Modeling HDL Cholesterol Levels: Insights from Age, BMI, Gender, Testosterone, and Smoking

Feifan Liu, Andrew Huang, Catherine Yan, Zhenyu Wang

Introduction

We aim to answer the following *research question*: Can we find a multiple regression model to predict HDL(high-density lipoprotein) cholesterol levels based on a combination of variables such as physical activity, gender, age, BMI, testosterone, and smoking in the American adult population over 20? Utilizing data from the National Library of Medicine's National Health and Nutrition Examination Survey (NHANES), we aim to quantify the predictive power of these variables using multiple regression analysis. In particular, our regression coefficients and our coefficient of determination can help us answer our question.

Prior research has laid the groundwork for this inquiry. Schwab et al. investigated the association of HDL with BMI, age, and sex; Kokkinos et al. explored the effect of exercise on HDL; HEISS et al. analyzed the impact of smoking; and Bagatell et al. examined HDL in relation to testosterone levels. While these studies have individually contributed to our understanding of HDL cholesterol, a comprehensive model integrating *all* these variables has been less explored.

Methods

We will initiate the process by utilizing the initial model from Project 1. Subsequently, we will eliminate predictors with p-values exceeding 0.05, employing the partial F test. Suppose the partial F test fails to provide evidence for rejecting the hypothesis, suggesting that the removed predictors lack significant linear relationships with the response. In that case, we will consider the reduced model as our first model. Otherwise, we will retain the full model as our first model.

Continuing, we will check the linear regression conditions. Initially, we will create scatter plots of the response versus fitted values, checking for any random diagonal scatter or easily identified non-linear trends. Subsequently, we will generate all predictor pairwise scatterplots, checking for curves or other non-linear patterns. We will consider our residual plot reliable if there are no such patterns.

Next, we will assess the four linear regression assumptions. If we discover any deviations, curving, or wiggling from the diagonal line in the QQ plot, there will be a violation of the normal error assumption. In such cases, considering a Box-Cox transformation for some numerical variables might be warranted. Then, we will construct residual versus fitted and residuals versus predicted residual plots. If systematic patterns other than linear emerge, a Box-Cox transformation for some numerical variables may be considered to fix the violation on linearity. Subsequently, we will check for systematic patterns such as clustering or sequential trends, potentially indicating a violation of the uncorrelated errors assumption. If present, alternative models may be contemplated. Lastly, we will evaluate constant error by examining for fanning

patterns with increasing or decreasing speed on the residual plots. If detected, a variance-stabilizing transformation may be applied to our response model.

For the second model, we will only transform our response variable. Then, for the third model, we will transform all the variables to the proper stage. Subsequently, we will repeat the steps from the initial stage for both models, which involve removing variables with p-values exceeding 0.05 and utilizing the partial F test to determine our second/third model.

Next, we will examine the dataset for potential issues, including leverage observations, outlying points, and influential observations. For leverage observations, which are those with a distance from the center of the X-space, we will use each observation's hat value as its measure. To identify outlying points—those situated far from the trend—we will employ standardized residuals as measures. For influential observations, we will use measures such as Cook's distance to evaluate if an observation influences the estimation of all fitted values. Additionally, we will use the Difference in Fitted Values (DFFITS) to assess the influence of a single observation on its own fitted value and the Difference in Betas (DFBETAS) to quantify the influence on at least one estimated coefficient. Throughout this process, we will calculate cutoff points based on the size of the dataset and the number of predictors. Any measure exceeding the cutoff points will be deemed a problematic observation. If any observation appears to be problematic in every way, we will exclude that particular observation from the dataset.

Following this, we will assess multicollinearity using the Variance Inflation Factor (VIF), employing a cutoff value of 5. If any variable surpasses a VIF greater than 5, we may consider removing this variable from the model.

