1.85, 0.07)	0.07
Age_CategoryAge4	-0.36	(-1.26, 0.54)	0.43
AlcoholYear_log	0.04	(0.02, 0.06)	0.00
DepressedSeveral:Age_CategoryAge1	-0.12	(-0.47, 0.24)	0.52
DepressedMost:Age_CategoryAge1	-1.16	(-1.86, -0.47)	0.00
DepressedSeveral:Age_CategoryAge2	-0.17	(-0.51, 0.18)	0.34
DepressedMost:Age_CategoryAge2	-1.04	(-1.73, -0.35)	0.00
DepressedSeveral:Age_CategoryAge3	0.08	(-0.26, 0.43)	0.63
DepressedMost:Age_CategoryAge3	-1.44	(-2.09, -0.79)	0.00
DepressedSeveral:Age_CategoryAge4	0.14	(-0.22, 0.51)	0.44
DepressedMost:Age_CategoryAge4	-1.14	(-1.8, -0.47)	0.00
BPSysAve:Age_CategoryAge1	0.00	(0, 0.01)	0.37
BPSysAve:Age_CategoryAge2	0.00	(-0.01, 0.01)	0.99
BPSysAve:Age_CategoryAge3	0.01	(0, 0.02)	0.02
BPSysAve:Age_CategoryAge4	0.01	(0, 0.01)	0.10

Note: R^2: 0.1215, Adjusted R^2: 0.1129

After performing residual analysis on the final model, we notice that the underlying assumptions for linear regression (LINE) have been met. The residual plot demonstrates homoscedasticity, meaning that the residuals are randomly scattered and have constant variation (e.g. no funneling effect). We also conclude from the histogram and Q-Q plot that the residuals approximate a normal distribution. Thus, we are confident in the fit of the model to the data. We also explore the prediction accuracy. Indeed, the values of mean squared error (MSE), root mean squared error (RSME), and mean absolute error (MAE) are relatively low, indicating a robust model. Specifically, we get MAE = 0.7726818, MSE = 1.016913, and RMSE = 1.008421. We expand on this in our discussion below.

Conclusions and Discussions

Using the model prediction process of backward selection we are able to get an appropriate linear regression model to predict High-density lipoprotein cholesterol (HDL-C). The significant predictors from this process were gender, race, depression, smoking status, blood pressure, and age. We also noticed that age interacting with depression and age interacting with blood pressure were significant and thus included in the model.

We'll evaluate some of the limitations of the results first. Even though we were able to find significant interaction terms, our initial model didn't include interactions. The interaction terms with age were only considered after the model. So, when performing backward selection we may have missed other possible significant interactions. In future analyses we consider other possible interactions in the initial model. Future analysis could explore more of the background into these variables and come up with possible interactions to include in the initial model. In general, focusing solely on main effect terms may have made our final model too simple. Since we are dealing with a lot of complex variables they may have some relation to each other. This may indicate that there are multicollinearity issues within the final model. However, as discussed above, we did check for multicollinearity and since the variance inflation factors (VIF) of the variables we selected are close to 1, it suggests that every variable is strongly uncorrelated among every predictor and the remaining predictor variables and hence the variance of the estimated regression coefficients are unlikely to be inflated.

According to the MAE (0.7726818), MSE (1.016913), RMSE (1.008421) given in the result, indicating that the final model provides reasonably accurate predictions with minimal deviation from the actual results. This suggests that the final model is well-suited for the task and can be considered reliable. Finally, we visualized the association between actual values and predicted values, Figure 1 shows a linear association between them.

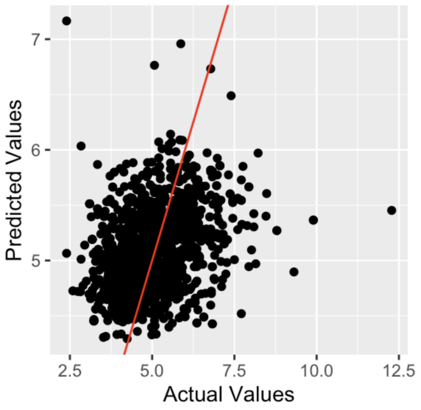
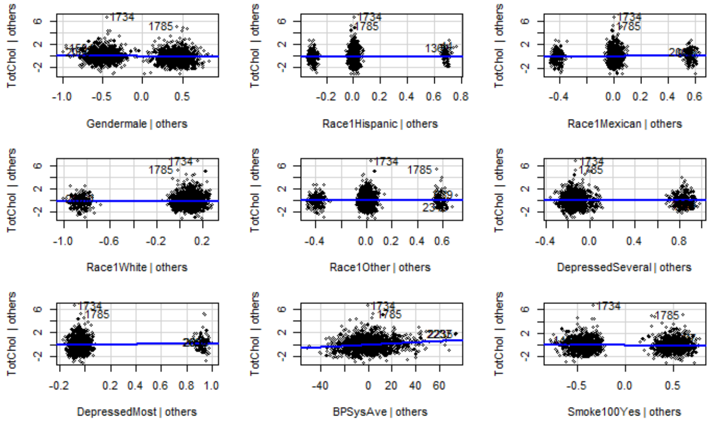


Figure 1. Plot of predicted vs actual values.

