Examining Predictors of HDL Cholestrol using NHANES Data

Edward J. Lee, Yusuf Emre, Vincent Mazz

2025-04-05

Introduction

Cardiovascular disease remains a leading cause of death worldwide, with elevated cholesterol levels serving as an important and preventable risk factor. As prevention becomes a cornerstone of public health policy, understanding what drives changes in cholesterol is of the most importance.

This study investigates which factors significantly influence total cholesterol levels in the adult population, using data from the National Health and Nutrition Examination Survey (NHANES). Specifically, it examines the role of age, weight, height, blood pressure, smoking habits, and physical activity as potential predictors.

Previous research provides a useful foundation. For example, Ferrara et al. (1997) found that cholesterol levels tend to decline in older adults. However, this study observed a weak but significant positive association between age and cholesterol, suggesting that additional lifestyle or metabolic factors may be at play. Next, Henriksson et al. (2001) reported a negative correlation between BMI and HDL cholesterol. Although BMI was not included directly in this model, height and weight were analyzed independently. The findings showed that weight alone lacked a strong relationship with cholesterol, partially contradicting earlier work. Finally, Kim et al. (2011) linked high blood pressure with poorer lipid profiles—a pattern repeated here, as both systolic and diastolic blood pressure were positively associated with cholesterol levels.

However, there remains inconsistency in how these predictors interact across populations and within multifactorial health profiles. This study addresses this gap by assessing the influence of each factor using multivariable regression.

Linear regression was chosen for its ability to estimate the relationship between a continuous outcome—total cholesterol—and multiple predictors. Diagnostic checks were used to evaluate key assumptions, including linearity, homoscedasticity, and normality of residuals. While some violations were observed (e.g., non-normal residuals and heteroscedasticity), potential remedies such as Box-Cox transformations were explored. Despite these limitations, linear regression remains a strong baseline method for identifying statistically significant predictors of cholesterol.

By applying regression analysis to nationally representative NHANES data, this study provides data-driven insights into the factors most strongly associated with cholesterol levels—insights that can help shape future public health strategies.

Data Description

NHANES is a built-in data package from R containing survey data collected by the US National Center for Health Statistics (NCHS) from 2009-2011. The target population of the National Health and Nutrition Examination Surveys (NHANES) is the “the non-institutionalized civilian resident population of the United States”. The sample population was obtained via a multistep sampling method, and intentionally oversamples certain minority groups to accurately weight their proportions. These efforts are made with the goal to assess and monitor the health and nutritional conditions of the US population via health checkups and interviews (Centers for Disease Control and Prevention, 2015).

Total high-density lipoprotein (HDL) cholesterol in mmol/L is a variable surveyed by NHANES, captured by the variable *TotChol* in the data package. The sample mean is 5.026 mmol/L and the median is 4.970

mmol/L based on 1289 observations. The nature of this variable being a continuous measurement which follows an approximately normal distribution allows it to be a suitable response variable for a linear regression model. Furthermore, the current literature has evidence that the predictor variables used in our model have an influence upon HDL levels.

Table 1: Summary Statistics of ***TotChol***

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.53	4.24	4.97	5.025725	5.64	10.29

The following is a brief summary of these predictor variables and what current research knows about its relationship to the HDL cholesterol levels in humans.

Age: Age of participant at time of screening in years. The approximately uniform distribution of this variable is indicative of good sampling design. Researchers find that HDL cholesterol levels tend to diminish with age, but there are many other factors that can contribute to this (Ferrera et al., 1997; Milman et al., 2014).

Height and ***Weight***: The standing height of the participant in centimetres, and the weight of the participant in kilograms respectively. The measurements of height follow an approximately standard normal distribution, once again indicative of good sampling techniques.

Weight follows moderate right skew with its measures of central tendency gathering around 80kg (M = 81.16kg, Mdn = 78.8). Seeing as how close the mean and median are to each other and how consistent these statistics are to population parameters captured by the NCHS (Fryar et al., 2021), it is probable that the skew is caused by outliers.

Researchers have shown that Body Mass Index (BMI) scores are negatively associated with HCL cholesterol levels (Lamon-Fava et al., 1996). Note that BMI is a function of height and weight, which both have their respective associations with cholesterol levels (Henriksson et al, 2001).

BPSysAve and ***BPDiaAve***: The systolic and diastolic blood pressure readings of participants in millimeters of mercury. Both follow approximately normal distributions, where BPSysAve (M = 121.67, Mdn = 119.0) has a right skew, and BPDiaAve (M = 70.38, Mdn = 71.0) has a slight left skew. Seeing as how the means and medians of these variables are very close to each other, these skews could be attributed to outliers or natural variation in the population. The current research shows evidence of healthier blood pressure readings in those with higher HDL cholesterol (Al-Jarallah, 2022; Kim et al., 2011, Lye et al., 2009).