Following that, we will employ likelihood measures (Adjusted R-squared, Akaike's Information Criteria (AIC), Bayesian Information Criteria (BIC), and corrected AIC (AICc)) to guide our model selection. If a model exhibits the highest Adjusted R-squared alongside the smallest values for AIC, BIC, and AICc, we will designate it as our preferred model. If no model satisfies this condition, we might choose the model based on its exact values of likelihood measures. Subsequently, we will reassess the linear regression conditions and assumptions. If all criteria are met, we will consider it our best model.

Results

Our initial model included all the variables:Age, Gender, BMI, the interaction between PhysActiveDays and Gender, Smoke100, and the interaction between Testosterone and Gender. The large F-statistics (147.577) and small p-value (< 2.2e-16) of model 1 (Table 1.1) allowed us to reject the null hypothesis, indicating a significant linear relationship between the response (DirectChol) and at least one of the predictors. However, we found that two variables, Smoke100 and PhysActiveDays:Gender, weren't significantly linearly related to the response because their p-value's were larger than 0.05. Consequently, we came up with a reduced model 2 (Table. 1.2). To validate the removal of the predictors, we conducted a partial F test and got a p-value of 0.5016. This suggests that we failed to reject the null hypothesis and concluded that the predictors we removed were all not significantly linearly related to the response. Therefore, it is appropriate to retain our reduced model.

Table 1. Summary of Models 1–2. The response variable is DirectChol, the cholesterol concentration. We have six variables, including Age, Gender, BMI, interaction between physical active days and gender, Smoke100, and interaction between Testosterone and gender. The stars beside a number suggest that it has a p-value less than 0.01, which means the number is significant.

	Dependent variable:	
	DirectChol	
	(1)	(2)
Age	0.003***	0.003***
	(0.0004)	(0.0004)
Male	-0.295***	-0.297***
	(0.027)	(0.025)
BMI	-0.020***	-0.020***
	(0.001)	(0.001)
Smoke100	-0.012	
	(0.012)	
Female*PhysActiveDays	-0.003	
	(0.004)	
Male*PhysActiveDays	-0.003	
	(0.004)	
Female*Testosterone	0.002***	0.002***
	(0.0004)	(0.0004)
Male*Testosterone	0.0001***	0.0001***
	(0.00005)	(0.00005)
Constant	1.892***	1.884***
	(0.036)	(0.035)
Observations	3,306	3,306
R2	0.264	0.263
Adjusted R2	0.262	0.262
Residual Std. Error	0.339 (df = 3297)	0.339 (df = 3300)
F Statistic	147.577*** (df = 8; 3297)	235.698*** (df = 5; 3300)
Note:		*p<0.1; **p<0.05; ***p<0.01

Then, we examined the assumptions (Fig. 1) and conditions (Fig. A1) of model 1. There are a few problematic observations and clusterings in the fourth and fifth plots, suggesting potential violations of correlated errors. Furthermore, we would suspect a violation of normality, as shown by the deviations from normality in the QQ plot. Because the log-likelihood of its Box-Cox

transformation of response is close to 0 (Fig. A2). Therefore, we fitted a transformed model1 (Table 2.1), applying a natural log transformation to the response variable (DirectChol).

