Original Article

A case study employing non-parametric regression to develop a novel triglycerides model based on HDL levels using non-normally distributed data

Wan Muhamad Amir W Ahmad[1*], Farah Muna Mohamad Ghazali[1], Mohamad Nasarudin Adnan[1], Nor Azlida Aleng[2], Firdaus Mohamad Hamzah[3]

ABSTRACT

Background

This research models high-density lipoprotein (HDL) and triglyceride consumption using non-parametric regression, bootstrap resampling, and data splitting into training and testing sets to improve comprehension of their complex connection and inform evidence-based health promotion efforts.

Objective

The goal of this study is to develop a non-parametric regression model that connects HDL levels with triglyceride levels. This will help make predictions more accurate for HDL levels in the patients that were studied by using diagnostic tools.

Materials and Methods

When the assumption of linear regression is not satisfied, the model that is constructed may produce biased estimates. To get around this problem, a non-parametric regression model is used in this study, along with an improved bootstrap method to get the best model estimates. In particular, the study uses the Kendall-Theil Sen Siegel slope, a robust estimator, to find the slope of the regression line. This reduces the effect of outliers or values that are too high or too low. This method makes it easy to explore the relationship between variables without having to make rigid assumptions about the parameters. This method was used to split the dataset into training and testing groups. The training dataset will be used to build the model, and the testing dataset will be used to make sure the model works.

Result

Based on the statistical analysis conducted using R, it was found that the non-parametric regression method performed superiorly in making predictions, particularly in instances where the data didn't adhere to the assumption of normality. Notably, both the training and testing datasets yielded high R-squared values of 79.2% and 73.6% respectively under the non-parametric regression model. In contrast, when employing simple parametric regression, the R-squared values for the training and testing datasets were 45.89% and 47.29% respectively. A significantly high level of performance was attained using the proposed methodology, as demonstrated by these results.

Conclusion

The outcome of the research underscores the exceptional performance exhibited by the hybrid model methodology utilized.

Keywords

Non-parametric regression; Kendall-Theil Sen Siegel; High-density lipoprotein (HDL); Triglyceride

INTRODUCTION

Triglycerides and high-density lipoprotein (HDL) have a complicated and multidimensional interaction that affects cardiovascular health[1]. HDL has anti-atherogenic properties because it is a scavenger that returns cholesterol from peripheral tissues to the liver for processing and excretion[2]. On the other hand, increased triglyceride levels have the potential to interfere with lipid metabolism and damage HDL function, which can result in lower HDL levels and worse reverse cholesterol transport. Furthermore, as seen in dyslipidemic conditions, elevated triglyceride levels are frequently linked to lower HDL cholesterol levels[3,4]. Because of the reciprocal relationship between HDL and triglycerides, which emphasizes their linked roles in lipid metabolism and the aetiology of cardiovascular disease, controlling these variables is crucial to reducing the risk of cardiovascular disease5. Elevated triglyceride levels have been linked to atherosclerosis, a condition where fat deposits build up in the arterial walls. This narrowing and hardening of the arteries can obstruct blood flow to essential

1. Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia.
2. Universiti Malaysia Terengganu (UMT), 21030 Kuala Nerus, Terengganu, Malaysia.
3. Universiti Pertahanan Nasional Malaysia (UPNM), Kem Sungai Besi, 57000 Kuala Lumpur, Malaysia.

Correspondence

Wan Muhamad Amir W Ahmad, School of Dental Sciences, Health Campus, Universiti Sains Malaysia, 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia
Email: wmamir@usm.my. Cell Phone: +60169869306

DOI: https://doi.org/10.3329/bjms.v25i2.88743

organs[6,7]. Moreover, low levels of HDL cholesterol, obesity, insulin resistance, metabolic syndrome, and other risk factors are frequently present in concert with excessive triglycerides, raising the combined risk of cardiovascular problems[33, 36].

Furthermore, hypertriglyceridemia may indicate underlying metabolic diseases such as thyroid issues and diabetes mellitus[6, 8, 9]. To lower the risk of cardiovascular disease and enhance overall health outcomes, it is crucial to monitor and manage triglyceride levels through lifestyle modifications, such as implementing a healthy diet, increasing physical activity, controlling weight, and, if necessary, medication under the supervision of a healthcare provider[10, 11]. Elevated triglyceride levels can significantly affect a person's cardiovascular and overall health. Elevated triglyceride levels are associated with an increased risk of heart attacks, strokes, and coronary artery disease. Elevated triglyceride levels have been linked to atherosclerosis, a condition where fat deposits build up in the arterial walls. This narrowing and hardening of the arteries can obstruct blood flow to essential organs[12, 13, 14, 15]. Moreover, low levels of HDL cholesterol, obesity, insulin resistance, metabolic syndrome, and other risk factors are frequently present in concert with excessive triglycerides, raising the combined risk of cardiovascular problems[10, 14, 15]. Monitoring and managing triglyceride levels through lifestyle modifications, such as adopting a healthy diet, increasing physical activity, controlling weight, and, if needed, medication under a healthcare provider's supervision, is essential for lowering the risk of cardiovascular disease and improving overall health outcomes. Increased levels of high-density lipoprotein (HDL), sometimes known as "good cholesterol," are associated with improved cardiovascular health. Because of its scavenger function, high levels of HDL reduce the likelihood of cardiovascular disease by removing excess cholesterol from tissues outside of the arteries and transferring it to the liver for processing and excretion[5, 6, 35]. The cardioprotective effects of HDL are enhanced by its antioxidant, anti-inflammatory, and vasoprotective capabilities. When HDL levels are high, it usually means that you're leading a healthy lifestyle, which includes not smoking, eating well, and getting plenty of exercise. Having a high HDL cholesterol level is linked to better cardiovascular outcomes and general well-being, thus it's typically seen as favourable. Overall cardiovascular health needs to maintain ideal levels of both triglycerides and high-density lipoprotein (HDL) cholesterol. Elevations in HDL cholesterol, commonly referred to as "good cholesterol," are linked to a decreased risk of heart disease[7, 16, 17]. By carrying extra cholesterol to the liver for excretion, HDL keeps plaque from accumulating in the arteries, lowers the risk of atherosclerosis, and lowers the risk of cardiovascular events like heart attacks and strokes[18, 19].