While we discussed the downsides of not including interaction terms in the original model, we specifically did this for many reasons. The main reason is, the original model can be used to check linearity, independence, normality and constant variance assumptions of this linear regression model. We used partial regression plots to check linearity assumption. Only the predictor BPSysAve shows a clear linear trend. Other plots do not show curved patterns or deviations from linearity. The plots of variables we need to add into our model later do not show curved or deviation from linearity either. From the Durbin-Watson test statistic of 1.398 and extreme small p value, it strongly indicates that there is a positive autocorrelation in the errors of our original model.



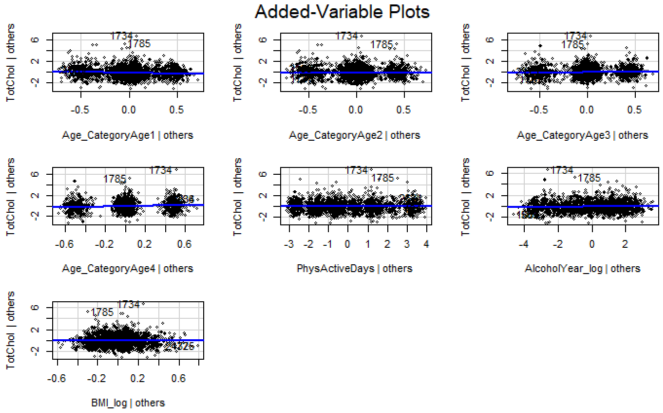
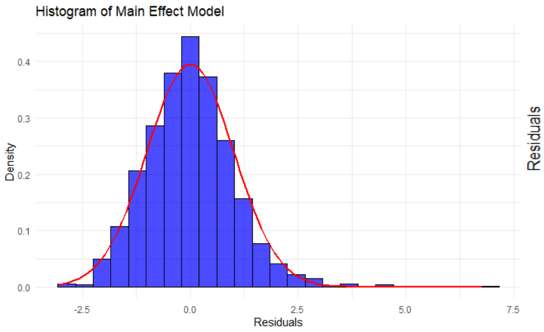


Figure 2. Partial Regression plots. **Figure 3.** Added Variable plots. .

Two approaches are applied to check the normality assumption: histogram and Q-Q plot. The histogram shows a shape of a zero-mean normal distribution which is symmetric, bell-shaped and light tailed and the Q-Q plot shows a true normal distribution. From the residuals vs fitted plot, the diagram is randomly scattered but there is no obvious cone shape shown.

Hence, linearity and normality assumptions are met while independence and constant variance assumptions are not.



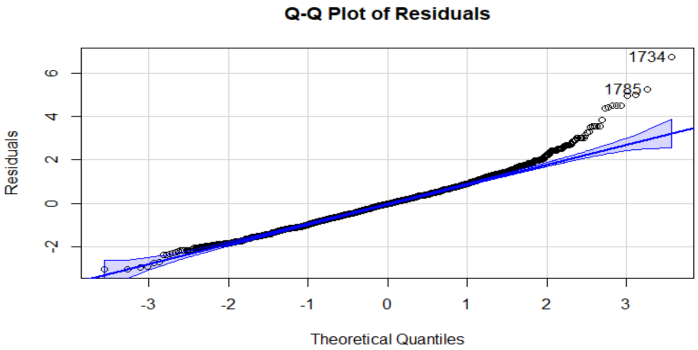
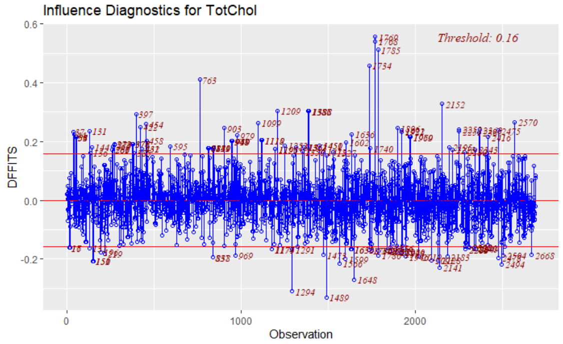


