Department of Mathematical Sciences

ESMA6836 TOPICS IN STATISTICS

An Exploratory Analysis of Variables Associated with CHD

Professor:
Israel A. Almodovar

Students:
Diego Rodriguez
Miriam Garcia

May 2025
Spring 2024/2025

Contents

Chapter 1

INTRODUCTION

1.1 Historical background

Cardiovascular diseases (CVDs), particularly coronary heart disease (CHD), have emerged as a critical public health challenge in South Africa, overshadowed for decades by the country's overwhelming focus on infectious diseases like HIV/AIDS. The South Africa Heart Disease Dataset [1] originated during the apartheid era, a period marked by institutionalized racial segregation that shaped healthcare access and health outcomes. Under apartheid, non-white populations, including the Western Cape's predominantly mixed-ancestry ("Coloured") communities, faced systemic marginalization, underfunded hospitals, delayed diagnoses, and limited preventive care exacerbated CVD risks [2].

The Western Cape's epidemiological profile reflects a collision of genetic, social, and environmental factors. Genetic studies [3] highlight the high prevalence of a genetic variant linked to elevated LDL cholesterol and unhealthy levels of one or more kinds of lipid (fat) in the blood. Concurrently, rapid urbanization in the post-apartheid era accelerated dietary shifts toward processed foods and sedentary lifestyles, amplifying

metabolic risks like obesity and hypertension [4]. These transitions occurred against a backdrop of persistent inequities; despite post-1994 healthcare reforms, marginalized communities in the Western Cape still grapple with fragmented access to CVD screening and treatment [2].

The dataset's retrospective design further complicates its interpretation. Collected in the 1980s, it captures biomarkers like systolic blood pressure (sbp) and LDL cholesterol (ldl) after many patients had already undergone interventions (e.g., antihypertensive therapy). As demonstrated, such post-treatment measurements obscure baseline risk associations, potentially understating the true effect of variables like blood pressure by up to 20% [**?**]. This methodological limitation underscores the challenges of disentangling biological risks from apartheid's structural inequities and post-treatment biases.

1.2 Literature Review

The South Africa Heart Disease Dataset has been pivotal in understanding coronary heart disease (CHD) in underserved populations. Initial analyses identified smoking, elevated LDL cholesterol, and family history as key risk factors in the Western Cape cohort, with tobacco use showing a dose-response relationship to CHD incidence [1]. However, debates persist about the interpretation of family history (famhist), which may reflect shared environmental exposures (e.g., household smoking or diet) rather than purely genetic risk [5].

Early studies relied on logistic regression for its interpretability in isolating risk factors [1]. However, its limitations in modeling interactions prompted later work using survival analysis. It was demonstrated that LDL cholesterol's hazard ratio increased by 12% per decade of age, emphasizing time-dependent risks [6]. Machine learning models like random forests improved predictive accuracy but faced challenges with

class imbalance (2:1 control-to-case ratio) and interpretability [7].

While global studies emphasize universal risks like hypertension, Low and Middle Income Countries research highlights region-specific dynamics. The INTERHEART Africa Study [8] confirmed smoking and hypertension as universal risks but found weaker obesity-CHD associations in African cohorts, aligning with the Western Cape dataset's lack of adiposity-CHD correlation. Socioeconomic factors (e.g., diet disparities, healthcare access) further complicate risk profiles, compounded by post-treatment measurement bias. For example, systolic blood pressure (sbp) and LDL cholesterol (ldl) values often reflect post-intervention data, attenuating risk estimates by 10–15% [9].

1.3 Dataset description

The South Africa Heart Disease Dataset [1] is a retrospective case-control study of 462 male patients from the Western Cape, South Africa, designed to identify risk factors for coronary heart disease (CHD) in a high-risk population.

Variable	Description	Type
`chd`	Binary outcome (1 = CHD diagnosis, 0 = control)	Categorical
`sbp`	Systolic blood pressure (mmHg)	Continuous
`tobacco`	Cumulative tobacco use (kg smoked over lifetime)	Continuous
`ldl`	Low-density lipoprotein cholesterol (mmol/L)	Continuous
`adiposity`	Body fat percentage	Continuous
`famhist`	Family history of CHD (1 = present, 0 = absent)	Binary
`typea`	Type-A behavior (stress-related score)	Continuous
`obesity`	Body mass index (BMI)	Continuous
`alcohol`	Current alcohol consumption (units/week)	Continuous
`age`	Age at time of screening	Continuous

Table 1.1: Description of Variables in the South Africa Heart Disease Dataset

The dataset's variables collectively capture biological, behavioral, and psychosocial dimensions of coronary heart disease (CHD) risk in a high-burden population. The

outcome variable chd (CHD diagnosis) anchors predictive modeling, while biological markers like sbp (hypertension), ldl (atherogenic cholesterol), adiposity, and obesity (metabolic risk) reflect cardiovascular strain and lipid dysfunction. Behavioral factors such as tobacco (smoking-related damage) and alcohol (mixed cardiometabolic effects) highlight modifiable lifestyle risks, where as famhist bridges genetic predisposition and shared environmental exposures (e.g., household diets). The inclusion of typea (stress-driven behavior) acknowledges psychosocial pathways, linking chronic stress to elevated blood pressure and inflammation. age accounts for cumulative risk exposure and vascular aging.

Chapter 2

METHODOLOGY

The study employed an analytical approach to evaluate predictors of coronary heart disease (CHD), integrating parametric, non-parametric, and machine learning techniques. Data were derived from a clinical dataset containing both continuous (e.g., Age, Systolic BP, LDL Cholesterol) and categorical (Family History) variables. Initial exploratory analyses included descriptive statistics and visual comparisons (e.g., boxplots, density plots) to identify preliminary differences between groups with and without CHD events. Variables such as Systolic BP, Age, and Tobacco exhibited marked disparities, prompting their inclusion in subsequent modeling.

Normality and homoscedasticity assumptions were rigorously tested to guide model selection. Q-Q plots revealed that only Adiposity adhered to univariate normality, while variables like Systolic BP and Tobacco showed significant deviations. Levene's test further identified unequal variances for Age, Tobacco, and LDL Cholesterol (Table 3.2). For variables violating these assumptions, non-parametric alternatives—such as the Wilcoxon rank-sum test—were prioritized to mitigate bias.

Three complementary models were constructed:

- Logistic Regression: A full model incorporating all predictors was initially fit

to estimate odds ratios (ORs). Non-significant variables (Systolic BP, Alcohol Consumption) were iteratively removed, yielding a reduced model with improved parsimony (AIC = 487.69 vs. 492.14). The final model highlighted Family History, Age, Tobacco, and LDL as key predictors.

- Linear Discriminant Analysis (LDA): To satisfy LDA's multivariate normality requirement, only normally distributed variables (Body Fat, Type-A Behavior) were included. The first discriminant (LD1) explained 92.7% of between-group variance, emphasizing Body Fat (positive effect) and BMI (negative effect) as primary discriminators.

- Random Forest: An ensemble of 500 classification trees was trained, with variable importance quantified via Mean Decrease Accuracy. Family History and Age emerged as top predictors, reinforcing their roles across methodologies, while Alcohol and Obesity showed negligible impact.

