A Project Report on

Predicting 10-Year Risk of Coronary Heart Disease:

A Logistic Regression Approach

SUBMITTED

By

Shanmukha Sree Veda Tippavajhala (300210776)

Anupam Sharma (300208103)

COURSE

STAT 315: Applied Regression Analysis

PROFESSOR

Longlong Huang

(Ph.D. Statistics, University of Calgary**)**

Date of Submission: 12th April 2024

Program of study
Data Analysis Post-baccalaureate certificate

University of the Fraser Valley

Table of Contents Page No:

1. Abstract

Cardiovascular diseases (CVDs) are a major global health concern, causing millions of deaths annually. Early prognosis of CVDs can significantly impact patient care and outcomes.

This project aimed to identify key risk factors for heart disease and predict the Ten-Year Coronary Heart Disease (CHD) risk using logistic regression. Three models were developed, and Model 3 emerged as the most optimal based on statistical criteria.

Model 3 incorporates a comprehensive set of predictors, including demographics, lifestyle factors, medical history, and physiological indicators. The inclusion of interaction terms in Model 3 captures the complex interplay between these factors in influencing CHD risk.

This optimized model has the potential to improve the accuracy of CHD risk prediction, enabling better preventive interventions and healthcare decisions. Further validation and refinement of Model 3 can pave the way for personalized risk assessment strategies, ultimately leading to improved patient outcomes in managing coronary heart disease.

2. Introduction

Coronary heart disease (CHD) remains a global health crisis, claiming millions of lives annually according to the World Health Organization. In developed countries like the United States, it is a leading cause of death. Early detection and risk assessment are crucial for preventing complications and improving patient outcomes. This study delves into the application of logistic regression to predict the 10-year risk of CHD.

Logistic regression offers a powerful statistical tool to analyze the relationship between a binary outcome (CHD in this case) and a set of predictor variables. By identifying the most relevant risk factors, we can create a model that estimates the likelihood of developing CHD within a 10-year timeframe. This information empowers both patients and healthcare professionals to make informed decisions regarding preventative measures and treatment strategies.

This research explores the development of multiple logistic regression models to predict 10-year CHD risk. We will compare the models based on various statistical criteria and identify the one that provides the most accurate and insightful risk assessment. The chosen model will incorporate a comprehensive set of potential risk factors encompassing demographics, lifestyle choices, medical history, and physiological indicators. Furthermore, we will explore the role of interaction terms within the model to reveal potential synergistic effects between different risk factors on CHD development.

By leveraging the power of logistic regression, this study aims to enhance the accuracy of 10-year CHD risk prediction, ultimately contributing to more effective preventative strategies and improved patient outcomes in managing coronary heart disease.

3. About Data Source

This project utilizes a publicly available dataset from the Framingham Heart Study, a long-term research project investigating cardiovascular health in residents of Framingham, Massachusetts. The dataset offers a wealth of information on over 3,000 individuals and includes 15 attributes, each a potential risk factor for coronary heart disease (CHD).

Here's a breakdown of the data categories:

Demographic Factors:

sex: Categorical variable indicating male ("M") or female ("F").

age: Continuous variable representing the patient's age in years. (Although recorded ages might be whole numbers, the underlying concept of age is continuous).

education: Categorical variable reflecting socioeconomic status. This factor can influence lifestyle choices and access to healthcare, both of which impact long-term cardiovascular health.

Behavioral Factors:

is_smoking: Categorical variable indicating whether the patient is a current smoker ("YES" or "NO").

CigsPerDay: Continuous variable representing the average number of cigarettes smoked per day by the patient.

Medical History:

BPMeds: Categorical variable indicating whether the patient is currently taking medication for blood pressure control ("YES" or "NO").

prevalentStroke: Categorical variable indicating whether the patient has a history of stroke ("YES" or "NO").

prevalentHyp: Categorical variable indicating whether the patient has a history of hypertension ("YES" or "NO").

diabetes: Categorical variable indicating whether the patient has diabetes ("YES" or "NO").