SmokeNow: 41.48% of participants (799 participants) have indicated that they currently smoke tobacco, with the remaining 57.52% of participants claiming they do not (590 participants). Several studies have shown evidence of negative associations with smoking and HDL cholesterol levels (Garrison et al., 1978; Nakamura et al., 2021).

DaysPhysActive: The self-reported number of days per week a participant engages in moderate to high intensity physical activity. Most participants claim they generally engage in physical activity 3 times a week. The current research generally tends to agree that consistent physical exercise promotes higher levels of HDL cholesterol (Kodama et al., 1960).

Table 2: Summary Statistics of Continuous Predictors

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Age	21.0	34.0	47.0	47.68425	60.0	80.0
Height	141.3	163.7	170.8	170.57502	177.3	199.6
Weight	41.1	67.6	78.8	81.15741	93.0	172.5
BPSysAve	81.0	110.0	119.0	121.66796	130.0	202.0
BPDiaAve	31.0	64.0	71.0	70.38169	78.0	116.0

Preliminary Model Diagnostics

Model Selection

Preliminary model diagnostics indicated the model would benefit from modifications to improve model fit based on the indications of violated linear regression assumptions. With the primary objective of a predictive model in mind, certain changes were implemented into the model.

A distinct convex curvilinear relationship is evident in Figure Scatterplot Matrix and Figure Residuals vs Age, indicative of a severe violation in linearity. The additional polynomial term *Age2*, the square of the *Age* variable vector, was included to capture this non-linear relationship between the *Age* predictor and dependent variable *TotChol*.

Following this change, the Box-Cox transformation was applied to the dependent variable *TotChol*. This transformation aims to address violations in normality and homoscedasticity as indicated by the right-tailed skew seen in Figure QQ Plot, and fanning patterns of residuals shown by Figure Residuals vs Fitted. Maximum likelihood was used to derive a lambda value ($\lambda = 0.1414$) for the transformation by using functions from the R package *MASS* (Venables & Ripley, 2002) and default built-in algebraic operators as shown in Formula 2. This transformation was not applied to the predictor variables to preserve interpretability.

$$Y^{(\lambda)} = \frac{Y^\lambda - 1}{\lambda}, \quad \lambda = 0.1414 \tag{2}$$

Remarkable improvements in model assumptions were noticed in the diagnostic plots of the transformed model, such as residual plots showing approximately null relationships with residuals more evenly and widely dispersed across the fitted values. Figure QQ Plot Transformed also now shows the effectiveness of the Box-Cox transformation with its resulting nearly perfect normal distribution in the residuals.

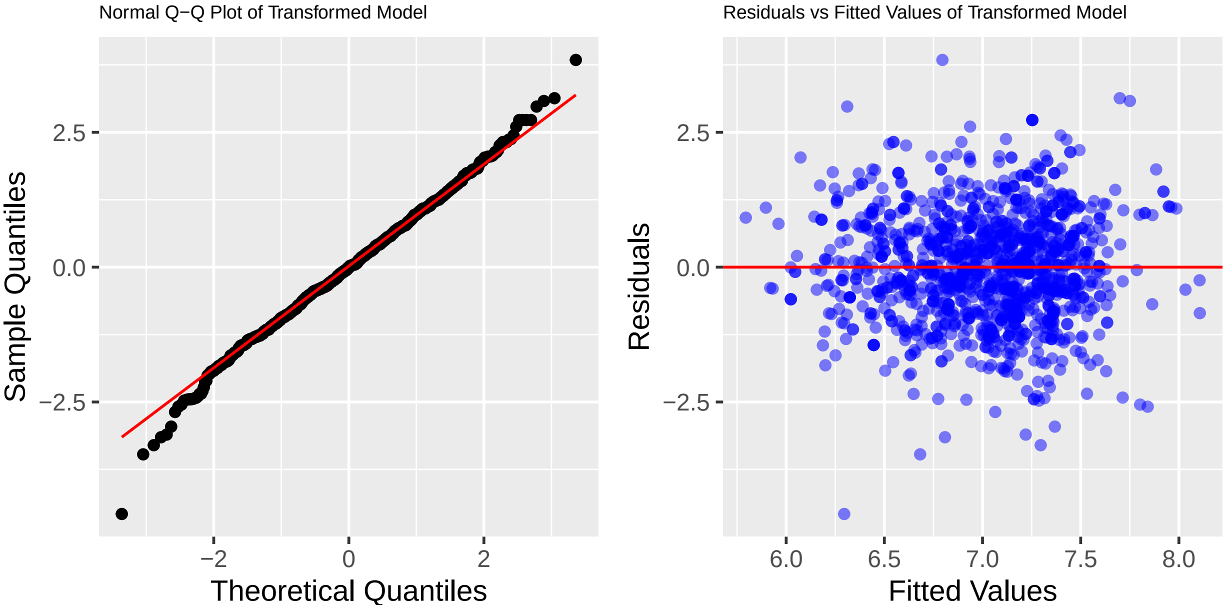


Figure 1: QQ-Plot and Residuals vs Fitted Plot of Transformed Model.