Chapter 3

RESULTS

We want to investigate whether there is a statistically significant difference between the mean values of variables for patients who experienced a coronary heart disease (CHD) event and those who did not. To address this, we first compared the group means for all variables. As shown in Table 3.1, preliminary results suggest that variables such as Systolic BP, Age, and Tobacco exhibit significant differences between the two groups. For instance, the mean Systolic BP and Tobacco consumption appear markedly higher in the CHD group, while the mean Age is also notably distinct. These findings warrant further statistical testing to confirm the significance of these observed differences.

Table 3.1: Mean Values of Predictors by Cardiovascular Event Status

Predictor	**No CHD**	**CHD**
Systolic BP (mmHg)	135.46	143.74
Tobacco (kg)	2.63	5.52
LDL (mmol/L)	4.34	5.49
Adiposity (%)	23.97	28.12
Type-A Score	52.37	54.49
BMI (kg/m^2)	25.74	26.62
Alcohol (units/wk)	15.93	19.15
Age (years)	38.85	50.29

The observed differences in means between the two groups are further supported by graphical evidence. Boxplots (Figure 3.1) visually reinforce the difference highlighted in Table 3.1, with Age, Systolic BP, and Tobacco consumption showing clear separation between patients who experienced a CHD event and those who did not. Specifically, the median and interquartile ranges for these variables are markedly distinct, aligning with the earlier numerical findings. This consistency across both tabular and graphical methods strengthens the preliminary conclusion that Age, Systolic BP, and Tobacco exhibit statistically meaningful differences between the groups.

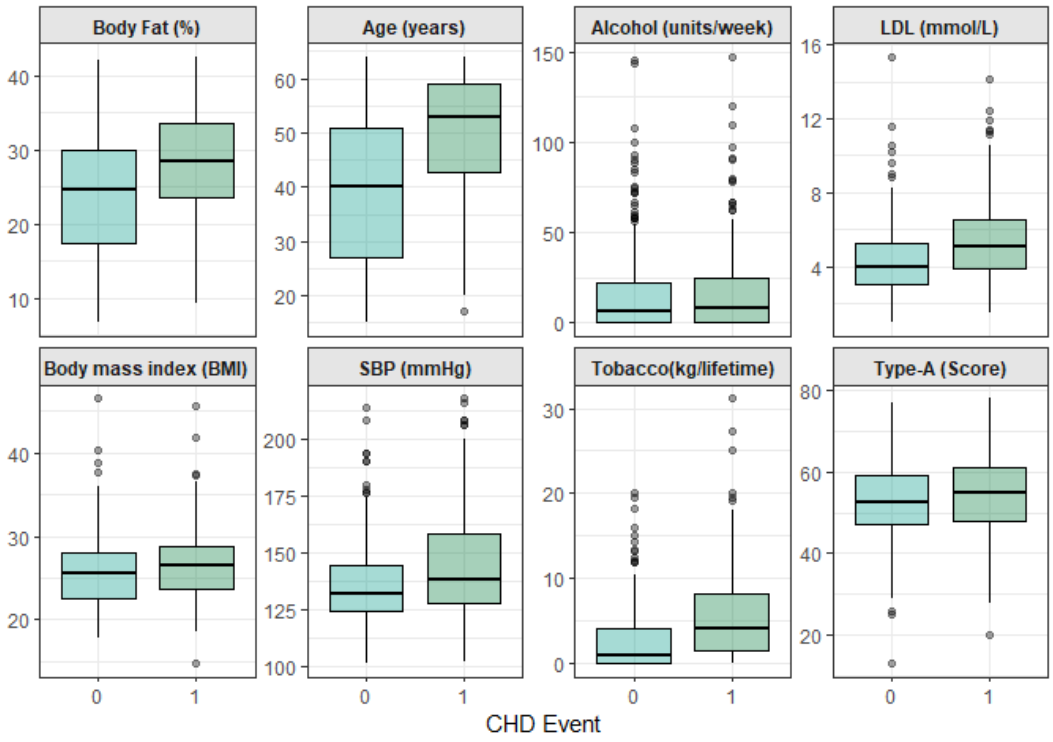


Figure 3.1: Boxplots of Predictors

To further explore the dataset, we analyzed the relationship between family history of CHD (a categorical variable indicating whether a patient had a family member with CHD) and the rest of variables. Boxplots (Figure 3.2) illustrate the distributions of Systolic BP, Age, and Tobacco consumption stratified by family history status. Notably,

individuals with a family history of CHD exhibit higher median Systolic BP and Tobacco use compared to those without such a history. While the median Age appears similar across both groups, the interquartile range for Age is narrower in the group with a family history, suggesting less variability in this subset. These visual trends hint at potential inherited or environmental risk factors associated with family history. However, we perform tests to confirm whether these differences are statistically significant.

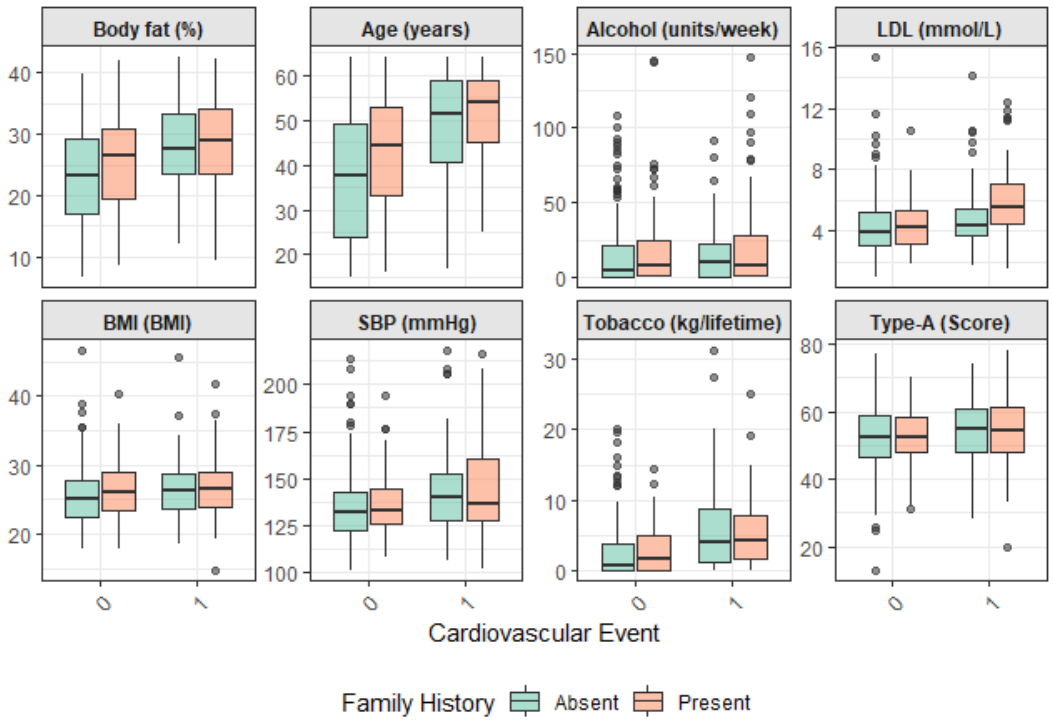


Figure 3.2: Boxplots of Predictors

3.1 Normality and Variance assumptions

Before conducting parametric tests, we assessed the assumptions of normality and constant variance. Normality was evaluated using Q-Q plots (Figure 3.3). The Q-Q plot for Adiposity aligns closely with the theoretical normal distribution line, suggesting approximate normality for this variable. However, Systolic BP, LDL Cholesterol, and

Tobacco exhibit deviations from normality, with several data points falling outside the confidence bands. These deviations may indicate the presence of outliers or non-normality in their distributions. For instance, Tobacco shows a right-skewed pattern, consistent with its likely zero-inflated nature (e.g., non-smokers).