On the other hand, high blood levels of triglycerides, a kind of fat, might harm cardiovascular health. Since high triglycerides contribute to the accumulation of plaque in the arteries, they are linked to an increased risk of atherosclerosis[16, 17, 18]. Moreover, low HDL cholesterol, obesity, insulin resistance, metabolic syndrome, and high triglyceride levels frequently coexist with one another as risk factors, hence raising the risk of cardiovascular disease[20, 37]. Thus, it is crucial to keep triglyceride and HDL cholesterol levels at appropriate ranges to protect cardiovascular health and lower the risk of heart disease. Understanding and controlling cardiovascular health greatly depends on modelling the link between triglycerides and high-density lipoprotein (HDL)[14, 15, 17]. Important elements of the lipid profile, HDL, and triglycerides are essential for lipid metabolism and the development of cardiovascular disease. Researchers and medical experts can more accurately determine cardiovascular risk and create focused strategies for treatment and prevention by clarifying the connection between these two lipid markers[21]. By identifying potential risk variables and underlying mechanisms, modeling this connection enables risk assessment and customized treatment options for cardiovascular disease. Moreover, by comprehending the interaction between HDL and triglycerides, pharmaceutical treatments, dietary adjustments, and lifestyle changes targeted at lowering cardiovascular risk and optimizing lipid levels can be made[15, 17, 21, 34]. Overall, improving patient outcomes, cardiovascular risk assessment, treatment, and HDL-triglyceride modeling is crucial. Researchers typically utilize regression analysis, such as linear regression or nonlinear regression, to model the relationship between HDL and triglycerides while controlling for potential confounding variables. The aim is to establish a mathematical equation that best describes how changes in HDL levels are associated with changes in triglyceride levels, or vice versa. Additionally, more advanced regression techniques like multiple regression or logistic regression may be employed to assess the impact of other factors, such as age, gender, diet, and lifestyle habits, on the

relationship between HDL and triglycerides. Through regression modeling, researchers can identify patterns, trends, and potential predictors of HDL and triglyceride levels, aiding in the understanding of lipid metabolism and cardiovascular disease risk[10, 18, 20, 21].

Parametric regression analysis can be employed to assess the relationship between HDL and triglycerides by fitting a predefined regression model, such as linear regression, which assumes a specific functional form for the relationship between the variables[5]. For instance, researchers may use parametric regression to determine if there is a linear association between HDL and triglyceride levels, accounting for potential confounders[22]. On the other hand, nonparametric regression techniques, like locally weighted scatterplot smoothing (LOWESS) or kernel regression, offer a flexible approach for modelling the relationship between HDL and triglycerides without assuming a specific functional form. This allows for the detection of more complex or nonlinear associations between the variables, which may not be captured by parametric models. Nonparametric regression can accommodate data that does not meet the assumptions of traditional parametric methods and provides valuable insights into the relationship between HDL and triglycerides across a wide range of data distributions and patterns[23]. Since parametric regression, like linear regression, is predicated on a particular mathematical model for the relationship between HDL and triglycerides, it is appropriate for determining linear correlations and estimating coefficients that have established meanings[5, 22].

However, because they do not assume a predetermined functional form, nonparametric regression techniques like kernel regression or locally weighted scatterplot smoothing (LOWESS) offer greater flexibility. This allows for the detection of complex or nonlinear relationships between HDL and triglycerides that parametric models might miss. The relationship between HDL and triglycerides across data distributions and patterns can be examined in greater detail using nonparametric regression, which is more complex and difficult to interpret than parametric regression[23, 24]

MATERIALS AND METHODS

The Data

This study utilized secondary data from the Hospital Universiti Sains Malaysia, with a sample size of fourteen individuals participating in the trial. Table 1 offers a concise summary of the data descriptions pertaining to the research variables.