Metrics of model fit in the transformed model ($R^2 = 0.1264$, $adjR^2 = 0.121,$) also showed a large increase when compared to the preliminary model ($R^2 = 0.08085$, $adjR^2 = 0.07583$). Despite the polynomial untransformed model yielding higher fit, severe violations in normality motivated the use of the Box-Cox cocurrently. These metrics compared to other iterations of model transformations can be found in Table of Models.

Table 3: Comparison of Linear Regression Models

Model	R2	Adj_R2	F_value
Preliminary	0.081	0.076	16.10
Polynomial	0.127	0.121	23.20
Box-Cox	0.079	0.074	15.71
Poly and Box-Cox	0.126	0.121	23.15

Upon fitting the transformed model, the dataset was screened for problematic observations. Initial data cleaning ensured the dataset excludes null entries and obvious misinputs, thus the criteria for removing problematic observations was only a matter of measures of influence.

Outliers were identified by checking standardized residuals, and influential observations were identified based on their measurements of leverage, Cook's Distance, Difference in Fits (DFFITS) and Differences in Beta Coefficients (DFBETAS). If an observation had any of these measures surpass their respective thresholds and were concurrently highlighted by the *influenceIndexPlot()* function from the R package *car* (Fox & Weisberg, 2019) they were flagged as problematic observations.

Based on this criteria, five potentially problematic observations were identified, of which only observation 823 and 728 were removed from the data set. A summary of these observations and their measures of leverage are present in Table 2.

Table 4: Measures of Influence and Model Fit After Removing Selected Observations

	St. Residual	Cook's Distance	Leverage	DFFITS	Adj.R2 with Obs Removed
10	1.818	0.01501	0.03927	0.3679	0.122
968	-4.662	0.01571	0.00646	-0.3791	0.120
724	-1.306	0.00588	0.03006	-0.2300	0.122
823	3.913	0.01094	0.00639	0.3156	0.125
728	-3.222	0.01468	0.01257	-0.3649	0.123

For each potentially problematic observation, the transformed model was fitted using the same dataset but with the exclusion of the observation under inspection. Models were then compared to determine which observations to remove for highest model fit.

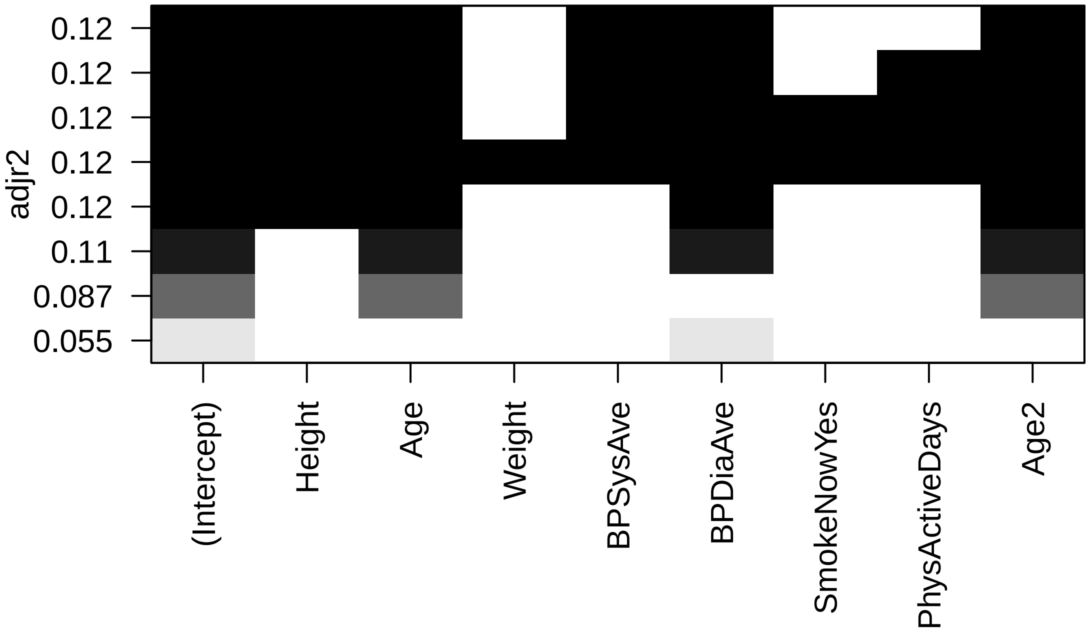
By this process, the exclusion of both observations 824 and 728 was found to induce the highest model fit ($R^2 = 0.129$, $adjR^2 = 0.1236$).The exclusion of any of the remaining problematic observations would decrease model fit (see Table 2), and thus with the motivation of a predictive model with high fit, the observations were retained in the dataset.