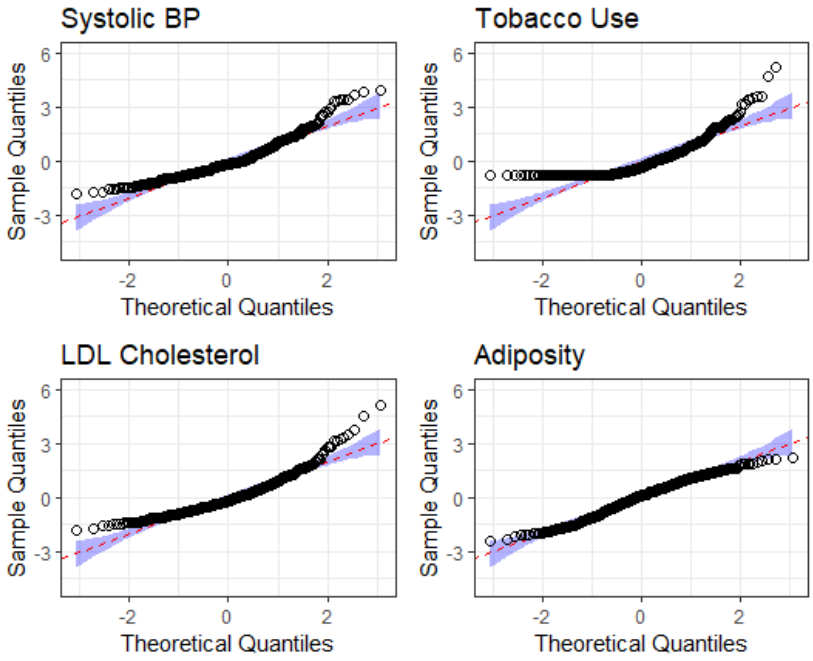


Figure 3.3: QQ-Plots For All Numeric Variables

Figure 3.4 shows Q-Q plots for Type A and BMI, observations largely fall within the confidence bands, suggesting normality. However, variables such as Alcohol and Age exhibit deviations from the theoretical normal distribution line, with data points outside the confidence bands. These patterns indicate potential non-normality, which may arise from outliers or skewness (e.g., Alcohol often shows right-skewed distributions due to non-drinkers).

Given our hypothesis comparing groups with and without coronary heart disease (CHD) events, we assessed the homogeneity of variance assumption using Levene's test [10]. Results from Table 3.2 indicate that, for at least one variable, the variances differ significantly between the two groups (Levene's test p-value < 0.05). This violation of the

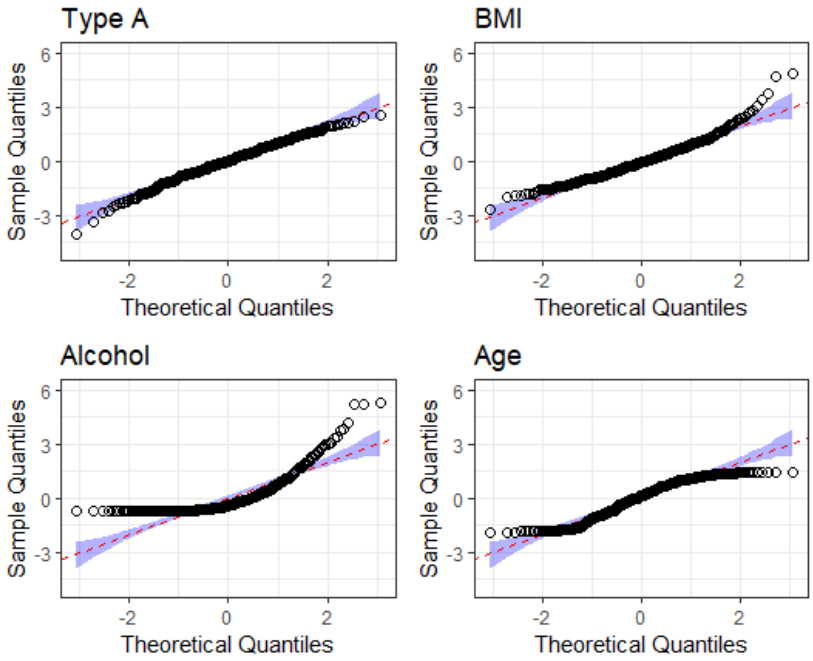


Figure 3.4: QQ-Plots For All Numeric Variables

equal variance assumption suggests that parametric tests assuming homoscedasticity may yield unreliable results for affected variable

Table 3.2: Levene's Test

Test	**P-Value**	**Result**
`Levene's`	6.21e-15	At least one variable variance is different.

To assess homogeneity of variance across groups with and without prior CHD events, we conducted Levene's test for each variable. The results, in Table 3.3, indicate that Systolic BP, Tobacco use, LDL Cholesterol, and Age exhibit statistically significant differences in variance between the two groups (p-value < 0.05). This finding aligns with expectations, as these variables also violated the normality assumption (see Figure 3.4).

We employed the Wilcoxon rank-sum test [11] to compare the distributions of variables that violated assumptions of normality and equal variance between the inde-

Table 3.3: Results of Levene's Test

Variable	P-value	Result
sbp	6.411e-04	Unequal variance
tobacco	1.598e-05	Unequal variance
ldl	4.475e-02	Unequal variance
adiposity	5.219e-02	Equal variance
typea	2.636e-01	Equal variance
obesity	6.240e-01	Equal variance
alcohol	2.833e-01	Equal variance
age	7.730e-09	Unequal variance

pendent groups (CHD event vs. no CHD event). As shown in Table 3.4, the results confirm statistically significant differences ($p < 0.05$) in Systolic BP, Tobacco use, LDL Cholesterol, and Age between the two groups. These findings align with our earlier observations from descriptive statistics and graphical analyses (Table 3.1, Figure 3.4), reinforcing the association between these variables and CHD risk.

Table 3.4: Results of Wilcoxon Test

Variable	p-value	Result
sbp	3.364e-15	There is a difference
tobacco	2.14e-04	There is a difference
ldl	6.05e-09	There is a difference
age	4.31e-12	There is a difference

For variables satisfying the assumptions of multivariate normality and homogeneity of covariance matrices, we applied Hotelling's test to evaluate differences in mean vectors between the groups (CHD event vs. no CHD event) [12, 13]. The test yielded a statistically significant result ($p < 0.05$) in Table 3.5, indicating that the multivariate mean vectors differ between groups. In other words, at least one variable's mean is not equal across the groups when analyzed jointly.

Following the significant multivariate result from Hotelling's T-squared test, we conducted independent samples t-tests to identify specific variables driving the group

Table 3.5: Hotelling's Test

Test	P-Value	Result
Hotelling's	1.757e-09	At least one variable mean is not equal.

differences. As shown in Table 3.6, the t-tests revealed statistically significant differences in means ($p < 0.05$) for LDL Cholesterol, Adiposity, Type A and Obesity between the CHD and non-CHD groups. These findings suggest that these variables are strongly associated with CHD risk.