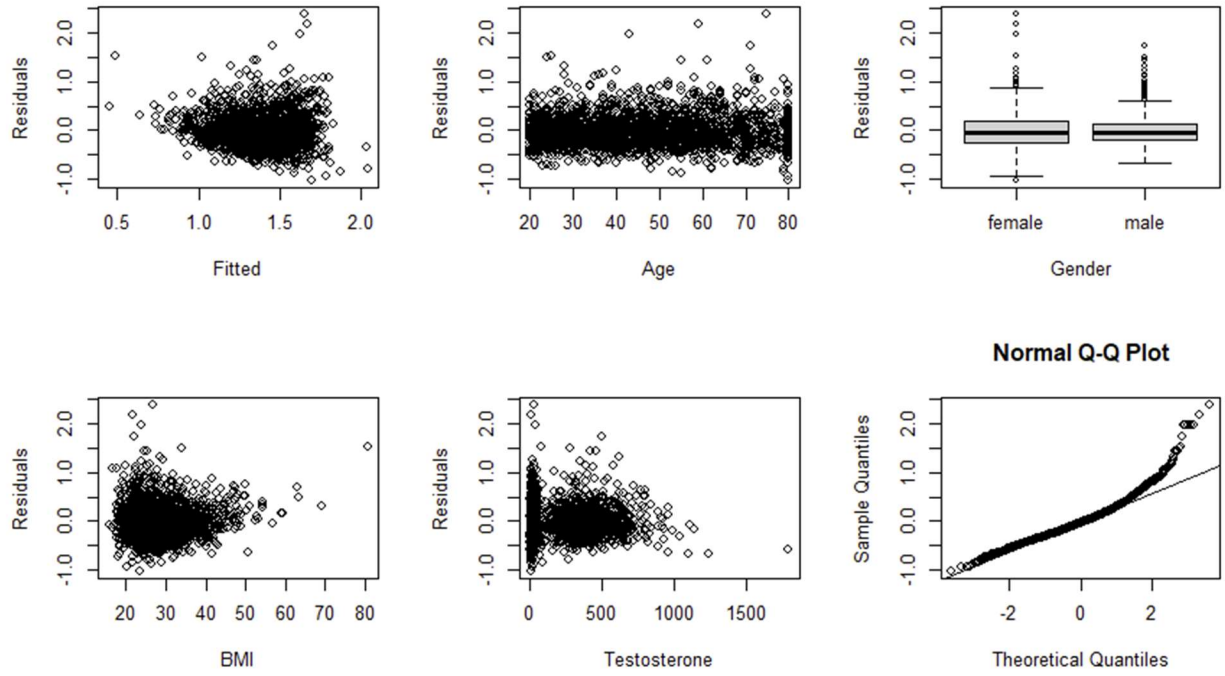

Figure 1. Assumption check for model 1. The first scatterplot is residual versus predictor, the second to fifth scatterplots are residual versus each predictor, and the last plot is a normal quantile-quantile (QQ) plot.

Similarly, when transforming response into ln(DirectChol), we saw the predictors Smoke100 and Gender: PhysActiveDays were not significantly linearly related to the response in transformed model 1 (Table 2.1), so we fitted a reduced transformed model 2 (Table 2.2). The partial F test for this model yielded a p-value of 0.4371, so the excluded predictors did not significantly linearly relate to the response. We then looked at potential transformations of the predictors. We found that numerical predictors such as age, BMI, and Testosterone could benefit from the power transformation. For computational simplicity, we rounded their transformation powers to ½, -(⅔), and natural log, respectively, and introduced these modified variables (named Age^(1/2), BMI^(-2/3), and ln(Testosterone)) into transformed model 3 (Table 2.3). Similarly, in transformed model 4 (Table 2.4), we removed non-significant variables, as indicated by a partial F test p-value of 0.3167.

Table 2. Summary of transformed models 1–4. The response variable lnChol was derived by applying a natural log transformation to DirectChol in models 3-6. There are six original variables, including Age, Gender, BMI, interaction between physical active days and gender, Smoke100, and interaction between Testosterone and gender. We transformed Age, BMI, and Testosterone, to Age^(1/2), BMI^(-2/3), and ln(Testosterone) in transformed models 3-4. Specifically, Age^(1/2) is the square root of Age, BMI^(-2/3) is BMI raised to the power of -⅔, and ln(Testosterone) is the natural log of Testosterone.