Table 1. The data description

Variable	Description
HDL	High-density lipoprotein (HDL)
Trig	Sodium intake in milligrams (mg) per day

The Statistical Approach

(a) ***Simple Non-parametric Regression***

There are several benefits to using non-parametric regression, especially when predicting HDL levels from triglycerides. Non-parametric regression offers flexibility in efficiently capturing intricate patterns because it does not presuppose a particular mathematical relationship between variables[23]. This adaptability is especially useful when there isn't a clear equation that accurately describes the relationship between variables, like HDL and triglycerides. By using kernel regression techniques, non-parametric regression can capture non-linearities or changes that typical parametric models would miss by adapting to the underlying structure of the data without making strict assumptions[25]. The proposed non-parametric regression model for the case is given as follows:

$$HDL = \hat{a}_0 + s(Trig) + \mathring{a}_i \quad (1)$$

where

HDL is the response variable for the *i*-th observation

$\hat{a}_0$ is the constant / intercept

s(Trig) is the non-parametric smooth function of Triglycerides

(b) ***Simple Linear Regression***

Employing a simple linear regression to predict HDL levels based on triglycerides, with the assumption of normality fulfilled, presents significant advantages. This method provides a clear understanding of how fluctuations in triglyceride levels correspond to changes in HDL levels, offering valuable insights into cardiovascular health[26]. The proposed non-parametric regression model for the case is given as follows:

$$HDL = \hat{a}_0 + \hat{a}_1 (Trig) + \mathring{a}_i \quad (2)$$

where

HDL is the response variable for the *i*-th observation

$\hat{a}_0$ and $\hat{a}_1$ are the constant

Trig is the independent variable

(c) ***Bootstrapping and Data Splitting***

Non-parametric regression and bootstrapping can produce reliable model performance and parameter estimates. A resampling technique called bootstrapping iteratively draws samples with replacements from the dataset to assess model variability and uncertainty. Combining non-parametric regression approaches with bootstrapping and partitioning the data into training and testing sets helps researchers and data scientists estimate model parameters, analyze variability, and improve generalizability to novel data instances. The testing dataset evaluates the model, whereas the training dataset trains it. Segregation helps evaluate the model's generalization to new data. Training and testing sets are essential for building robust models that generalize well to new data, preventing overfitting and maintaining model reliability in real-world scenarios. Robust modeling requires bootstrapping and splitting data into training and testing sets. Bootstrapping, which generates numerous resampled datasets, is crucial for examining model parameters and performance statistic variability. Bootstrapping evaluates the model's stability and dependability by drawing samples with replacements from the original data. This method helps calculate confidence intervals, quantify model estimate uncertainty, and improve the statistical validity of findings. Bootstrapping also helps validate models by assessing their applicability to varied data subsets and identifying bias or overfitting[27, 28, 29].

Partitioning data into training and testing sets is crucial for model evaluation and development. This segregation helps evaluate a model's performance on unknown data, revealing its generalization beyond the training set[30]. This method helps identify overfitting, where a model performs well on training data but badly on new observations. The testing set is also essential for hyperparameter optimization, which optimizes model configurations for new data. The separation of data into training and testing sets eliminates data leakage and ensures that information from the testing set does not affect model training, resulting in a more accurate assessment of the model's real-world prediction skills[28, 31]. In conclusion, bootstrapping and data partitioning help create accurate, stable, and generalizable models that work across applications.

RESULTS AND DISCUSSION

In the results section, an in-depth analysis of the study carried out at Hospital Universiti Sains Malaysia is presented. Initially, the assessment of regression assumptions was undertaken to ascertain the appropriateness of either a parametric or non-parametric approach for the given scenario. Subsequently, upon evaluation, it was determined that the assumptions were not met, primarily due to the identification of unstandardized residuals with a *p*-value below 0.05.

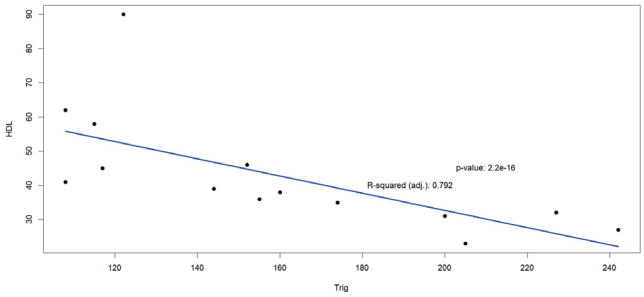

Figure 1. Non-parametric regression plot of HDL and Triglycerides

The first part of the analysis is the development of a non-parametric regression model. In this non-parametric regression model, the association between high-density lipoprotein (HDL) levels and triglyceride concentration is delineated by a flexible smoothing function denoted as '*s*(Triglycerides)'. Unlike traditional parametric regression techniques that presuppose a fixed functional structure, this non-parametric methodology affords heightened adaptability in capturing potentially nonlinear and intricate relationships between the variables.