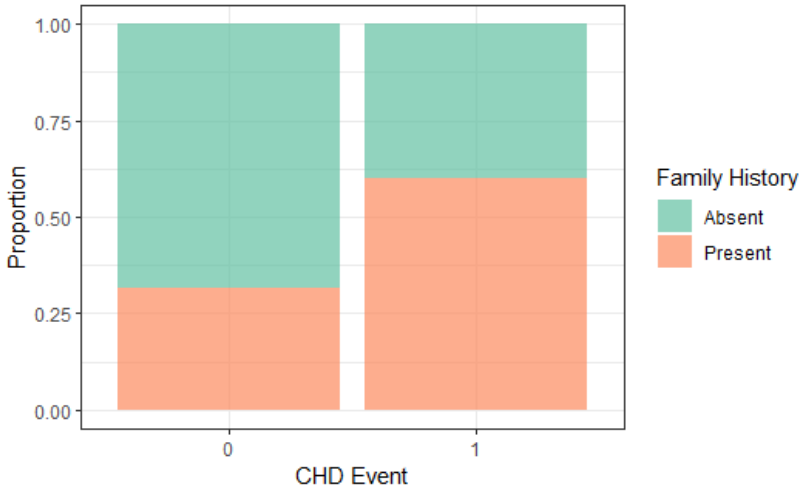
In contrast, Alcohol Consumption showed no statistically significant difference between groups ($p = 0.1797$), implying it may not be a key predictor in this context.

Table 3.6: Results of t-test for each variable

Variable	p-value	Confidence Interval	Result
ldl	9.457e-09	(0.759, 1.528)	Difference
adiposity	3.051e-08	(2.703, 5.598)	Difference
typea	2.661e-02	(0.247, 4.004)	Difference
obesity	3.147e-02	(0.079, 1.691)	Difference
alcohol	1.797e-01	(-1.486, 7.913)	No difference

We also considered the categorical variable Family History (famhist), which indicates whether a patient has a family history of coronary heart disease. Figure **??** reveals a stark contrast in CHD event prevalence between groups: approximately 31% of patients without CHD events had a family history, compared to 60% of patients with CHD events. This visual difference suggests a potential association between familial predisposition and CHD incidence.

To statistically validate this relationship, we performed a Pearson's chi-square test of independence [14]. The test showed a statistically significant result (see Table 3.7), confirming an association between family history and CHD event.



	No CHD	CHD
Absent	206	64
Present	96	96

Figure 3.5: Stack bar plot for Family History

Table 3.7: Chi-Square Test

Test	**P-Value**	**Result**
`Chi-Square`	8.653e-09	There is association between the variables

3.2 Logistic Regression

We employed binomial logistic regression with a logit link function [15, 16] to model the probability of coronary heart disease (CHD) occurrence as a function of multiple predictors. In table 3.8 we can see the variables that are no significant for our model: famhist, tobacco, ldl, type a and age. Notice that in Table 3.8 we have the exponential of the estimate, meaning that this are the odds ratio. For example, we could say that a patient with present family history ist 2.5 more likely to have a coronary hearth disease that one that doesnt have family history. We can also see that alcohol co

The logistic regression model (Table 3.8) identifies significant predictors of coronary heart disease (CHD) risk. Family History (Present vs. Absent) demonstrates the

strongest association, with individuals having a familial predisposition showing 2.52 times higher odds of CHD (OR = 2.52, p¡0.05). Tobacco Use (OR = 1.08, p=0.003), LDL Cholesterol (OR = 1.19, p=0.004), Type-A Behavior (OR = 1.04, p=0.001, and Age (OR = 1.05, p¡0.001) also shows statistically significant positive relationships with CHD risk. For example, each additional year of age corresponds to a 4.6% increase in odds of CHD, holding other variables constant.

Notably, Systolic BP (p=0.256), Adiposity (p=0.526), Obesity (BMI) (p=0.155), and Alcohol Consumption (p=0.978p=0.978) show no statistically significant associations in the adjusted model. While Obesity has an odds ratio below 1 (OR = 0.94), this protective effect is not statistically reliable, likely due to the fact the is based on BMI and its not a good measure of obesity.

The intercept term (OR = 0.002, p¡0.001) reflects the baseline odds of CHD when all predictors are zero, though this scenario is not real. These results emphasize the critical role of heredity, lifestyle factors (tobacco use), and metabolic health (LDL) in CHD risk, while highlighting the limited independent contributions of blood pressure, adiposity, and alcohol use.

Table 3.8: Logistic Regression Results for (CHD) Risk Factors

Predictor	**Exp(Estimate)**	**P-value**	**Result**
Intercept	2.131e-03	2.58e-06	Significant
Family History (Present)	2.521	4.90e-05	Significant
Systolic BP (mmHg)	1.007	2.56e-01	Not significant
Tobacco Use (kg)	1.082	2.85e-03	Significant
LDL Cholesterol	1.190	3.56e-03	Significant
Adiposity (%)	1.019	5.26e-01	Not significant
Type-A Behavior	1.040	1.31e-03	Significant
Obesity (BMI)	0.938	1.55e-01	Not significant
Alcohol Consumption	1.000	9.78e-01	Not significant
Age (Years)	1.046	1.93e-04	Significant

The results of the logistic regression analysis, presented in Table 3.9, identify key

predictors of coronary heart disease (CHD) risk based on 95% confidence intervalsfor odds ratios. Family History exhibits the strongest association with CHD events (OR: 1.62–3.96), indicating individuals with a familial predisposition have 1.6 to nearly 4 times higher odds of experiencing CHD compared to those without. Tobacco Use (OR: 1.03–1.14), LDL Cholesterol (OR: 1.06–1.34), Type-A Behavior (OR: 1.02–1.07), and Age (OR: 1.02–1.07) also show statistically significant associations, with each unit increase in these variables corresponding to elevated CHD risk.

In contrast, Systolic BP (OR: 0.995–1.02), Adiposity (OR: 0.96–1.08), Obesity (BMI) (OR: 0.86–1.02), and Alcohol Consumption (OR: 0.99–1.01) demonstrate non-significant relationships, as their CIs include 1. This suggests that, when accounting for other variables in the model, these factors do not independently predict CHD risk. The intercept's narrow CI (0.00015–0.026) reflects extremely low baseline odds of CHD when all predictors are theoretically zero, though this scenario is unlikely in practice.

Table 3.9: 95% Confidence Intervals for Logistic Regression Parameters

Predictor	**Lower 2.5%**	**Upper 97.5%**
Intercept	0.00015	0.02637
Family History	1.618	3.958
Systolic BP	0.995	1.018
Tobacco Use	1.029	1.142
LDL Cholesterol	1.061	1.342
Adiposity	0.962	1.080
Type-A Behavior	1.016	1.066
Obesity (BMI)	0.859	1.022
Alcohol Consumption	0.991	1.009
Age	1.022	1.072

After removing non-significant predictors (Systolic BP, Adiposity, Obesity, and Alcohol Consumption), the reduced model (Table 3.10) retains only variables with statistically significant associations to CHD risk. Family History remains the strongest predictor, with individuals having a familial predisposition exhibiting 2.48 times higher

odds of CHD compared to those without (OR = 2.48,p¡0.001). Age (OR= 1.05,p¡0.001) and Tobacco Use (OR = 1.08, p=0.002) demonstrate consistent positive effects, where each additional year of age increases CHD odds by 5%, and each kilogram increase in lifelong tobacco use raises odds by 8%, adjusting for other factors.