Figure 4. Histogram of Main Effects Model **Figure 5.** Q-Q Plot of Residuals main effects



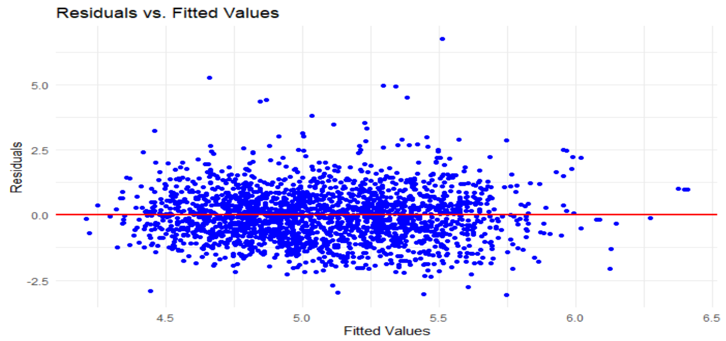


Figure 6. Influence Diagnostics main effects. **Figure 7.** Plot of Residuals vs Fitted Values main effects.

To identify the influential observations, we applied influence diagnostics.The outlier and leverage diagnostics for Total Cholesterol indicates there are many observations with large leverage and some outlier observations. Then we used DFFITS to quantify influence. The influence diagnostics for the Total Cholesterol graph indicates only several observations have DFFITS values exceeding threshold and most of the observations stay inside the threshold.

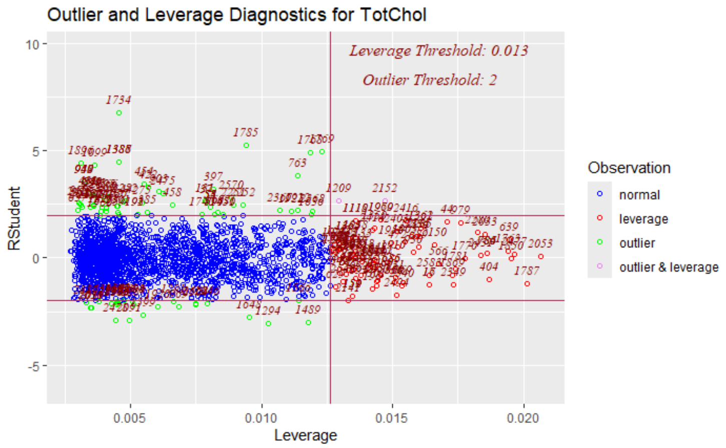
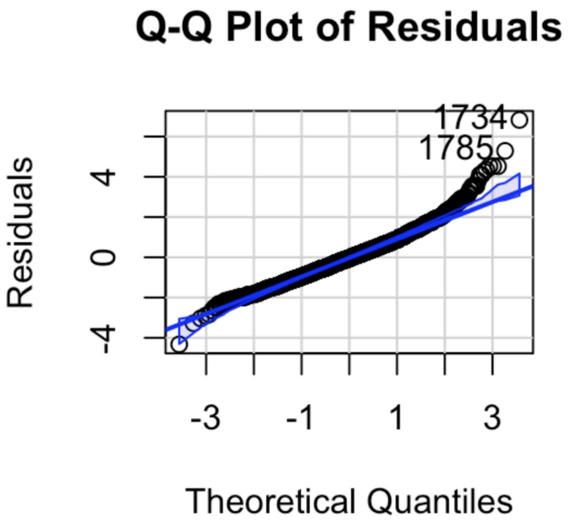


Figure 8. Outlier and Leverage Diagnostics

The backwards model selection identified a final model that effectively predicts Total Cholesterol while excluding variables that did not contribute significantly to the model. We removed BMI_log from the original model since it provided the lowest AIC reduction (78.512) and the removed PhysActiveDays with AIC=77.541. At this point, no further variables provide a significant improvement in the AIC score compared to the remaining predictors. The significant variables retained are: Gender, Race1, Depressed, BPSysAve, Smoke100, Age_Category and AlchoholYear_log.



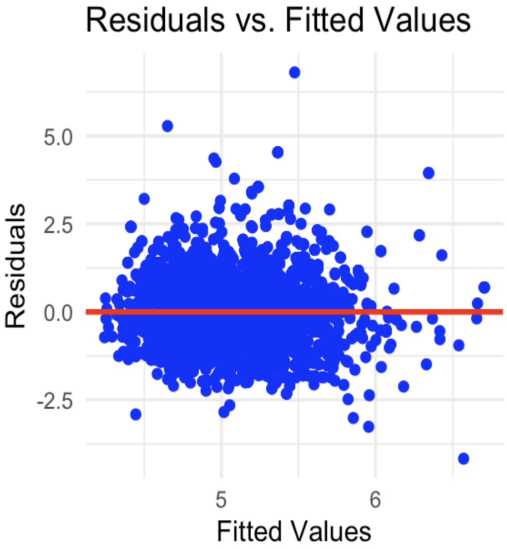


Figure 9. Q-Q plot for full model.