Non-parametric Simple Regression

Using the training dataset, the model given as HDL = 42.5656 + 8.901·*s*(Trig) + $\mathring{a}_i$, which predicts HDL levels as a function of triglycerides, where 42.5656 denotes the baseline HDL level in the absence of triglycerides, and 8.901 serves as the coefficient quantifying the triglycerides' influence on HDL levels. The 's' function molds to the data without imposing rigid shape assumptions, facilitating the discernment of subtle variations and complexities in the triglycerides-HDL relationship. The model which uses the testing data is given as HDL = 42.4400+ 8.899 ·*s*(Trig) + $\mathring{a}_i$. This nonparametric approach enriches

Table 2. Regression analysis of HDL with Triglycerides reading

NON-PARAMETRIC REGRESSION					
Dependent Variable (*Y*)		Independent Variable (*X*)	Std. Coefficient Beta (β)	t-value* F-test#	*p*-value
Training Dataset	HDL	Constant	42.5656	173	<2e-16 ***
		s(Triglycerides)	8.901	382.3	<2e-16 ***
R^2 for the Training Dataset: 0.792					
The model : HDL = 42.5656 + 8.901 *s*(Trig) + $\mathring{a}_i$ ……..(3)					
Testing Dataset	HDL	Constant	42.4400	54.02	<2e-16 ***
		s(Triglycerides)	8.899	31.61	<2e-16 ***
R^2 for the Testing Dataset: 0.736					
The model : HDL = 42.4400+ 8.899 · *s*(Trig) + $\mathring{a}_i$ ……..(4)					
PARAMETRIC REGRESSION					
Dependent Variable (*Y*)		Independent Variable (*X*)	Std. Coefficient Beta (β)	t-value* F-test#	*p*-value
Training Dataset	HDL	Constant	80.08752	19.218	<2e-16 ***
		Triglycerides	-0.257810	-27.60	<2e-16 ***
R^2 for the Training Dataset: 0.4589					
The model : HDL = 80.08752-0.257810(Trig) + $\mathring{a}_i$ ……..(5)					
Testing Dataset	HDL	Constant	79.54118	19.62	<2e-16 ***
		Triglycerides	-0.23966	-9.376	<2e-16 ***
R^2 for the Testing Dataset: 0.4729					
The model : HDL = 79.54118-0.23966 (Trig) + $\mathring{a}_i$ ……..(6)					

The data employed did not follow a normal distribution.

our comprehension of the nuanced interplay between triglycerides and HDL levels, thereby contributing to a more thorough characterization of lipid metabolism dynamics. An R-squared value serves as a metric for gauging the extent to which the model accounts for the variability in the analyzed variable. Ranging between 0 and 1, with 1 representing a perfect fit, it quantifies the proportion of variance explained by the model. In this analysis, the model achieving R-squared values of 0.792(78.2%) for training and 0.736(73.6%) for testing suggests it performs admirably across both datasets, indicating a substantial level of explanatory power in

capturing the variability of the target variable.

Parametric Simple Regression

For the training model, the model was given as HDL = 80.08752-0.257810(Trig) + $\hat{a}_i$. The coefficient for Trig, -0.257810, indicates the slope of the line, illustrating the change in HDL levels associated with a one-unit increase in triglyceride levels. The intercept term, 80.08752, denotes the expected HDL level when triglyceride levels are zero, though this scenario might not be practically relevant. Overall, this model suggests a linear relationship between HDL and triglyceride levels, with HDL decreasing by approximately 0.257810 units for every one-unit increase in triglycerides. The R^2 is given as 45.89%. In this case, with an R-squared value of 45.89%, approximately 45.89% of the variability in HDL levels can be attributed to the linear relationship with triglyceride levels as captured by the model, suggesting a moderate level of explanatory power in elucidating the variation in HDL levels based on triglyceride levels.

For the testing model, the model was given as HDL = 79.54118-0.23966 (Trig) + $\hat{a}_i$. In this simple linear regression model, the HDL (high-density lipoprotein) levels are predicted based on triglyceride levels, as denoted by the equation HDL = 79.54118 - 0.23966(Trig). The coefficient for triglycerides, -0.23966, indicates the expected change in HDL levels for every one-unit increase in triglyceride levels, while the intercept term, 79.54118, represents the predicted HDL level when triglyceride levels are zero, though this scenario may not be practically relevant. The model's performance is evaluated using the R-squared value, which measures the proportion of variance in HDL levels explained by the model. With an R-squared value of 47.29%, approximately 47.29% of the variability in HDL levels can be accounted for by the linear relationship with triglyceride levels as depicted by the model, indicating a moderate degree of explanatory power in capturing the variation in HDL levels based on triglyceride levels.

A comparison was conducted between parametric and nonparametric simple regressions by evaluating the R-squared values of both the training and testing datasets. The initial examination was centered on the R-squared values, revealing that nonparametric regression consistently yields higher values compared to simple linear regression. Specifically, in nonparametric regression, both the training and testing datasets exhibit the highest R-squared values when contrasted with parametric regression. This observation underscores the superiority of nonparametric simple regression, particularly in scenarios where the data deviates from a normal distribution.