	Dependent variable:			
	lnChol			
	(1)	(2)	(3)	(4)
Age	0.002***	0.002***		
	(0.0002)	(0.0002)		
Age^(1/2)			0.037***	0.037***
			(0.003)	(0.003)
Male	-0.226***	-0.226***	-0.272***	-0.274***
	(0.019)	(0.017)	(0.074)	(0.073)
BMI	-0.014***	-0.014***		
	(0.001)	(0.001)		
BMI^(-2/3)			6.756***	6.723***
			(0.276)	(0.276)
Smoke100	-0.009		-0.011	
	(0.008)		(0.008)	
Female*PhysActiveDays	-0.002		-0.003	
	(0.003)		(0.002)	
Male*PhysActiveDays	-0.002		-0.002	
	(0.003)		(0.003)	
Female*Testosterone	0.001***	0.001***		
	(0.0003)	(0.0003)		
Male*Testosterone	0.0001***	0.0001***		
	(0.00003)	(0.00003)		
Female*ln(Testosterone)			0.059***	0.058***
			(0.009)	(0.009)
Male*ln(Testosterone)			0.043***	0.043***
			(0.011)	(0.011)
Constant	0.659***	0.652***	-0.786***	-0.786***
	(0.025)	(0.024)	(0.052)	(0.052)
Observations	3,306	3,306	3,306	3,306
R2	0.280	0.279	0.301	0.300
Adjusted R2	0.278	0.278	0.299	0.299
Residual Std. Error	0.234 (df = 3297)	0.234 (df = 3300)	0.230 (df = 3297)	0.230 (df = 3300)
F Statistic	160.046*** (df = 8; 3297)	255.551*** (df = 5; 3300)	177.468*** (df = 8; 3297)	283.196*** (df = 5; 3300)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3. Likelihood measures for three reduced models. Transformed model4 has the highest adjusted R squared, and lowest AIC, BIC and AICc.

Model	R Squared	Adjusted R Squared	AIC	BIC	AICc
Model2	0.263	0.262	-7142.23	-7105.61	-7142.19
Transformed Model2	0.279	0.278	-9602.02	-9565.4	-9601.99
Transformed Model4	0.300	0.299	-9700.37	-9663.75	-9700.34

After obtaining the models, we wanted to select the best model among the three reduced models (model2, transformed model 2, and transformed model 4). We first assessed for problematic points, including leverage observations, outliers, and influential observations. Notably, there are no points that are problematic in all checks. However, model 2 exhibited a higher number of outliers compared to the other two, with the transformed model 2 having the fewest.