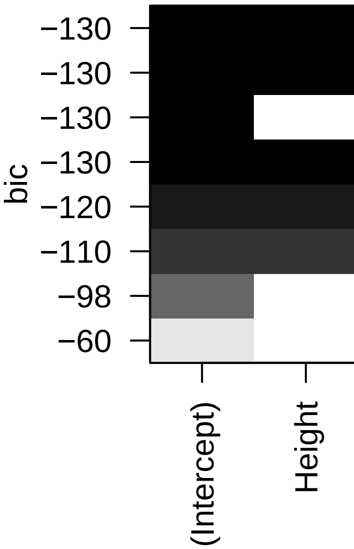
Several methods of variable selection were employed, such as full and partial F-tests, t-tests for individual predictors, and stepwise regression for AIC and BIC. All of these methods unanimously arrived at the same conclusion of finding the predictors *Age, Age2, Height, BPSysAve* and *BPDiaAve* to be statistically significant at $\alpha = .05$.

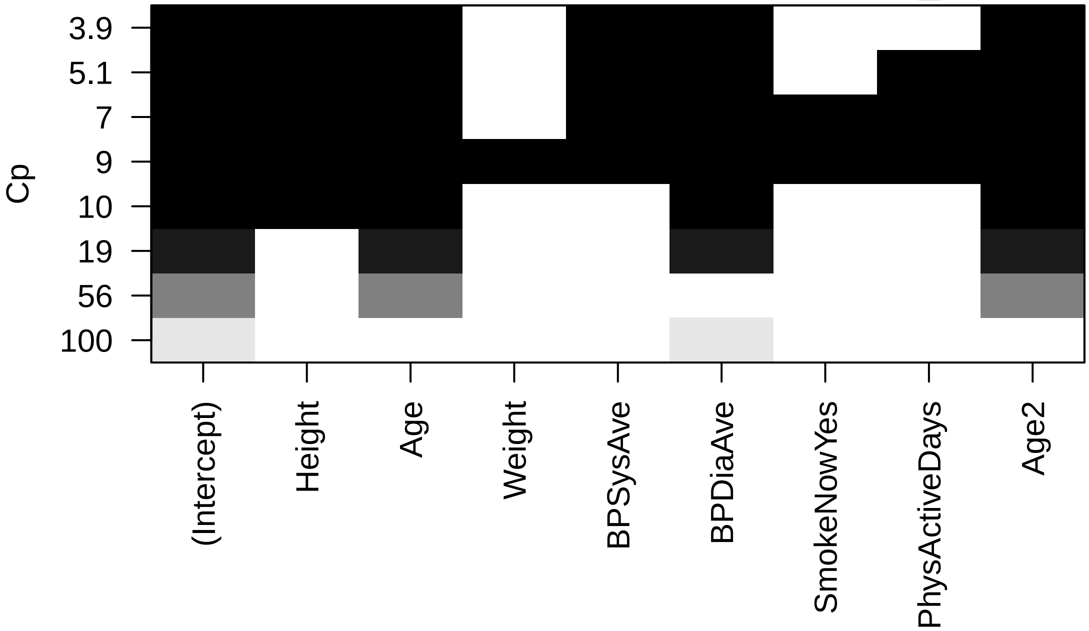
Table 5: ANOVA Table of Transformed Model