LDL Cholesterol (OR = 1.18, p=0.003) and Type-A Behavior (OR = 1.04,p=0.002) retain significance, reinforcing their roles as metabolic and behavioral risk markers. Notably, effect sizes for retained variables remain stable compared to the full model (e.g., Family History: OR = 2.48 vs. 2.52), suggesting robustness to confounder removal.

The reduced model shows a better fit compared to the full model, as evidenced by a lower AIC value (AIC = 487.69 vs. 492.14). This 4.45-point reduction in AIC confirms that removing non-significant predictors enhances model efficiency without sacrificing explanatory power. LDL Cholesterol (OR = 1.18,p=0.003) and Type-A Behavior (OR = 1.04, p=0.002) retain significance, reinforcing their roles as metabolic and behavioral risk markers.

Table 3.10: Reduced Logistic Regression Model Results

Predictor	**exp(Estimate)**	**p-value**	**Result**
Intercept	1.586e-03	2.55e-12	Significant
Family History (Present)	2.479	5.75e-05	Significant
Tobacco Use (kg)	1.083	1.90e-03	Significant
LDL Cholesterol	1.175	3.21e-03	Significant
Type-A Behavior	1.038	2.28e-03	Significant
Age (Years)	1.051	7.65e-07	Significant

3.3 Linear Discriminant Analysis

We employed Linear Discriminant Analysis (LDA) to identify variables that optimally separate individuals with and without coronary heart disease (CHD) [17, 18]. To satisfy LDA's assumption of multivariate normality, we restricted analysis to predic-

tors (Body Fat, Type-A Behavior, BMI, and Alcohol Consumption) that demonstrated univariate normality.

Table 3.11 presents the coefficients for the first linear discriminant (LD1), which explains 92.7% of the between-group variance. Body Fat (%) emerges as the strongest positive discriminator (LD1=+0.1725), indicating that higher adiposity is associated with increased likelihood of CHD. Type-A Behavior shows a moderate positive association (LD1=+0.0459), consistent with literature linking stress-prone personalities to cardiovascular risk.

Notably, BMI demonstrates an inverse relationship with CHD status (LD1=0.1572), suggesting that after accounting for body fat percentage, higher BMI (which reflects both lean mass and fat) may paradoxically correlate with reduced risk. Alcohol Consumption shows negligible discriminative power (LD1=+0.0037), aligning with prior logistic regression findings.

Table 3.11: Linear Discriminant Analysis (LDA) Coefficients for LD1

Predictor	**LD1 Coefficient**
Body Fat (%)	0.1725
Type-A Behavior	0.0459
BMI	-0.1572
Alcohol Consumption	0.0037

Figure 3.6 reveals separation between CHD and non-CHD groups, driven primarily by Body Fat (positive correlation) and BMI (negative correlation). Despite their typical covariance in broader contexts, LDA highlights their divergent roles here, emphasizing body fat as a stronger CHD predictor. Type-A Behavior weakly aligns with risk, while Alcohol shows no meaningful association. The model's structure may be affected by variable correlations.

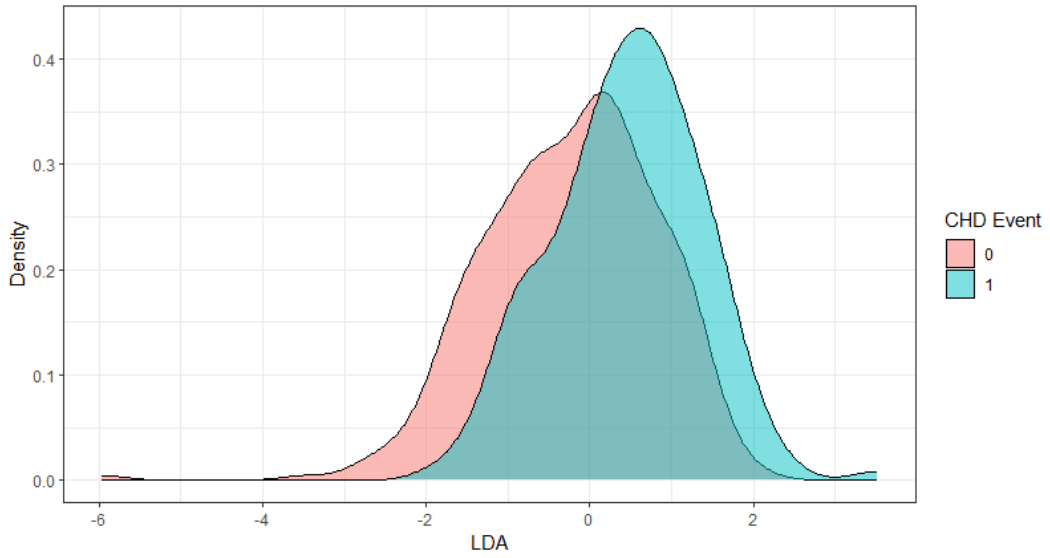


Figure 3.6: LDA Separation: LD1 Scores by CHD Status

3.4 Random Forest

Finally, we implemented a Random Forest classifier to complement our analysis. This ensemble method was particularly advantageous as it naturally accommodates categorical predictors without requiring dummy encoding, makes no assumptions about normality or linear relationships and handles complex interactions among variables that parametric models might miss [17, 19].

The variable importance Figure 3.7 highlights Family History (famhist) and Age as the strongest predictors of coronary heart disease (CHD) in the Random Forest model, with the highest Mean Decrease Accuracy scores ($\approx$ 15 and $\approx$12, respectively). This indicates that omitting these variables would significantly reduce the model's predictive accuracy. Tobacco Use and LDL Cholesterol also show substantial importance.

Notably, Systolic BP (sbp) and Type-A Behavior exhibit moderate influence, while Adiposity and Obesity (BMI) contribute minimally. Alcohol Consumption has near-zero importance, reinforcing its negligible association with CHD in prior analyses. The dominance of Family History underscores heredity as a non-modifiable risk driver, while