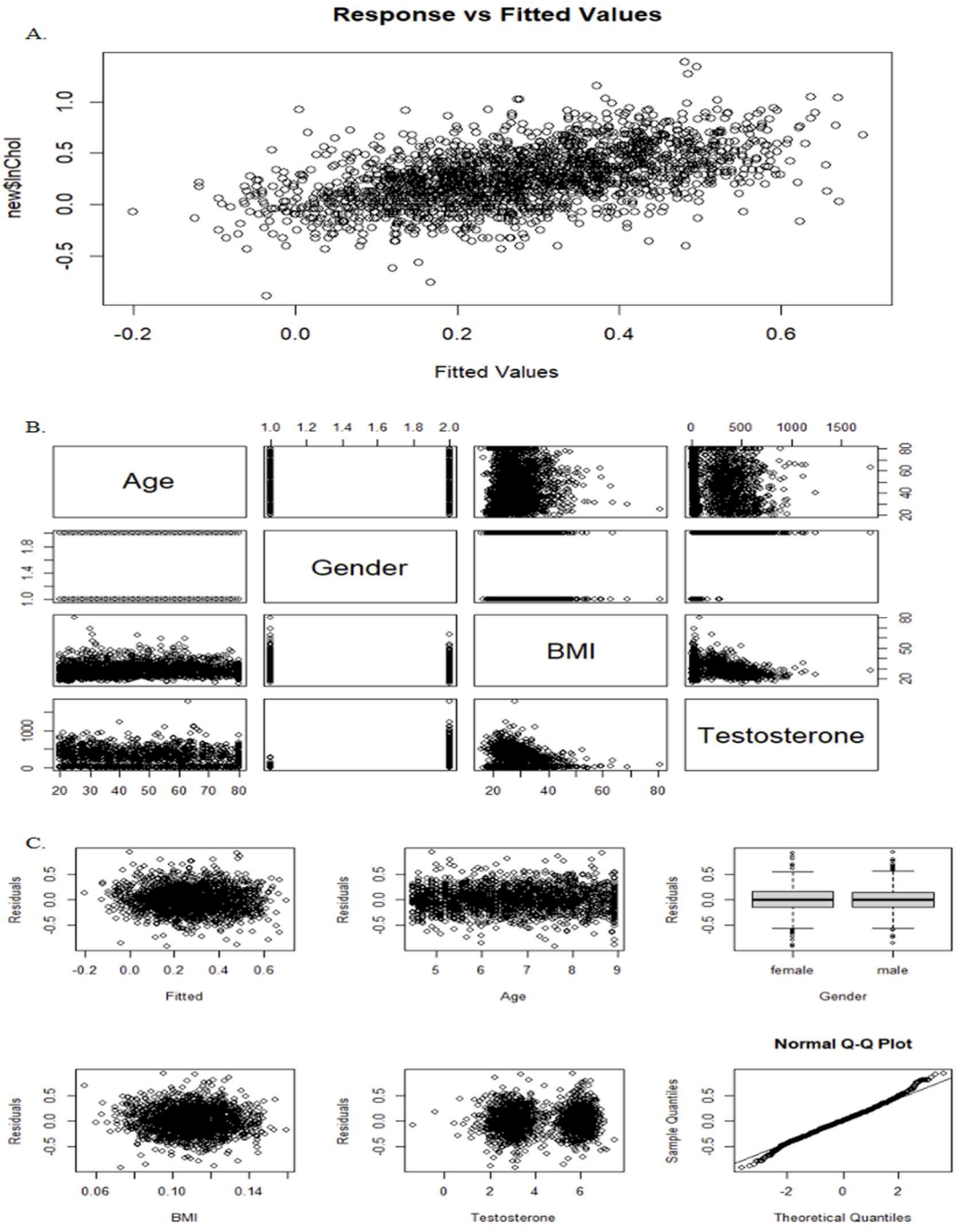

Figure 2. Conditions and Assumptions check of our selected model, transformed model4. (A) Conditional mean response, (B) Conditional mean predictors, (C) Assumptions (linearity, uncorrelated errors, constant variance, and normality) check.

Furthermore, we checked the multilinearity among each model (Table A1). The VIF between Gender and Gender:Testosterone2 is higher than 5 in transformation model 4, indicating a strong multicollinearity between them. However, based on our research assumptions, we acknowledged that testosterone levels vary by gender and aimed to explore this relationship. Therefore, we decided not to remove this variable due to its relevance to our study.

In addition, we look into their likelihood measures to further select the models. Transformed model 4 demonstrated superior performance with the highest adjusted R-squared and the lowest AIC, BIC, and AICc values (Table 3). Based on all these metrics above, we selected transformed model 4 as our preferred model.

Finally, we ensured that the selected transformed model 4 adhered to all linear regression conditions and assumptions (Fig. 2). This evaluation revealed no violations of conditions (Fig 2A, B), and notably, the normality of the model significantly improved following the transformation (Fig. 2C). Although we recognized a potential assumption violation with respect to testosterone, indicated by the presence of two distinct clusters (Fig. 2C), we used the interaction between gender and testosterone as our predictor variable to relieve the problem. As a result, we could confidently conclude that transformed model 4 is not only our best model but also meets all the necessary assumptions for linear regression.

Discussion

Conclusion of the research question

Our final model (transformed model 4) is a multiple linear regression model with (natural) log-transformed HDL cholesterol ($ln(cholesterol)$) where cholesterol is measured in mmol/L) as the dependent variable. The independent variables are:

- $\sqrt{Age}$
- $Gender$
- $(BMI)^{-2/3}$
- $Gender \times ln(Testosterone)$

Our final model was selected from a group of other models that integrated various predictors such as physical activity, BMI, sex, testosterone levels, and smoking status. It was chosen based on its adherence to linear regression assumptions and its superior model performance metrics, including the highest adjusted R-squared and the lowest AIC, BIC, and AICc values.