	Sum Sq	Df	F value	Pr(>F)	Significant
(Intercept)	60.5215002	1	63.5727641	0.0000000	Yes
Age	74.6467831	1	78.4101901	0.0000000	Yes
Age2	65.3772613	1	68.6733342	0.0000000	Yes
Weight	0.0004749	1	0.0004989	0.9821836	No
Height	8.3331807	1	8.7533080	0.0031473	Yes
BPSysAve	7.8837350	1	8.2812030	0.0040724	Yes
BPDiaAve	19.4749518	1	20.4568049	0.0000067	Yes
SmokeNow	0.0860365	1	0.0903741	0.7637509	No
PhysActiveDays	0.8114999	1	0.8524126	0.3560450	No
Residuals	1216.6605992	1278	NA	NA	NA

```
## Subset selection object
## Call: regsubsets.formula(pb.TotChol ~ ., data = clean.frame, nvmax = 8,
##     nbest = 1, really.big = TRUE, method = "exhaustive")
## 8 Variables  (and intercept)
##                Forced in Forced out
## Height             FALSE      FALSE
## Age                FALSE      FALSE
## Weight             FALSE      FALSE
## BPSysAve           FALSE      FALSE
## BPDiaAve           FALSE      FALSE
## SmokeNowYes        FALSE      FALSE
## PhysActiveDays     FALSE      FALSE
## Age2               FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          Height Age Weight BPSysAve BPDiaAve SmokeNowYes PhysActiveDays Age2
## 1  ( 1 ) " "    " " " "    " "      "*"      " "         " "            " "
## 2  ( 1 ) " "    "*" " "    " "      " "      " "         " "            "*"
## 3  ( 1 ) " "    "*" " "    " "      "*"      " "         " "            "*"
## 4  ( 1 ) "*"    "*" " "    " "      "*"      " "         " "            "*"
## 5  ( 1 ) "*"    "*" " "    "*"      "*"      " "         " "            "*"
## 6  ( 1 ) "*"    "*" " "    "*"      "*"      " "         "*"            "*"
## 7  ( 1 ) "*"    "*" " "    "*"      "*"      "*"         "*"            "*"
## 8  ( 1 ) "*"    "*" "*"    "*"      "*"      "*"         "*"            "*"
```







```
## Start:  AIC=-54.33
## pb.TotChol ~ Age + Age2 + Weight + Height + BPSysAve + BPDiaAve +
##     SmokeNow + PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - Weight          1     0.000 1216.7 -56.334
## - SmokeNow        1     0.086 1216.8 -56.244
## - PhysActiveDays  1     0.811 1217.5 -55.476
## <none>                        1216.7 -54.335
## - BPSysAve        1     7.884 1224.5 -48.022
## - Height          1     8.333 1225.0 -47.550
## - BPDiaAve        1    19.475 1236.1 -35.897
## - Age2            1    65.377 1282.0  11.028
## - Age             1    74.647 1291.3  20.300
```

```
##
## Step:  AIC=-56.33
## pb.TotChol ~ Age + Age2 + Height + BPSysAve + BPDiaAve + SmokeNow +
##     PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - SmokeNow        1     0.088 1216.8 -58.241
## - PhysActiveDays  1     0.811 1217.5 -57.476
## <none>                        1216.7 -56.334
## + Weight          1     0.000 1216.7 -54.335
## - BPSysAve        1     7.936 1224.6 -49.967
## - Height          1    10.536 1227.2 -47.237
## - BPDiaAve        1    19.546 1236.2 -37.823
## - Age2            1    65.904 1282.6   9.557
## - Age             1    75.216 1291.9  18.868
##
## Step:  AIC=-58.24
## pb.TotChol ~ Age + Age2 + Height + BPSysAve + BPDiaAve + PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - PhysActiveDays  1     0.811 1217.6 -59.384
## <none>                        1216.8 -58.241
## + SmokeNow        1     0.088 1216.7 -56.334
## + Weight          1     0.003 1216.8 -56.244
## - BPSysAve        1     8.071 1224.8 -51.731
## - Height          1    10.615 1227.4 -49.062
## - BPDiaAve        1    19.459 1236.2 -39.821
## - Age2            1    66.037 1282.8   7.779
## - Age             1    75.131 1291.9  16.872
##
## Step:  AIC=-59.38
## pb.TotChol ~ Age + Age2 + Height + BPSysAve + BPDiaAve
##
##                  Df Sum of Sq    RSS     AIC
## <none>                        1217.6 -59.384
## + PhysActiveDays  1     0.811 1216.8 -58.241
## + SmokeNow        1     0.088 1217.5 -57.476
## + Weight          1     0.000 1217.6 -57.384
## - BPSysAve        1     7.982 1225.5 -52.974
## - Height          1    10.444 1228.0 -50.391
## - BPDiaAve        1    19.562 1237.1 -40.870
## - Age2            1    65.411 1283.0   5.965
## - Age             1    74.398 1292.0  14.949
```

Results from a stepwise regression using AIC values further confirmed the need to remove redundant predictors *Weight*, *SmokeNow* and *PhysActiveDays*. This allowed for a final iteration of the fitted model with the optimal values of $AIC = -59.38$ and $BIC = -133.85$.