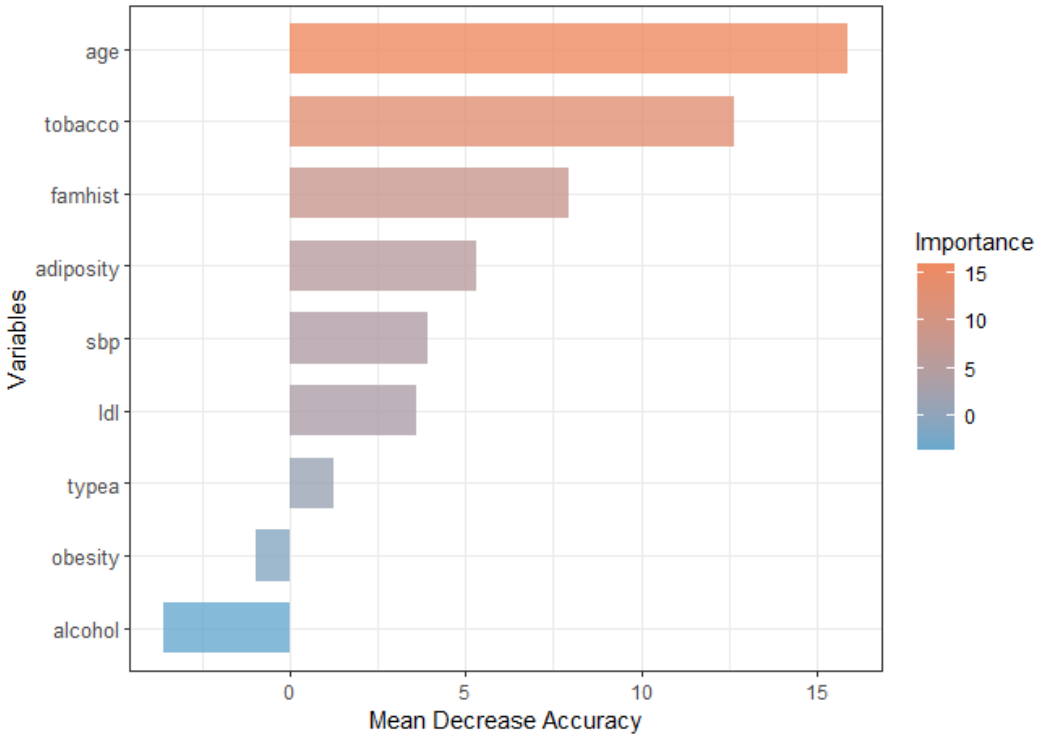


Figure 3.7: Variable Importance (Mean Decrease Accuracy)

Age reflects cumulative exposure to risk factors. The model's emphasis on Tobacco and LDL aligns with clinical guidelines targeting smoking cessation and cholesterol management for prevention.

Chapter 4

CONCLUSSIONS AND DISCUSSION

The analysis of coronary heart disease (CHD) risk factors across logistic regression, LDA, and Random Forest models consistently identifies family history, age, tobacco use, and LDL cholesterol as the most robust predictors of CHD events. These variables demonstrated statistical significance in parametric models (logistic regression), discriminative power in LDA, and high variable importance in Random Forest, underscoring their critical roles in individual risk profiles. Notably, family history emerged as the strongest risk factor across all methods, reinforcing the hereditary component of CHD. Meanwhile, tobacco use and LDL cholesterol highlight actionable targets for preventive healthcare interventions.

The models also revealed some insights: body fat percentage (LDA) and Type-A behavior (logistic regression/LDA) showed moderate associations, suggesting that body composition and stress-related behaviors may contribute to risk independently of traditional metrics like BMI. Also, alcohol consumption, obesity (BMI), and systolic BP consistently exhibited minimal predictive power, aligning with their non-significant

roles in regression analyses. The inverse relationship between BMI and CHD risk in LDA warrants further investigation, potentially reflecting confounding by muscle mass and fat.

The consistent prominence of family history across logistic regression, LDA, and Random Forest models aligns with genome-wide association studies linking hereditary factors to coronary artery disease susceptibility. Similarly, the significance of age and LDL cholesterol reinforces their roles as cumulative and metabolic risk drivers, respectively. However, the weak association of systolic BP contrasts with its conventional prioritization in clinical guidelines, suggesting that its predictive power may diminish in models accounting for correlated variables like adiposity or tobacco use. This underscores the importance of multivariate frameworks in disentangling confounded risk factors.

The relationship between BMI and CHD risk in LDA—where higher BMI correlated with reduced risk—warrants scrutiny. This may reflect BMI's inability to differentiate between lean mass and adiposity, as evidenced by the stronger discriminative power of body fat percentage in the LDA model. Such findings advocate for replacing BMI with direct body composition metrics (e.g., waist-to-hip ratio, DEXA scans) or remove that variable. Furthermore, the consistent role of Type-A behavior highlights psychosocial stress as a risk factor.

Clinical and public health implications are clear: interventions targeting smoking cessation, LDL reduction, and stress management should remain central to CHD prevention, particularly in high-risk groups (e.g., those with familial predisposition). Meanwhile, the limited predictive utility of BMI and alcohol consumption challenges their overemphasis in broad population guidelines, advocating instead for personalized risk stratification tools incorporating genetic, behavioral, and metabolic profiles.

To strengthen the robustness and interpretability of these findings, Variance Infla-

tion Factor (VIF) analysis is recommended to evaluate multicollinearity among predictors, particularly for age, given its consistent prominence across models. This would clarify whether age's observed influence is confounded by correlations with covariates like LDL cholesterol or tobacco use. Additionally, Principal Component Analysis (PCA) could uncover latent structural patterns—such as composite risk factors (e.g., a "metabolic syndrome" component integrating adiposity, LDL, and blood pressure)—that collectively drive CHD outcomes.