The intercept of our model (-0.786) has no meaningful interpretation since zero-valued predictor variables are never observed in the real world. Our coefficients represent the change in mean $ln(cholesterol)$ per unit increase in the corresponding independent variable, **holding all others constant**. A one-unit increase in $\sqrt{Age}$ corresponds to a 0.037 unit increase in mean $ln(cholesterol)$. For unit increase in $(BMI)^{-2/3}$, 6.723 increase in mean response. Being male is associated with a -0.274 change in mean $ln(cholesterol)$. Lastly, for each unit increase in $ln(Testosterone)$, the change in expected $ln(cholesterol)$ for males is 0.043, and for females is 0.058. We were unable to find a model where physical activity and smoking were significant variables.

Our model suggests a significant relationship between HDL cholesterol levels and BMI, gender, age, and testosterone levels among adults over 20 in the United States. Within our dataset, our model can explain 30% of the variance of $ln(cholesterol)$.

Our results are mostly in line with the results of Schwab et al., Bagatell et al., and Heiss et al., particularly with Gender, Age, and Testosterone. The coefficient for BMI seems to contradict the results of Heiss et al. because in our model, a decrease in BMI results in an increase in$(BMI)^{-2/3}$, which is associated with an increase in the mean response. We could not find a good comparison with Kokkinos et al. as our model was not able to incorporate physical activity levels.

Limitation of analysis

Firstly, the data is solely from the United States, potentially limiting the findings' applicability to other populations. Second, the relatively small adjusted R-square value indicates that while our model identifies general trends in HDL cholesterol levels, its predictive accuracy at an individual level is limited. Therefore, the results should be interpreted as indicative of broader patterns rather than precise predictors of HDL levels. These limitations underscore the need for further research with more varied datasets and more comprehensive models.

In the context of health studies, an R-squared value of 0.300 might be considered high. A high R-squared may be achieved by overfitting to the specific dataset, which can lead to poor generalization of new data. We were not able to test any model validation techniques on our model.

Ethics Discussion

We chose to use manual model selection instead of automatic model selection.

From an ethical perspective, we feel there is a duty of *beneficence* to find the best possible model. An accurate and capable model would be able to help people make educated decisions for their health. We believed manual model selection would be best here because we had a wider range of available techniques and could also use domain-relevant knowledge to help inform our decisions during model-making. Automatic model selection methods would be inferior in some aspects because they cannot include transformation, interaction variables, and assumption checks.

We are also compelled to perform an even greater duty of *non-maleficence*. Publishing results that are statistically invalid or misleading could lead to harmful decisions. Thus, when it comes to scientific studies made in health research, we believe that it is important to understand the motivation behind the results. To avoid this, it was important for us to do our due diligence. We felt that manual methods allowed for more control over this process because they are more *transparent* than automatic methods.

From a *professional responsibility* standpoint, it is important to be able to take ownership of the results of your work - and be able to back them up - instead of offloading such responsibilities to

an automated tool. Doing otherwise could pose legal and moral risks. Thus, we believe that manual methods are superior in this fashion.

Citation

Schwab, K. O., Doerfer, J., Naeke, A., Rohrer, T., Wiemann, D., Marg, W., Hofer, S. E., Holl, R. W., & Germán. (2009). Influence of food intake, age, gender, HbA1c, and BMI levels on plasma cholesterol in 29 979 children and adolescents with type 1 diabetes - reference data from the German diabetes documentation and quality management system (DPV). Pediatric Diabetes, 10(3), 184–192. https://onlinelibrary-wiley-com.myaccess.library.utoronto.ca/doi/full/10.1111/j.1399-5448.2008.00469.x

Kokkinos, P., & Fernhall, B. (1999). Physical activity and high density lipoprotein cholesterol levels. Sports Medicine, 28(5), 307–314. https://doi.org/10.2165/00007256-199928050-00002

HEISS, GERARDO, JOHNSON, NORMAN, REILAND, SUSAN, DAVIS, C. & TYROLER, HERMAN. (1980). The Epidemiology of Plasma High-density Lipoprotein Cholesterol Levels: The Lipid Research Clinics Program Prevalence Study Summary. Circulation, 62, IV-116-IV-136. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=ovfta&NEWS=N&AN=00003017-198011001-00016.