Final Model Inference and Results

The final regression model provides meaningful insight into the research question of identifying key predictors of total cholesterol levels.As the response variable is Box-Cox transformed, all interpretations are made on the transformed scale rather than in original mmol/L units.

$$\hat{Y}^{(\lambda)} = \hat{\beta}_{Age}x_{Age} + \hat{\beta}_{Age2}x^2_{Age2} + \hat{\beta}_{Height}x_{Height} + \hat{\beta}_{BPSysAve}x_{BPSysAve} + \hat{\beta}_{BPDiaAve}x_{BPDiaAve} + \hat{\epsilon}$$

Intercept (4.8098341) This is the expected value of the transformed total cholesterol when all predictors are 0. Since values like Age or Height can't realistically be zero, the intercept itself isn't meaningful on its own. Age (0.0974977) When all other predictors are held constant, a 1-year increase in age is associated with an increase of 0.0975 units in the transformed cholesterol levels. Age^2 (-0.0009263) This negative coefficient reflects the curvilinear relationship of age and cholesterol levels. As age increases further, the rate of increase in cholesterol decreases and after a certain age, cholesterol begins to decline. This quadratic effect answers the research question by confirming that age influences cholesterol negatively and positively at different times in life. Together, Age and Age^2 imply that middle-aged adults are likely to have the highest cholesterol while younger and older individuals have lower cholesterol. Height (-0.0098211) Each additional centimeter in height is associated with a decrease of about 0.0098 units in transformed cholesterol. This implies that taller individuals tend to have lower total cholesterol. Average Systolic Blood Pressure (BPSysAve, 0.0056469) A 1 mmHg increase in systolic blood pressure is associated with an increase of about 0.0056 units in transformed cholesterol. This suggests that higher systolic pressure is linked with higher cholesterol, supporting the idea that these cardiovascular risk factors are correlated. Average Diastolic Blood Pressure (BPDiaAve, 0.0127101) A 1 mmHg increase in diastolic blood pressure is associated with an increase of about 0.0127 units in transformed cholesterol. Notably, this effect is more than double the effect size of systolic pressure, suggesting that diastolic pressure may be a stronger predictor in this context.