Figure 10. Residuals vs. fitted for full model.

By adding interaction terms, we got our final model. The multiple R-squared is 0.1215 which suggests that our final model explains 12.15% of the variation of the total cholesterol. The adjusted R-squared drops to 0.1129 suggests that some predictors may not add substantial explanatory power. Both R-squared values are relatively low, indicating that most of the variation in total cholesterol remains unexplained. The F statistic is 14.17 on 26 and 2663 degrees of freedom with an extremely small p-value indicating that our final model, as a whole, is statistically significant. It suggests that at least one predictor in the final model significantly contributes to explain total cholesterol.

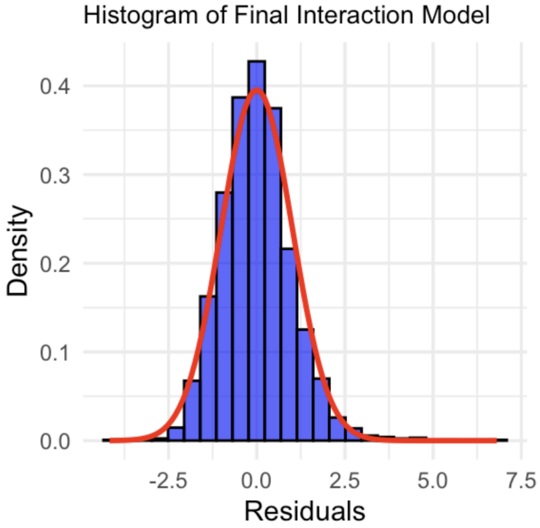


Figure 11. Histogram of final interaction model.

Bibliography

1. Anagnostis, Panagiotis, et al. "Effects of Menopause, Gender and Age on Lipids and High-Density Lipoprotein Cholesterol Subfractions." *Maturitas*, vol. 81, no. 1, May 2015, pp. 62–68. *PubMed*, https://doi.org/10.1016/j.maturitas.2015.02.262.
2. Hageman, Susan M., and Saurabh Sharma. "Low HDL Cholesterol." *StatPearls*, StatPearls Publishing, 2024. *PubMed*, http://www.ncbi.nlm.nih.gov/books/NBK560749/.
3. Jeong, Wonseok. "Association between Dual Smoking and Dyslipidemia in South Korean Adults." *PloS One*, vol. 17, no. 7, 2022, p. e0270577. *PubMed*, https://doi.org/10.1371/journal.pone.0270577.
4. Kostapanos, Michael S., and Moses S. Elisaf. "High Density Lipoproteins and Type 2 Diabetes: Emerging Concepts in Their Relationship." *World Journal of Experimental Medicine*, vol. 4, no. 1, Feb. 2014, pp. 1–6. *PubMed Central*, https://doi.org/10.5493/wjem.v4.i1.1.
5. Rajagopal, G., et al. "High-Density Lipoprotein Cholesterol: How High." *Indian Journal of Endocrinology and Metabolism*, vol. 16, no. Suppl 2, Dec. 2012, pp. S236–38. *PubMed Central*, https://doi.org/10.4103/2230-8210.104048.
6. Terry, Ana, et al. Plan and Operations of the National Health and Nutrition Examination Survey, August 2021–August 2023. stacks.cdc.gov, https://stacks.cdc.gov/view/cdc/151927. Accessed 8 Dec. 2024.
7. Yu, Lianlong, et al. "The Effect of BMI on Blood Lipids and Dyslipidemia in Lactating Women." *Nutrients*, vol. 14, no. 23, Dec. 2022, p. 5174. *PubMed*, https://doi.org/10.3390/nu14235174.

Contribution

Yingzhen Wang: R coding, including data preprocessing, model selection, data visualization, and diagnoses. Participate in making slides. Presentation.

Tong Liu: Literature collection, as well as the Methods section of the report.

Jackie Lindstrom: Abstract and background, participation in making slides, editing of results and discussion sections.

Gabriel Alwan: Participation in making slides, editing/reviewing of intro sections, results section, part of conclusion and discussion, materials section.

Qinglan Miranda Ouyang: Literature collection, reviewing of the abstract section, participation of conclusion and discussion section.