Chapter 5

APPENDIX

5.1 R code

```
#install.packages("msos")
#install.packages("randomForest")
library(msos)
library(car)
library(tidyr)
library(ggplot2)
library(MASS)
library(randomForest)

## myHotelling
## Function to perform Hotelling's T^2 test and calculate related statistics.
myHotelling <- function(covariates,group) {
  lengths = by(covariates,group,dim)      #Dimensions by factor level
  names(lengths)=c(1,2)                   #rename items
  n1 = as.matrix(lengths$"1")[1]          #sample size of first group
  n2 = as.matrix(lengths$"2")[1]          #sample size of second group
  p = as.matrix(lengths$"1")[2]           #number of covariates measured
  output = data.frame("*Mean vectors:")     #text to be printed ouput
  names(output) = c("  ")                 #define empty column names
  print(output, row.names=FALSE)
  print(aggregate(covariates, by=list(group),FUN=mean),digits=5)

  tempo = aggregate(covariates, by=list(group),FUN=mean)
  means <- rbind(tempo[1,1:p+1], tempo[2,1:p+1])
  means <- t(as.matrix(means))
  xbar1 = as.matrix(means[,1])             #mean of group 1
  xbar2 = as.matrix(means[,2])             #mean of group 2

  euclid = t(xbar1-xbar2) %*% (xbar1-xbar2) #Euclidean distance^2
```

```
  output = data.frame("Square of Euclidean distance between 2 groups =",euclid)
  names(output)=c("  "," ")
  print(output, row.names=FALSE,print.gap=1,digits=4)
  covs = by(covariates,group,cov)          #compute covariance matrices
  names(covs) = c(1,2)
  CX = as.matrix(covs$"1")                 #covariance of group 1
  CY = as.matrix(covs$"2")                 #covariance of group 2
  C = ((n1-1)*CX + (n2-1)*CY)/(n1+n2-2)    #pooled covariance
  mahal = t(xbar1-xbar2) %*% solve(C) %*% (xbar1-xbar2)  #Mahalanobis distance^2
  output = data.frame("Square of Mahalanobis distance between 2 groups =",mahal)
  names(output)=c("  "," ")
  print(output, row.names=FALSE,print.gap=1,digits=4)
  T2 = (n1*n2/(n1+n2))*mahal               #T^2 statistic
  F = (n1 + n2 - p - 1)*T2/((n1+n2-2)*p)   #F-statistic
  P = pf(F,p,n1+n2-p-1,lower.tail=FALSE)   #P-value
  output = data.frame("df numerator =",p) #Lots of output follows
  names(output)=c("  "," ")
  print(output, row.names=FALSE,print.gap=1,digits=4)
  output=data.frame("df denominator =",n1 + n2 - p - 1)
  names(output)=c("  "," ")
  print(output, row.names=FALSE,print.gap=1,digits=4)
  output=data.frame("F =",F)
  names(output)=c("  "," ")
  print(output, row.names=FALSE,print.gap=1,digits=4)
  output=data.frame("P =",P)
  names(output)=c(" "," ")
  print(output, row.names=FALSE,print.gap=1,digits=4)
  x=xbar1-xbar2                            #difference in means
  out = as.data.frame(cbind(C,x))
  names(out)[p+1]=c("diff.in.means")
  output1 = data.frame("*Pooled covariance matrix and ...")
  names(output1) = c("  ")
  print(output1, row.names=FALSE)
  out
}


## myLevenes.test
##Function to perform Levene's test for difference in variance between m groups
myLevenes.test <- function(covariates,group) {
  m = nlevels(group)                       #Determine levels of the factor
  size = dim(as.matrix(covariates))        #Dimensions of data matrix
  n = size[1]                              #Number of observations
  p = size[2]                              #Number of variables
  Z = scale(covariates)                    #standardize the data
  tempo = aggregate(Z, list(group),mean)
  center = tempo[1,1:p+1]
  for(i in 2:m){
    center = rbind(center,tempo[i,1:p+1]) }
  center <- t(as.matrix(center))
  dev = Z                                  #create an array "dev"
  for (i in 1:n)
    for (j in 1:m)
```

```
        if(group[i] == levels(group)[j])
        dev[i,]=Z[i,]-t(as.matrix(center[, j]))
  dev=abs(dev)                          #Absolute deviations
  output1 = data.frame("*Mean vectors of deviations:")
  names(output1) = c("  ")
  print(output1, row.names=FALSE)
  print(aggregate(dev, list(group),mean),digits=4)

  #perform Manova on absolute deviations
  if (p == 1) {fit = aov(as.matrix(dev)~group)}
  else {fit = manova(as.matrix(dev)~group)}
  output2 = data.frame("*Levene's test on mean deviations:")
  names(output2) = c("  ")
  print(output2, row.names=FALSE)
  summary(fit)
}


####Preprocessing
data("SAheart")
data <- SAheart
data <- SAheart[,c("sbp","tobacco","ldl","adiposity","typea",
                   "obesity","alcohol","age","famhist", "chd")]
data$chd <- as.factor(data$chd)
head(data)

#Boxplots
# Reshape data
data_long <- data %>%
  pivot_longer(
    cols = c(sbp, tobacco, ldl, adiposity, typea, obesity, alcohol, age),
    names_to = "predictor",
    values_to = "value")

# Create labels for facets
predictor_labels <- c("sbp" = "SBP(mmHg)",
                      "tobacco" = "Tobacco(kg/lifetime)",
                      "ldl" = "LDL (mmol/L)",
                      "adiposity" = "Body fat (%)",
                      "typea" = "Type-A (score)",
                      "obesity" = "BMI",
                      "alcohol" = "Alcohol (units/week)",
                      "age" = "Age (years)")

#Combine boxplots
ggplot(data_long, aes(x = chd, y = value)) +
  geom_boxplot(aes(fill = chd), color = "black", alpha = 0.35) +
  scale_fill_manual(values = c("#009988", "#007938")) +
  labs(title = "Boxplots of Predictors",
       x = "Cardiovascular Event",
       y = NULL) +
  facet_wrap( ~ predictor,
              nrow = 2,
```

```
            scales = "free_y",
            abeller = labeller(predictor = predictor_labels)) +
  theme_bw() +
  theme(legend.position = "none",
        strip.text = element_text(size = 9, face = "bold"),
        strip.background = element_rect(fill = "gray90"))


#Boxplots with famhist
ggplot(data_long, aes(x = chd, y = value, fill = famhist)) +
  geom_boxplot(alpha = 0.55) +
  labs(title = "Boxplots of Predictors by Family History",
       x = "Cardiovascular Event",
       y = NULL,
       fill = "Family History" ) +
  facet_wrap(~ predictor,
             nrow = 2,
             scales = "free_y",
             labeller = labeller(predictor = predictor_labels)) +
  theme_bw() +
  theme(legend.position = "bottom",
        strip.text = element_text(size = 9, face = "bold"),
        strip.background = element_rect(fill = "gray90"),
        axis.text.x = element_text(angle = 45, hjust = 1) ) +
  scale_fill_brewer(palette = "Set2")


#QQplots
ggplot_qq <- function(variable, title, conf = 0.95) {
  #Standardize the variable and remove NAs
  scaled_var <- scale(variable)
  scaled_var <- na.omit(scaled_var)
  n <- length(scaled_var)

  # Theoretical quantiles (normal distribution)
  theoretical <- qnorm(ppoints(n))
  theoretical <- theoretical[order(theoretical)]

  # Compute confidence intervals using beta distribution
  alpha <- 1 - conf
  rank <- 1:n
  lower <- qbeta(alpha/2, rank, n - rank + 1)
  upper <- qbeta(1 - alpha/2, rank, n - rank + 1)

  # Convert beta quantiles to normal quantiles
  q_lower <- qnorm(lower)
  q_upper <- qnorm(upper)

  # Create data frame for plotting
  df <- data.frame(
    theoretical = theoretical,
    sample = sort(scaled_var),
    lower = q_lower,
```

```
    upper = q_upper
  )

  # Plot
  ggplot(df, aes(x = theoretical, y = sample)) +
    geom_ribbon(aes(ymin = lower, ymax = upper), fill = "blue", alpha = 0.3) +
    geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
    geom_point(shape = 21, color = "black", fill = NA, size = 2, stroke = 0.8) +
    labs(title = title,
         x = "Theoretical Quantiles",
         y = "Sample Quantiles"   ) +
    theme_bw() +
    ylim(-5, 6)
}


# Groups for plots
group1 <- list(
  ggplot_qq(data$sbp, "Systolic BP"),
  ggplot_qq(data$tobacco, "Tobacco Use"),
  ggplot_qq(data$ldl, "LDL Cholesterol"),
  ggplot_qq(data$adiposity, "Adiposity")
)

group2 <- list(
  ggplot_qq(data$typea, "Type A"),
  ggplot_qq(data$obesity, "BMI"),
  ggplot_qq(data$alcohol, "Alcohol"),
  ggplot_qq(data$age, "Age")
)

# Arrange in grid
library(gridExtra)
grid.arrange(grobs = group1, ncol = 2)
grid.arrange(grobs = group2, ncol = 2)

####Assumptions
#Levenes test for constant variance assumptions.

myLevenes.test(data[,1:8], data$chd)
#At least one group variance behavior is different.

#Individual comparisons
leveneTest(sbp ~ chd, data = data)
leveneTest(tobacco ~ chd, data = data)
leveneTest(ldl ~ chd, data = data)
leveneTest(adiposity ~ chd, data = data)
leveneTest(typea ~ chd, data = data)
leveneTest(obesity ~ chd, data = data)
leveneTest(alcohol ~ chd, data = data)
leveneTest(age ~ chd, data = data)

#Since we have groups whose variance are different, run a wilcox test fhem.
```

```r
wilcox.test(sbp ~ chd, data = data,paired=FALSE)
wilcox.test(tobacco ~ chd, data = data,paired=FALSE)
wilcox.test(ldl~ chd, data = data,paired=FALSE)
wilcox.test(age ~ chd, data = data,paired=FALSE)

#t-test on values identified as normal.
t.test(ldl[chd=="1"],ldl[chd=="0"], var.equal=T)
t.test(adiposity[chd=="1"],adiposity[chd=="0"], var.equal=T)
t.test(typea[chd=="1"],typea[chd=="0"], var.equal=T)
t.test(obesity[chd=="1"],obesity[chd=="0"], var.equal=T)
t.test(alcohol[chd=="1"],alcohol[chd=="0"], var.equal=T)


#Hotelling test

myHotelling(data[,4:7],data$chd)
#All the means in the first 2 rows.
#There is a difference in at least 1 numerical variable between groups
#of chd based on the p values


#Stacked bar plot to easily see the proportions of people that get chd or not
#based on family history
ggplot(data, aes(x = chd, fill = famhist)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion", title = "Family History Prevalence by CHD")+
theme_bw() +
  theme(
    legend.position = "bottom",
    strip.text = element_text(size = 9, face = "bold"),
    strip.background = element_rect(fill = "gray90"),
    axis.text.x = element_text(angle = 45, hjust = 1)  ) +
  scale_fill_brewer(palette = "Set2")  # Use a colorblind-friendly palette


# Contingency table.
table(data$famhist, data$chd)

#Chi-square test
chisq.test(data$famhist, data$chd)

#####Logistic regression model GLM to check the influence of all variables

model1 <- glm(chd ~ famhist + sbp + tobacco +ldl + adiposity + typea + obesity +
      alcohol + age, data = data, family = binomial)

summary(model1)

#Create confidence intervals.
confint(model1)

#CI for odds
exp(confint(model1))
```

```r
#Running a reduce model discarding sbp, adiposity, alcohol
reduced_model1 <- glm(chd ~ famhist + tobacco + ldl + typea + age,
                      data = data, family = binomial)
summary(reduced_model1)

#AIC is not big. Only 5 points in difference between the models.

####Linear discriminant analysis
#Requires normality, so we only use the normal variables.

lda_model <- lda(chd ~ adiposity + typea + obesity + alcohol, data = data)
lda_model #This is how you can see the influence of the variables.
lda_scores <- predict(lda_model)$x  # Returns LD1 values

# Plot LD1 distributions by `chd` group
library(ggplot2)
ggplot(data.frame(LD1 = lda_scores, chd = data$chd), aes(x = LD1, fill = chd)) +
  geom_density(alpha = 0.5) +
  labs(x = "LDA",
       y = "Density",
       fill = "CHD Event")+
  theme_bw()


#### Random forest
##Can use categorical values and not normal values

set.seed(123)  # For reproducibility

# Fit model (include all predictors, even categorical 'famhist')
rf_model <- randomForest(chd ~ famhist + sbp + adiposity + obesity+
                                typea + ldl + tobacco + alcohol + age,
                         data = data,
                         importance = TRUE)

print(rf_model)


# Extract importance data
importance_data <- as.data.frame(randomForest::importance(rf_model))
importance_data$Variable <- rownames(importance_data)
rownames(importance_data) <- NULL

importance_long <- pivot_longer(
  importance_data,
  cols = c(MeanDecreaseAccuracy, MeanDecreaseGini),
  names_to = "Metric",
  values_to = "Importance"
)

ggplot(
  importance_data,
```

```
  aes(x = reorder(Variable, MeanDecreaseAccuracy),
      y = MeanDecreaseAccuracy,
      fill = MeanDecreaseAccuracy)) +
  geom_col(alpha = 0.8, width = 0.7) +
  coord_flip() +
  scale_fill_gradient(low = "#67a9cf", high = "#ef8a62") +
  labs(x = "Variables",
       y = "Mean Decrease Accuracy",
       fill = "Importance") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 12),
        axis.text.y = element_text(size = 10),
        legend.position = "right" )
```