10-Fold cross validation was applied to the final model to assess predictability performance and overall model validation. As anticipated, the final model yielded the most optimal values of root mean square error ($RMSE = 0.9751559$) and mean absolute error ($MAE = 0.7694394$) when compared to other models. These values in comparison to other models can be found in Table of Model Validation below.

```
## [1] 0.9542581
```

```
## [1] 0.9768613
```

```
## Linear Regression
##
## 1287 samples
##    5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1158, 1159, 1158, 1158, 1158, 1159, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.9751559  0.1373103  0.7694394
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
## Linear Regression
##
## 1287 samples
##    8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1159, 1160, 1160, 1157, 1159, 1158, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.9751863  0.1249023  0.7700059
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.
```

```
## Linear Regression
##
## 1287 samples
##    1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1159, 1159, 1159, 1159, 1158, 1158, ...
## Resampling results:
##
##   RMSE      Rsquared  MAE
##   1.040114  NaN       0.8219892
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
## Linear Regression
##
## 1289 samples
##    4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1161, 1159, 1162, 1160, 1160, 1159, ...
## Resampling results:
##
##   RMSE      Rsquared    MAE
##   1.040968  0.08495365  0.811146
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
## Linear Regression
##
## 1289 samples
##    7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1159, 1160, 1161, 1160, 1160, 1160, ...
## Resampling results:
##
##   RMSE      Rsquared    MAE
##   1.042695  0.07761385  0.8123572
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

These values (on the Box-Cox transformed scale) suggest that the final model's predictions for the value of TotChol deviates by roughly 0.98 mmol/L with the absolute error on average being approximately 0.77 mmol/L. The R-squared value suggests only 13.7% of the variance in the transformed TotChol variable is explained by the predictors in the model. These values, despite being modest at best, are a significant improvement when compared to the preliminary model ($RMSE = 1.042796$, $(R^2) = 0.0717131$, $MAE = 0.8143652$).

Discussion

In this study, we investigated the factors that influence the amount of total cholesterol levels in blood using NHANES (National Health and Nutrition Examination Survey) with a multiple linear regression model. Our final model included age, height, average diastolic and systolic blood pressure measurements, as well as a polynomial term for age as predictors. These predictors achieved the highest $R^2 = 0.1264$, our best attempt at modelling total cholesterol levels in blood.

Age was the most significant predictor among all. The positive coefficient for age indicates that on average, the total amount of cholesterol levels increase as one gets older. However, the negative coefficient for Age^2 indicates a non-linear relationship where total cholesterol levels increase with age up until a certain age at which they begin to decline afterwards. This behaviour agrees with findings by Ferrara et al. (1997), who also observed decreasing cholesterol levels in elderly population.

Height was found to be negatively associated with total cholesterol levels. While this may seem counterintuitive at first glance, findings by Henriksson et al. (2001) support this finding by indicating that shorter individuals may have unfavorable lipid profiles. This relationship may be mediated by other biological or socioeconomic factors that require further exploration.

Both diastolic and systolic blood pressures were positively associated with total cholesterol. Our observations are consistent with previous research by Kim et al. (2011) and Lye et al. (2009) who have shown that dyslipidemia and hypertension often occur together. Their interconnected nature may point to shared endocrinologic impacts such as systemic inflammation, and metabolic syndrome.

On the other hand, weight, physical activity, and smoking status were not retained in the final model. The exclusion of weight as a significant predictor is particularly noteworthy. Despite its known role in cardiovascular health, weight showed no significant relationship with cholesterol levels in our final model. This suggests that without accounting for its interaction with height, weight may not be an adequate measure for predicting cholesterol levels. This is consistent with the literature which often uses BMI rather than weight alone.

Similarly, the number of physically active days and smoking status did not show significant contributions either. This may be due to the bias in self-reported measures or the presence of mediating factors not captured in our model. For example, intensity and duration of exercise were not included, only their frequencies. Smoking status did not demonstrate its effects on cholesterol neither in our preliminary nor final model. This could stem from underreporting or variation in the duration and intensity of smoking habits among participants.

Key findings

We initially fitted a basic linear model which violated multiple linear regression assumptions. These violations included heteroscedasticity, non-normality of residuals, and under a broader scope, the presence of influential points. Through applying Box-Cox transformation, adding a quadratic term for age, and removing influential points, our final model got better at predicting total cholesterol levels. The residual plots also showed better homoscedasticity, and the Q-Q plot indicated improved normality despite some residual skewness.

Additionally, we used multiple validation techniques to evaluate the performance of our model which included model validation, 10-fold cross-validation as well as comparisons across reduced and full models. The root mean squared error for our final model was 0.977, and the cross-validation R^2 reached 13.73% with the final model. While these values imply a poor fit, tuning our preliminary model greatly improved its ability to explain the variance in total cholesterol levels.

Interestingly, the results of our Lasso regression landed on a similar subset of predictors (Age, Age^2, Height, BPSysAve, BPDiaAve) which further confirmed our choice of the final model.

Recommendations

Height as a predictive factor should be more closely studied. While it's an immutable trait, understanding why shorter individuals tend to have higher total cholesterol levels could inform early interventions or stratification in risk prediction models.