CONCLUSION

In comparing nonparametric and parametric regressions, it's evident that nonparametric models exhibit higher R-squared values, indicating better explanatory power. For the nonparametric regression models, both training and testing datasets yield notably high R-squared values-79.2% and 73.6%, respectively suggesting strong fits. The nonparametric models, represented by HDL = 42.5656 + 8.901 *s*(Trig) for training and HDL = 42.4400 + 8.899 *s*(Trig) for testing, utilize smoothing functions to capture complex relationships between HDL and triglyceride levels. Conversely, parametric regression models demonstrate lower R-squared values, with 45.89% for training and 47.29% for testing, indicating comparatively weaker fits. The parametric models, represented by HDL = 80.08752 - 0.257810(Trig) for training and HDL = 79.54118 - 0.23966(Trig) for testing, assume a predefined linear relationship between HDL and triglycerides. Overall, the results suggest that nonparametric regression offers superior performance, particularly in capturing the nuances of the relationship between HDL and triglyceride levels compared to parametric approaches.

Non-parametric regression analysis provides a practical solution in scenarios where conventional regression assumptions may be invalid, sample sizes are restricted, or outliers significantly impact observations, thus diminishing the reliability of the least squares method. Real-world data often strays from the idealized normal distribution, challenging traditional regression methodologies. This discrepancy underscores the need for robust regression techniques capable of accommodating non-normally distributed data and attenuating outlier effects[23, 25]. Non-parametric regression methods, devoid of specific parametric assumptions, offer enhanced adaptability and resilience in managing diverse data distributions, rendering them particularly advantageous in contexts where underlying data characteristics are less predictable or deviate from conventional norms. Non-parametric methods entail fewer assumptions regarding the population distribution underlying the sample compared to parametric methods[22]. While parametric approaches

often assume a predetermined distributional form, nonparametric methods do not, enabling greater adaptability to intricate datasets. This flexibility renders nonparametric techniques particularly useful for real-world data exhibiting atypical distributional characteristics. Consequently, nonparametric methods are employed when the distribution of data is unknown. This statistical modeling approach is characterized by its flexibility and robustness.

In conclusion, the employment of non-parametric regression models to explore the associations between HDL and Triglycerides offers a methodologically sound approach. Utilizing non-parametric methods, such as bootstrap resampling, ensures the robustness and reliability of estimating uncertainties linked with the model. Bootstrap resampling involves repeatedly sampling observations from the dataset to construct multiple datasets, thereby providing a more accurate estimation of model uncertainty. Furthermore, by incorporating both training and testing datasets, this modeling strategy permits the evaluation of predictive performance and generalizability. Training data are used to develop the model while testing data are utilized to assess how well the model can predict outcomes on new, unseen data. This practice offers valuable insights into the model's potential performance in real-world applications. The amalgamation of nonparametric regression, bootstrap resampling, and meticulous consideration of training and testing datasets enhances the precision and applicability of the models. This comprehensive approach facilitates a deeper understanding of the complex relationships between dietary factors and health outcomes, thereby contributing to advancements in health research and policy-making.

Acknowledgement

The authors wish to extend their appreciation to Universiti Sains Malaysia (USM) for generously funding this study via the Ministry of Higher Education (MOHE) Fundamental Research Grant Scheme (FRGS/1/2022/STG06/USM/02/10).

Conflicts of Interest

The authors affirm that they have no conflicts of interest to disclose.

Data Availability

This paper is an original work, and all data contained within it are accessible solely for research purposes through the principal investigators.

Author's Contribution

Data gathering and idea owner of this study: WMAWA, FMMG, MNA, NAA.

Study design: WMAWA, FMMG, MNA, NAA, FMH.

Data gathering: WMAWA, FMMG, MNA, NAA, FMH.

Writing and submitting a manuscript: WMAWA, FMMG, MNA, NAA, FMH.

Editing and approval of final draft: WMAWA, FMMG, MNA, NAA.

Appendix

 # simple nonparametric regression

if(!require(psych)){install.packages("psych")}

if(!require(mblm)){install.packages("mblm")}

if(!require(quantreg)){install.packages("quantreg")}

if(!require(rcompanion)){install.packages("rcompanion")}

if(!require(mgcv)){install.packages("mgcv")}

if(!require(lmtest)){install.packages("lmtest")}

if(!require(Rfit)){install.packages("Rfit")}

data=read.table(header=TRUE, stringsAsFactors=TRUE, text="

HDL Trig

39 144

23 205

32 227

38 160

90 122")

Order factors by the order in data frame

Otherwise, R will alphabetize them

mydata <- rbind.data.frame(data, stringsAsFactors = FALSE)

iboot <- sample(1:nrow(mydata),size=1000, replace = TRUE)