Physiological Indicators (Current Measurements):

totChol: Continuous variable representing the patient's total cholesterol level.

sysBP: Continuous variable representing the patient's systolic blood pressure (measured in millimeters of mercury (mm Hg)). This reflects the maximum force exerted on arterial walls during a heartbeat. A healthy systolic BP is generally less than or equal to 120 mm Hg.

diaBP: Continuous variable representing the patient's diastolic blood pressure (measured in mm Hg). This reflects pressure in the arteries when the heart is resting between beats. A healthy diastolic BP is generally less than or equal to 80 mm Hg.

BMI: Continuous variable representing the patient's Body Mass Index.

heartRate: Continuous variable representing the patient's heart rate. (While technically discrete, heart rate is often considered continuous in medical research due to the large number of possible values).

glucose: Continuous variable representing the patient's blood glucose level.

Response Variable:

TenYearCHD: This is the binary target variable we aim to predict. It indicates whether the patient has a 10-year risk of developing coronary heart disease (CHD). "1" represents a positive risk ("Yes"), and "0" represents no risk ("No").

The richness of this dataset allows us to explore the relationships between various risk factors and their combined influence on the likelihood of developing CHD within a 10-year timeframe. By employing logistic regression analysis, we can leverage this data to build a model that accurately predicts CHD risk, paving the way for more effective preventative strategies and improved patient outcomes.

4. Logistic Regression Model

The 10% significance level is considered for all the models.

4.1.Model 1

$$\mathbf{\log\left(\frac{\widehat{\theta(x)}}{1-\widehat{\theta(x)}}\right) = -9.274979 + 0.066171 \ast age + 0.488752 \ast I(sex = Male) + 0.022857 \ast cigsPerDay + 0.946820 \ast I(prevalentStorke = 1) + 0.003188 \ast totChol + 0.015838 \ast sysBP + 0.008698 \ast glucose}$$

- $\log\left(\frac{\widehat{\theta(x)}}{1-\widehat{\theta(x)}}\right)$, this term represents the logit function, which transforms the predicted probability (θ(x)) of a patient developing CHD within 10 years into a log-odds ratio. The log-odds ratio allows us to interpret the effect of each predictor variable on the odds of CHD rather than just the raw probability.
- Intercept: The odds of TenYearCHD occurring for the patients is 9.374×10^{-5} when all other predictor variables are equal to zero (which is unlikely in reality).
- The odds of the TenYearCHD occurring will be multiplied by 1.0684 for every 1-year increase in age, with all other variables fixed.
- The odds of TenYearCHD occurring for Male patients are 1.6302 times the odds of TenYearCHD occurring for Female patients with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0231 for every additional 1 CigsPerDay, with all other variables fixed.
- The odds of TenYearCHD occurring for patients with prevalentStroke are 2.5775 times the odds of TenYearCHD occurring for patients without prevalentStroke with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0032 for every 1 mg/dL increase in totChol, with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0159 for every 1 mmHg increase in sysBP, with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0087 for every 1mg/dL increase in glucose, with all other variables fixed.

age	sex	CigsPerDay	prevalentStroke	totChol	sysBP	glucose
1.204754	1.210579	1.257597	1.007914	1.059288	1.15901	1.013249

Table: variance inflation factor (VIF)

- These variance inflation factors do not exceed 5, therefore there is no multi-collinearity.