The lack of a significant effect for physical activity may be due to the lack of details . More detailed data on duration and intensity (e.g., minutes of activity, type of exercise) would likely capture its impact on cholesterol more accurately.

Weight alone is insufficient. Future studies should focus more on composite indices like BMI or body fat percentage rather than relying on weight as a standalone predictor.

Smoking status needs validation. Given the inconsistencies in smoking's effects on cholesterol in this and previous studies, future work should consider serum measurements to quantify the effects of smoking.

Limitations and Areas for Improvement

Low R^2 value: The final model explained 12.64% of the variance in cholesterol levels. This indicates that our model is missing predictors that could potentially do a better job at explaining cholesterol levels.

Measurement bias: NHANES data relies on self-reported records for key variables. Incorporating more objective measures like wearables for physical activity or anonymized hospital records for serum levels would likely provide more reliable data.

Lack of interaction terms: The final model did not explore interactions between predictors such as age and blood pressure, which may uncover more nuanced relationships. These could be explored in future analyses.

Focus on total cholesterol: Total cholesterol is useful, but it can hide important differences. Two people might have the same total HDL cholesterol, but very different breakdowns of LDL and triglycerides. Since LDL is more closely tied to heart disease risk, looking at it specifically could give better insights into how things like age, blood pressure, or height affect cholesterol profiles.

Suggestions for future studies:

Modeling non-linear and interaction effects: Using methods such as polynomial regression or tree-based models could capture non-linearities and interactions linear regression is unable to capture.

Expanding the predictor set: Including variables such as alcohol consumption and medication usage could significantly improve model fit.

Modelling for multiple response variables: Rather than modeling total HDL cholesterol alone, a multivariate approach that predicts LDL and triglycerides separately could be more thorough.

Stratified analysis: Examining how predictors influence cholesterol levels differently by gender, ethnicity, or socioeconomic status may improve accuracy.

Bibliography

Al-Jarallah, A., & Babiker, F. (2022). High Density Lipoprotein Reduces Blood Pressure and Protects Spontaneously Hypertensive Rats Against Myocardial Ischemia-Reperfusion Injury in an SR-BI Dependent Manner. Frontiers in Cardiovascular Medicine, 9, 825310–825310. https://doi.org/10.3389/fcvm.2022.825310

Centers for Disease Control and Prevention. (2015). National Health and Nutrition Examination Survey data [Data set]. National Center for Health Statistics. https://www.cdc.gov/nchs/nhanes/index.htm

Ferrara, A., Barrett-Connor, E., & Shan, J. (1997). Total, LDL, and HDL cholesterol decrease with age in older men and women. The Rancho Bernardo Study 1984-1994. Circulation, 96(1), 37–43. https://doi.org/10.1161/01.cir.96.1.37

Fox, J., & Weisberg, S. (2019). An R companion to applied regression (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fryar, C. D., Carroll, M. D., Gu, Q., Afful, J., & Ogden, C. L. (2021). Anthropometric reference data for children and adults: United States, 2015–2018 (Vital and Health Statistics, Series 3, No. 46). National Center for Health Statistics. https://www.cdc.gov/nchs/products/index.htm

Garrison, R. J., Kannel, W. B., Feinleib, M., Castelli, W. P., McNamara, P. M., & Padgett, S. J. (1978). Cigarette smoking and HDL cholesterol the Framingham offspring study. Atherosclerosis, 30(1), 17–25. https://doi.org/10.1016/0021-9150(78)90149-1

Henriksson, K. M., Lindblad, U., Ågren, B., Nilsson-Ehle, P., & Råstam, L. (2001). Associations between Body Height, Body Composition and Cholesterol Levels in Middle-Aged Men. The Coronary Risk Factor Study in Southern Sweden (CRISS). European Journal of Epidemiology, 17(6), 521–526. https://doi.org/10.1023/A:1014508422504

Kim, N. H., Cho, H. J., Kim, Y. J., Cho, M. J., Choi, H. Y., Eun, C. R., Kim, J.-H., Yang, S. J., Yoo, H. J., Kim, H. Y., Seo, J. A., Kim, S. G., Baik, S. H., Choi, D. S., & Choi, K. M. (2011). Combined Effect of High-Normal Blood Pressure and Low HDL Cholesterol on Mortality in an Elderly Korean Population: The South-West Seoul (SWS) Study. American Journal of Hypertension, 24(8), 918–923. https://doi.org/10.1038/ajh.2011.78

Kodama, S., Tanaka, S., Saito, K., Shu, M., Sone, Y., Onitake, F., Suzuki, E., Shimano, H., Yamamoto, S., Kondo, K., Ohashi, Y., Yamada, N., & Sone, H. (2007). Effect of Aerobic Exercise Training on Serum Levels of High-Density Lipoprotein Cholesterol: A Meta-analysis. Archives of Internal Medicine (1960),167(10), 999–1008.https://doi.org/10.1001/archinte.167.10.999

Lamon-Fava, S., Wilson, P. W. F., & Schaefer, E. J. (1996). Impact of Body Mass Index on Coronary Heart Disease Risk Factors in Men and Women: The Framingham Offspring Study. Arteriosclerosis, Thrombosis, and Vascular Biology, 16(12), 1509–1515. https://doi.org/10.1161/01.ATV.16.12.1509 Lye, H.-S., Kuan, C.-Y., Ewe, J.-A., Fung, W.-Y., & Liong, M.-T. (2009). The improvement of hypertension by probiotics: Effects on cholesterol, diabetes, renin, and phytoestrogens. International Journal of Molecular Sciences, 10(9), 3755–3775. https://doi.org/10.3390/ijms10093755

Milman, S., Atzmon, G., Crandall, J., & Barzilai, N. (2014). Phenotypes and genotypes of high density lipoprotein cholesterol in exceptional longevity. Current vascular pharmacology, 12(5), 690–697. https://doi.org/10.2174/1570161111666131219101551

Muscella, A., Stefàno, E., & Marsigliante, S. (2020). The effects of exercise training on lipid metabolism and coronary heart disease. American Journal of Physiology-Heart and Circulatory Physiology, 319(1). https://doi.org/10.1152/ajpheart.00708.2019

Nakamura, M., Yamamoto, Y., Imaoka, W., Kuroshima, T., Toragai, R., Ito, Y., Kanda, E., Schaefer,

E. J., & Ai, M. (2021). Relationships between Smoking Status, Cardiovascular Risk Factors, and Lipoproteins in a Large Japanese Population. Journal of Atherosclerosis and Thrombosis, 28(9), 942–953. https://doi.org/10.5551/jat.56838

Venables, W. N., & Ripley, B. D. (2002). Modern applied statistics with S (4th ed.). Springer. https://www.stats.ox.ac.uk/pub/MASS4/