Bootdata <- mydata[iboot,]

library(psych)

headTail(Bootdata)

str(Bootdata)

```
summary(Bootdata)
index = sample(1:nrow(Bootdata),round(0.90*nrow(Bootdata)))
train_data <- as.data.frame(Bootdata[index,])
test_data <- as.data.frame(Bootdata[-index,])
# Print Data
print(train_data)
print(test_data)
############### Modeling Train Data ###############################
library(mgcv)
model.g = gam( HDL~ s(Trig),
          data = train_data,
          family=gaussian())
summary(model.g)
model.null = gam(HDL ~ 1,
            data = train_data,
            family=gaussian())
anova(model.g,
    model.null)
train_data$HDL = as.numeric(train_data$HDL)
train_data$HDL2 = train_data$HDL ^ 2
model.g = lm(HDL ~ Trig, data = train_data)
R2_Training<-efronRSquared(actual=train_data$HDL,
residual=model.g$residuals)
R2_Training
print(R2_Training)
############### Plotting Data ################################
library(lmtest)
lrtest(model.g,
     model.null)
library(rcompanion)
plotPredy(data  = Bootdata,
       x    = Trig ,
       y    = HDL,
       model = model.g,
       xlab  = "Trig",
       ylab  ="HDL" )
Pvalue   = 2.2e-16
R2       = 0.742
t1    = paste0("p-value: ", signif(Pvalue, digits=3))
t2    = paste0("R-squared (adj.): ", signif(R2, digits=3))
text(210, 43, labels = t1, pos=3.8)
text(180, 40, labels = t2, pos=4)
############### Prediction Testing Data ####################
options(warn=-1)
library(mgcv)
model.g1 = gam( HDL~ s(Trig),
          data = test_data,
          family=gaussian())
summary(model.g1)
R2_Testing<-efronRSquared(residual=model.g1$residuals, predicted=model.g1$fitted.values)
R2_Testing
print(R2_Testing)
# R-square for Training data and R-square for Testing data
print(paste(R2_Training,R2_Testing))
#Plots of residuals
 x = residuals(model.g )
if(!require(rcompanion)){install.packages("rcompanion")}
 library(rcompanion)
 plotNormalHistogram(x)
##########################################################
# for simple linear regression
Model <- lm(HDL~Trig,data=train_data)  # build the model
summary(Model)
R2_Training1<-efronRSquared(actual=train_data$HDL,
residual=Model$residuals)
```

```
R2_Training1
print(R2_Training1)
##########################################################################
Model1 <- lm(HDL~Trig,data=test_data)  # build the model
summary(Model1)
R2_Testing1<-efronRSquared(residual=Model1$residuals, predicted=Model1$fitted.values)
R2_Testing1
print(R2_Testing1)
# R-square for Training data and R-square for Testing data
print(paste(R2_Training1,R2_Testing1))
```