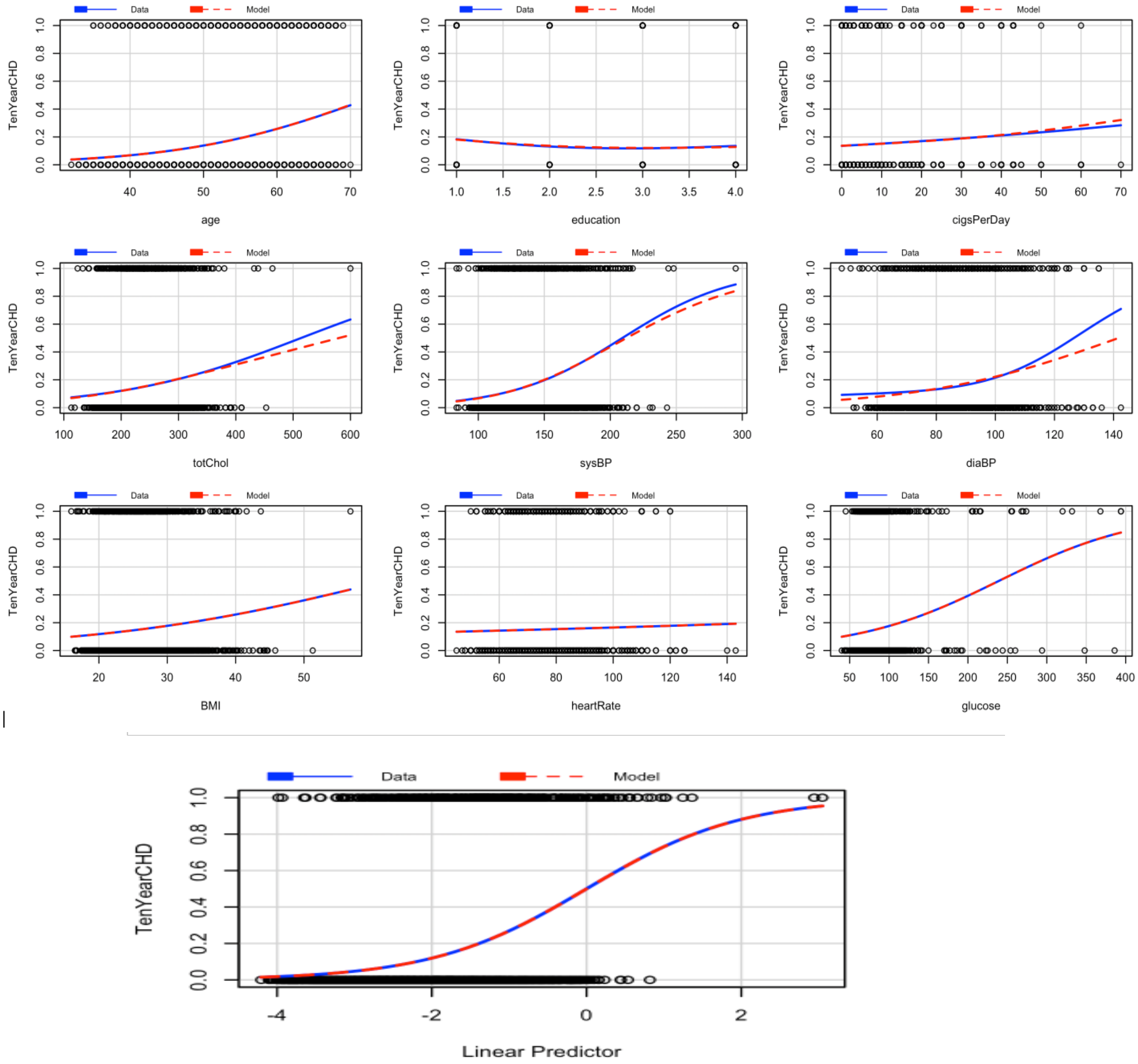


Figure1: Marginal Model Plots for Model 1

Each of the nonparametric estimates (marked as solid curves) matches (are very close to) the fitted values (marked as dashed curves). Model 1 is a valid model for the data.