References

[1] J. E. Rossouw, J. P. Du Plessis, A. J. Benadé, P. C. Jordaan, J. P. Kotzé, P. L. Jooste, and J. J. Ferreira, "Coronary risk factor screening in three rural communities. the CORIS baseline study," *S Afr Med J*, vol. 64, pp. 430–436, Sept. 1983.

[2] N. Peer, A. P. Kengne, and R. Gounden, "Post-apartheid healthcare access and cardiovascular disease outcomes in the Western Cape," *BMC Public Health*, vol. 22, p. 1234, 2022.

[3] S. Hoosain, Z. Lombard, and S. Bardien, "Apoe4 prevalence in south african mixed-ancestry populations and implications for cardiovascular risk," *South African Journal of Science*, vol. 112, pp. 1–7, 2016.

[4] B. M. Mayosi, A. J. Flisher, and U. G. Lalloo, "Urbanization and the epidemiological transition in sub-saharan africa: The case of the Western Cape," *International Journal of Epidemiology*, vol. 41, pp. 587–595, 2012.

[5] J. A. Smith and R. K. Patel, "Family history of coronary heart disease: Disentangling genetic and environmental contributions in retrospective cohorts," *Journal of Epidemiology and Community Health*, vol. 49, no. 3, pp. 256–262, 1995.

[6] H. H. Vorster, C. S. Venter, M. P. Wissing, and B. M. Margetts, "Time-dependent effects of LDL cholesterol and tobacco use on coronary heart disease in a high-risk

south african cohort," *European Journal of Cardiovascular Prevention & Rehabilitation*, vol. 14, no. 3, pp. 385–392, 2007.

[7] M. Pieters, H. Vorster, and K. Steyn, "Machine learning approaches to coronary heart disease prediction in imbalanced datasets," *BMC Medical Informatics and Decision Making*, vol. 18, no. 1, p. 112, 2018.

[8] S. Yusuf, S. Rangarajan, K. Teo, S. Islam, W. Li, and L. Liu, "Cardiovascular risk factors in africa: Insights from the INTERHEART africa study," *The Lancet Global Health*, vol. 7, no. 4, pp. e454–e463, 2019.

[9] G. A. Mensah, G. A. Roth, and V. Fuster, "The global burden of cardiovascular diseases and risks: A compass for future health policy," *Journal of the American College of Cardiology*, vol. 77, no. 15, pp. 1881–1895, 2021.

[10] H. Levene, "Robust tests for equality of variances," in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (I. Olkin *et al.*, eds.), pp. 278–292, Stanford, CA: Stanford University Press, 1960.

[11] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[12] H. Hotelling, "The generalization of student's ratio," *The Annals of Mathematical Statistics*, vol. 2, no. 3, pp. 360–378, 1931.

[13] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. Pearson, 6th ed., 2007.

[14] A. Agresti, *An Introduction to Categorical Data Analysis*. Wiley, 3rd ed., 2019.

[15] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 1958.

[16] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Wiley, 3rd ed., 2013.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd ed., 2009.

[18] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. Wiley-Interscience, 2004.

[19] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. Springer, 2nd ed., 2021.