REFERENCES

1. Von Bibra, H., Saha, S., Hapfelmeier, A., Müller, G., & Schwarz, P. E. H. (2017). Impact of the Triglyceride/High-Density Lipoprotein Cholesterol Ratio and the Hypertriglyceremic-Waist Phenotype to Predict the Metabolic Syndrome and Insulin Resistance. *Hormone and Metabolic Research*, **49**(7): 542-549. https://doi.org/10.1055/s-0043-107782
2. Girona, J., Amigó, N., Ibarretxe, D., Plana, N., Rodríguez-Borjabad, C., Heras, M., Ferré, R., Gil, M., Correig, X., & Masana, L. HDL Triglycerides: A New Marker of Metabolic and Cardiovascular Risk. *International journal of molecular sciences*, 2019;**20**(13): 3151. https://doi.org/10.3390/ijms20133151
3. Iwani, N. A. K. Z., Jalaludin, M. Y., Zin, R. M. W. M., Fuziah, M. Z., Hong, J. Y. H., Abqariyah, Y., Wan Nazaimoon, W. M. Triglyceride to HDL-C ratio is Associated with Insulin Resistance in Overweight and Obese Children. Scientific Reports, 2017;**7**(1): Iwani, N. A., Jalaludin, M. Y., Zin, R. M., Fuziah, M. Z., Hong, J. Y., Abqariyah, Y., Mokhtar, A. H., & Wan Nazaimoon, W. M. Triglyceride to HDL-C Ratio is Associated with Insulin Resistance in Overweight and Obese Children. *Scientific reports*, 7, 40055. https://doi.org/10.1038/srep40055
4. Vrigazova, B. The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems. *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy*, 2021;**12**(1): 228-242. https://doi.org/10.2478/bsrj-2021-0015
5. Ferrando, C., Wang, S., & Sheldon, D. (2022, May). Parametric Bootstrap for Differentially Private Confidence Intervals. *In International Conference on Artificial Intelligence and Statistics*, 1598-1618.
6. Albers, J. J., Slee, A., Fleg, J. L., O'Brien, K. D., & Marcovina, S. M. Relationship of baseline HDL subclasses, small dense LDL and LDL triglyceride to cardiovascular events in the AIM-HIGH clinical trial. *Atherosclerosis*, 2016;**251**:454–459. https://doi.org/10.1016/j.atherosclerosis.2016.06.019
7. Braunwald, E. (). Gliflozins in the Management of Cardiovascular Disease. *The New England journal of medicine*, 2022;**386**(21): 2024–2034. https://doi.org/10.1056/NEJMra2115011
8. Zhang, T., Tang, X., Mao, L., Chen, J., Kuang, J., Guo, X., Xu, D., Peng, D., & Yu, B. (). HDL-associated apoCIII plays an independent role in predicting postprandial hypertriglyceridemia. *Clinical biochemistry*, 2020;**79**: 14–22. https://doi.org/10.1016/j.clinbiochem.2020.02.004
9. Lanktree, M. B., Thériault, S., Walsh, M., & Paré, G. HDL Cholesterol, LDL Cholesterol, and Triglycerides as Risk Factors for CKD: A Mendelian Randomization Study. *American journal of kidney diseases : the official journal of the National Kidney Foundation*, 2018;**71**(2): 166–172. https://doi.org/10.1053/j.ajkd.2017.06.011
10. Pantoja-Torres, B., Toro-Huamanchumo, C. J., Urrunaga-Pastor, D., Guarnizo-Poma, M., Lazaro-Alcantara, H., Paico-Palacios, S., Del Carmen Ranilla-Seguin, V., Benites-Zapata, V. A., & Insulin Resistance and Metabolic Syndrome Research Group. High triglycerides to HDL-cholesterol ratio is associated with insulin resistance in normal-weight healthy adults. *Diabetes & metabolic syndrome*, 2019;**13**(1): 382–388. https://doi.org/10.1016/j.dsx.2018.10.006
11. Yeh, W. C., Tsao, Y. C., Li, W. C., Tzeng, I. S., Chen, L. S., & Chen, J. Y. Elevated triglyceride-to-HDL cholesterol ratio is an indicator for insulin resistance in middle-aged and elderly Taiwanese population: a cross-sectional study. *Lipids in health and disease*, 2019;**18**(1):176. https://doi.org/10.1186/s12944-019-1123-3
12. Weinstein, S., Maor, E., Kaplan, A., Hod, T., Leibowitz, A., Grossman, E., & Shlomai, G. Non-Interventional Weight Changes Are Associated with Alterations in Lipid Profiles and in the Triglyceride-to-HDL Cholesterol Ratio. *Nutrients*, 2024;**16**(4): 486. https://doi.org/10.3390/nu16040486
13. Timmis, A., Townsend, N., Gale, C., Grobbee, R., Maniadakis, N., Flather, M., Wilkins, E., Wright, L., Vos, R., Bax, J., Blum, M., Pinto, F., Vardas, P., & ESC Scientific Document Group. European Society of Cardiology: Cardiovascular Disease Statistics 2017. *European heart journal*, 2017;**39**(7): 508–579. https://doi.org/10.1093/eurheartj/ehx628
14. Guan, M., Wu, L., Cheng, Y., Qi, D., Chen, J., Song, H., Hu, H., & Wan, Q. (2024). Defining the threshold: triglyceride to high-density lipoprotein cholesterol (TG/HDL-C) ratio's non-linear impact on tubular atrophy in primary membranous nephropathy. *Frontiers in endocrinology*, 2024;**15**: 1322646. https://doi.org/10.3389/fendo.2024.1322646
15. Jacobs, D. R., Jr, Woo, J. G., Sinaiko, A. R., Daniels, S. R., Ikonen, J., Juonala, M., Kartiosuo, N., Lehtimäki, T.,

Magnussen, C. G., Viikari, J. S. A., Zhang, N., Bazzano, L. A., Burns, T. L., Prineas, R. J., Steinberger, J., Urbina, E. M., Venn, A. J., Raitakari, O. T., & Dwyer, T. Childhood Cardiovascular Risk Factors and Adult Cardiovascular Events. *The New England journal of medicine*, 2022;**386**(20): 1877–1888. https://doi.org/10.1056/NEJMoa2109191