Bagatell, C. J., Knopp, R. H., Vale, W., Rivier, J., & Bremner, W. J. (1992). Physiologic testosterone levels in normal men suppress High-Density lipoprotein cholesterol levels. Annals of Internal Medicine, 116(12_Part_1), 967–973. https://doi.org/10.7326/0003-4819-116-12-967

Appendix

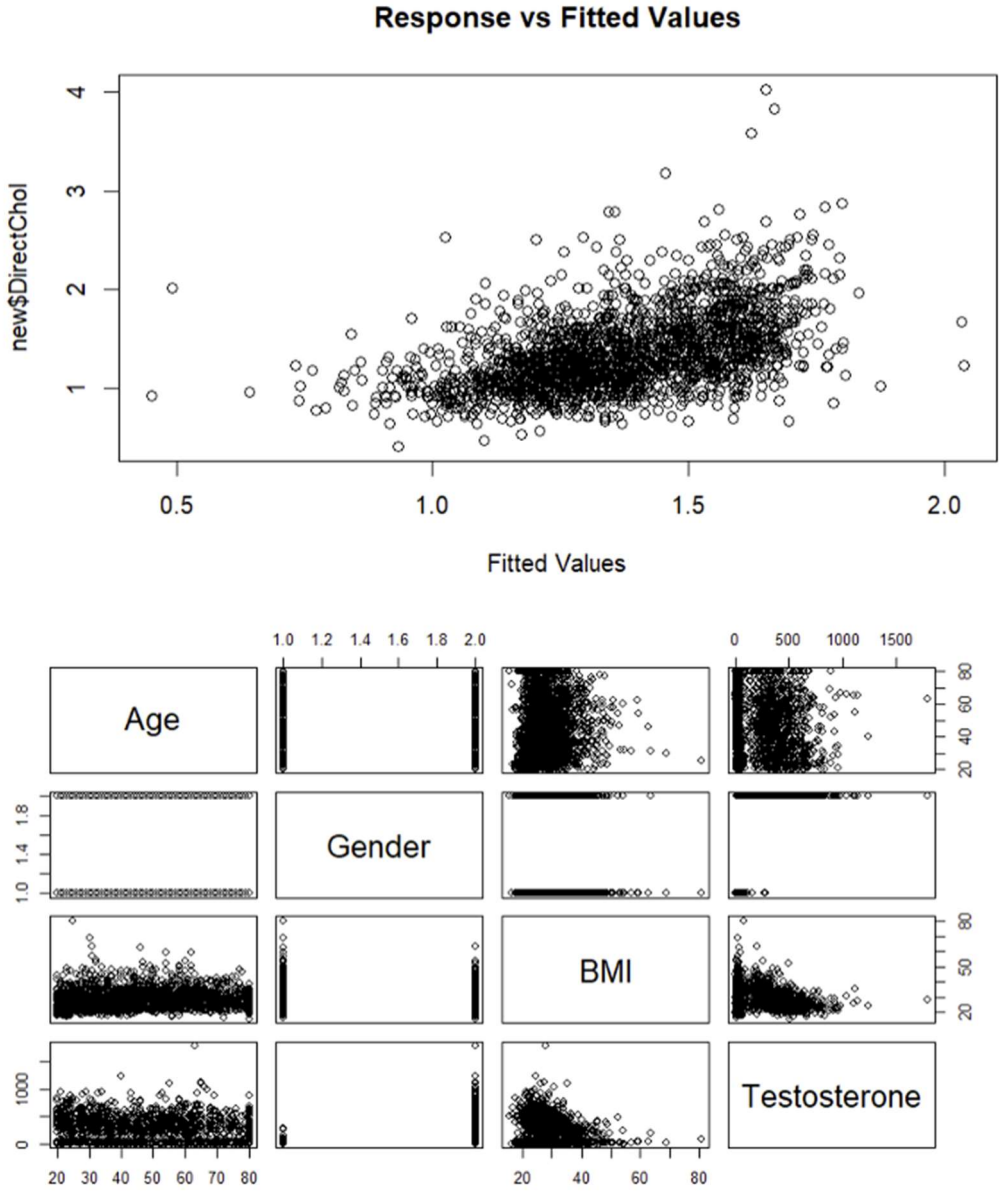

Figure A1. Condition checks of model 2. The upper figure is conditional mean response while the lower is conditional mean predictors. There are a few potential problematic observations in conditional mean response because they are away from the diagonal line.