16. Puri, R., Nissen, S. E., Shao, M., Elshazly, M. B., Kataoka, Y., Kapadia, S. R., Tuzcu, E. M., & Nicholls, S. J. (2016). Non-HDL Cholesterol and Triglycerides: Implications for Coronary Atheroma Progression and Clinical Events. *Arteriosclerosis, thrombosis, and vascular biology*, 2016;**36**(11): 2220–2228. https://doi.org/10.1161/ATVBAHA.116.307601
17. Sultani, R., Tong, D. C., Peverelle, M., Lee, Y. S., Baradi, A., & Wilson, A. M. Elevated Triglycerides to High-Density Lipoprotein Cholesterol (TG/HDL-C) Ratio Predicts Long-Term Mortality in High-Risk Patients. *Heart, lung & circulation*,2020; **29**(3): 414–421. https://doi.org/10.1016/j.hlc.2019.03.019
18. Čížek, P., & Sadıkoğlu, S. Robust Nonparametric Regression: A Review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2020;**12**(3): e1492. https://doi.org/10.1002/wics.1492
19. Ma, L. Y., Chen, W. W., Gao, R. L., Liu, L. S., Zhu, M. L., Wang, Y. J., Wu, Z. S., Li, H. J., Gu, D. F., Yang, Y. J., Zheng, Z., & Hu, S. S. China cardiovascular diseases report 2018: an updated summary. *Journal of geriatric cardiology : JGC*, 2020;**17**(1): 1–8. https://doi.org/10.11909/j.issn.1671-5411.2020.01.001
20. Kjeldsen S. E. (2018). Hypertension and cardiovascular risk: General aspects. *Pharmacological research*, 129, 95–99. https://doi.org/10.1016/j.phrs.2017.11.003
21. Simar, L., Wilson, P. W. Hypothesis Testing in Nonparametric Models of Production Using Multiple Sample Splits. *Journal of Productivity Analysis*, 2020;**53**, 287-303. https://doi.org/10.1007/s11123-020-00574-w
22. Toth P. P. Triglyceride-rich lipoproteins as a causal factor for cardiovascular disease. *Vascular health and risk management*, 2016;**12**: 171–183. https://doi.org/10.2147/VHRM.S104369
23. Cavaliere, G., Gonçalves, S., Nielsen, M., Zanelli, E. (2023). Bootstrap Inference in The Presence of Bias. *Journal of the American Statistical Association*, 1-26.
24. Wakabayashi, I., & Daimon, T. Comparison of discrimination for cardio-metabolic risk by different cut-off values of the ratio of triglycerides to HDL cholesterol. *Lipids in health and disease*, 2019;**18**(1): 156. https://doi.org/10.1186/s12944-019-1098-0
25. Chen, C., & Dai, J. L. Triglyceride to high-density lipoprotein cholesterol (HDL-C) ratio and arterial stiffness in Japanese population: a secondary analysis based on a cross-sectional study. *Lipids in health and disease*, 2018;**17**(1): 130. https://doi.org/10.1186/s12944-018-0776-7
26. Tada, H., Nohara, A., & Kawashiri, M. A. Serum Triglycerides and Atherosclerotic Cardiovascular Disease: Insights from Clinical and Genetic Studies. *Nutrients*, 2018;**10**(11): 1789. https://doi.org/10.3390/nu10111789
27. Budoff M. Triglycerides and Triglyceride-Rich Lipoproteins in the Causal Pathway of Cardiovascular Disease. *The American journal of cardiology*, 2016;**118**(1): 138–145. https://doi.org/10.1016/j.amjcard.2016.04.004
28. Duran, E. K., Aday, A. W., Cook, N. R., Buring, J. E., Ridker, P. M., & Pradhan, A. D. Triglyceride-Rich Lipoprotein Cholesterol, Small Dense LDL Cholesterol, and Incident Cardiovascular Disease. *Journal of the American College of Cardiology*, 2020;**75**(17): 2122–2135. https://doi.org/10.1016/j.jacc.2020.02.059
29. Taillie, L. S., Bercholz, M., Popkin, B., Reyes, M., Colchero, M. A., & Corvalán, C. . Changes in food purchases after the Chilean policies on food labelling, marketing, and sales in schools: a before and after study. *The Lancet. Planetary health*, 2021;**5**(8): e526–e533. https://doi.org/10.1016/S2542-5196(21)00172-8
30. Nordestgaard B. G. Triglyceride-Rich Lipoproteins and Atherosclerotic Cardiovascular Disease: New Insights From Epidemiology, Genetics, and Biology. *Circulation research*, 2016;**118**(4): 547–563. https://doi.org/10.1161/CIRCRESAHA.115.306249
31. Skalidis, I., Muller, O., & Fournier, S. CardioVerse: The cardiovascular medicine in the era of Metaverse. *Trends in cardiovascular medicine*, 2023;**33**(8): 471–476. https://doi.org/10.1016/j.tcm.2022.05.004
32. Ankaralı, H., Pasin, Ö., Gönenç, S., & Al-Mahmood, A. K. Interaction between numerical variables in regression model, and its graphical interpretation. *Bangladesh Journal of Medical Science*, 2023;**22**(1): 189–194. https://doi.org/10.3329/bjms.v22i1.63078
33. Al-Mahmood, A. K., Ismail, A. A., R, F. A., Wan Bebakar, W. M., & Tai, E. S. The metabolic syndrome in normal weight Malay subjects. Bangladesh Journal of Medical Science, 2016;**15**(1): 123–128. https://doi.org/10.3329/bjms.v15i1.27149
34. Afrin, S. F., Mahmood, A. K. A., Bari, K. F., Rahman, F., & Hassan, Z. Pattern of lipid levels of subjects seeking laboratory services in an established laboratory in the Dhaka city. Bangladesh Journal of Medical Science, 2017;**16**(3): 375–379. https://doi.org/10.3329/bjms.v16i3.32849
35. Goel, S., Garg, P. K., Malhotra, V., Madan, J., Mitra, S., & Grover, S. Dyslipidemia in Type II Diabetes Mellitus - An assessment of the main lipoprotein abnormalities. *Bangladesh Journal of Medical Science*, 2016;**15**(1): 99–102. https://doi.org/10.3329/bjms.v15i1.21170
36. Al-Mahmood, A. K. SF Afrin, N Hoque. Dyslipidemia in Insulin Resistance: Cause or Effect. 2014.*Bangladesh Journal of Medical Biochemistry* **7**(1): https://doi.org/10.3329/bjmb.v7i1.18576
37. Al-Mahmood, A. K. SF Afrin, N Hoque. Metabolic Syndrome and Insulin Resistance: Global Crisis 2013.Bangladesh Journal of Medical Biochemistry 2013;**4**(1): https://doi.org/10.3329/bjmb.v4i1.13779