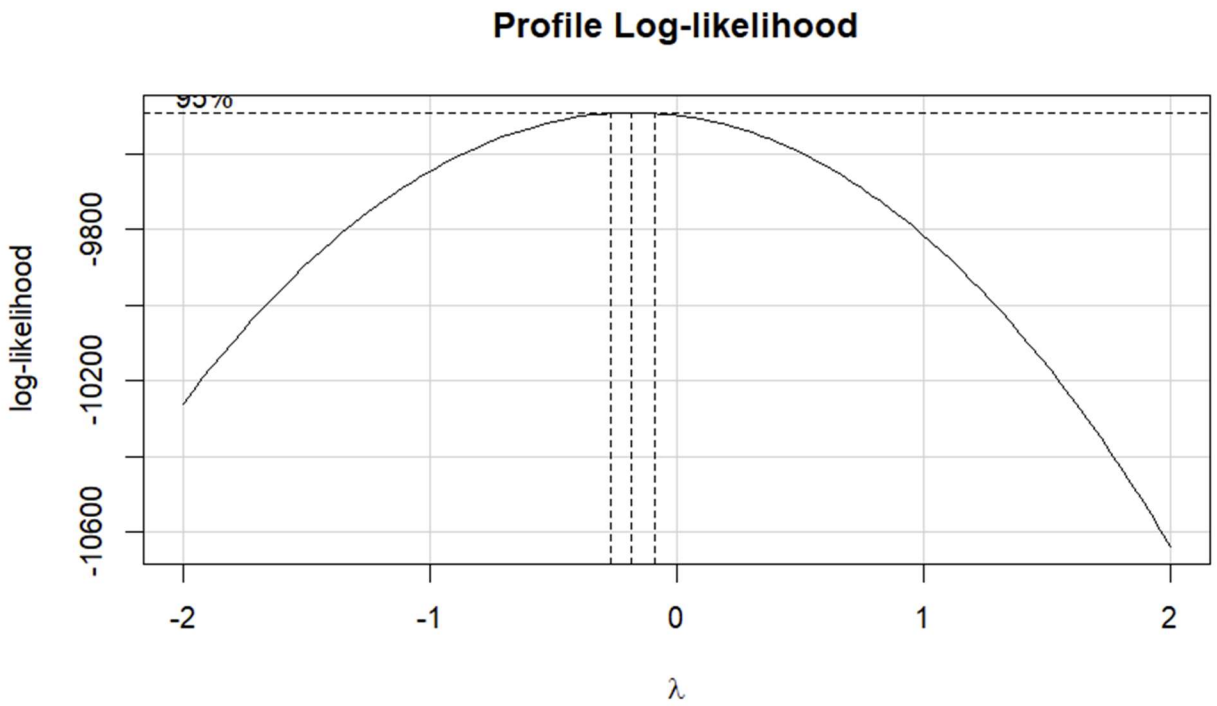

Figure A2. Box-Cox transformation on response (DirectChol). The rounded BoxCox power is -0.15, which is close to 0.

Table A1. Variance Inflation Factor (VIF) of predictor variables in model2, transformed model2, and transformed model4. The Age, BMI, and testosterone is transformed in Transformed model4. The cutoff value equals to 5.

VIF	Age	Gender	BMI	Gender:Testosterone
Model2	1.025	4.572	1.064	4.744
Transformed Model2	1.025	4.572	1.064	4.744
Transformed Model4	1.066	83.798	1.067	88.555