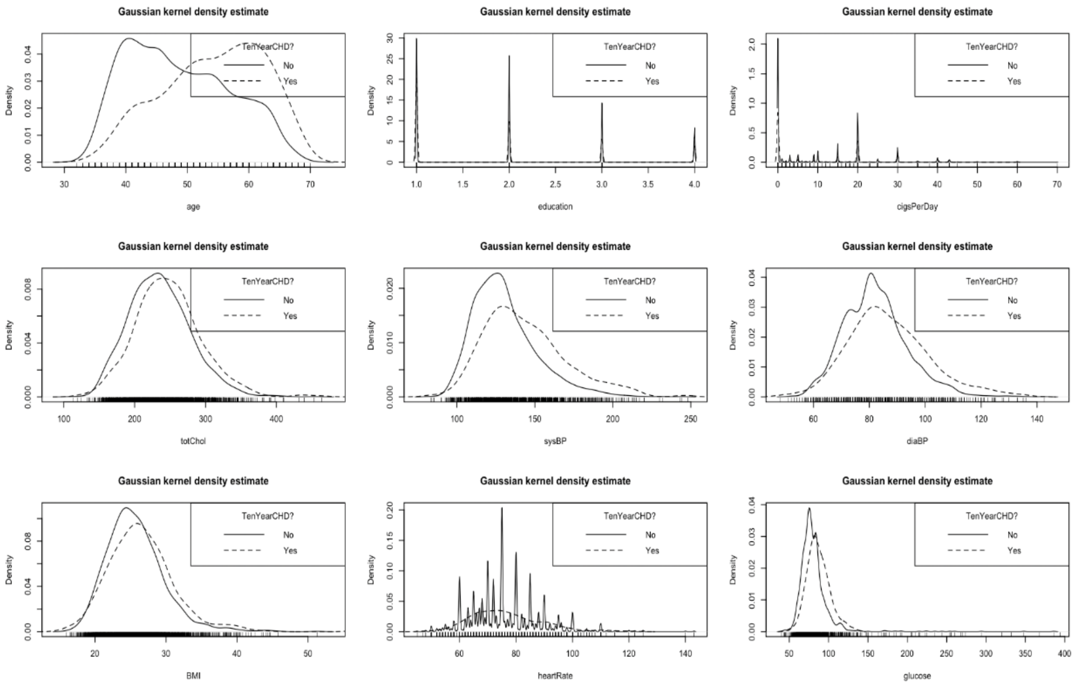


Figure 2: Gaussian kernel density curves for Model 1

The densities of age, diaBP, and glucose are skewed, the log of odds can depend on age, diaBP, glucose and log(age), log(diaBP), log(glucose).

4.2. Model 2

$$\log\left(\frac{\widehat{\theta(x)}}{1-\widehat{\theta(x)}}\right) = 9.004615 + 3.330674 * \log(age) + 0.517066 * I(sex = Male) + 0.022305 * cigsPerDay + 0.972265 * I(prevalentStorke = 1) + 0.003467 * totchol + 0.018031 * sysBP + 0.085000 * diaBP - 8.017965 * \log(diaBP) + 0.008373 * glucose$$

- Intercept: The odds of TenYearCHD occurring for the patients is 8140.5661 when all other predictor variables are equal to zero (which is unlikely in reality).
- The odds of the TenYearCHD occurring will be multiplied by 27.9572 for every 1-year increase in log(age), with all other variables fixed.
- The odds of TenYearCHD occurring for Male patients are 1.6771 times the odds of TenYearCHD occurring for Female patients with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0225 for every additional 1 CigsPerDay, with all other variables fixed.

- The odds of TenYearCHD occurring for patients with prevalentStroke are 2.6439 times the odds of TenYearCHD occurring for patients without prevalentStroke with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0035 for every 1 mg/dL increase in totChol, with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0182 for every 1 mmHg increase in sysBP, with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0887 for every 1 mmHg increase in diaBP, with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 0.00032 for every 1 mmHg increase in diaBP, with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0084 for every 1mg/dL increase in glucose, with all other variables fixed.

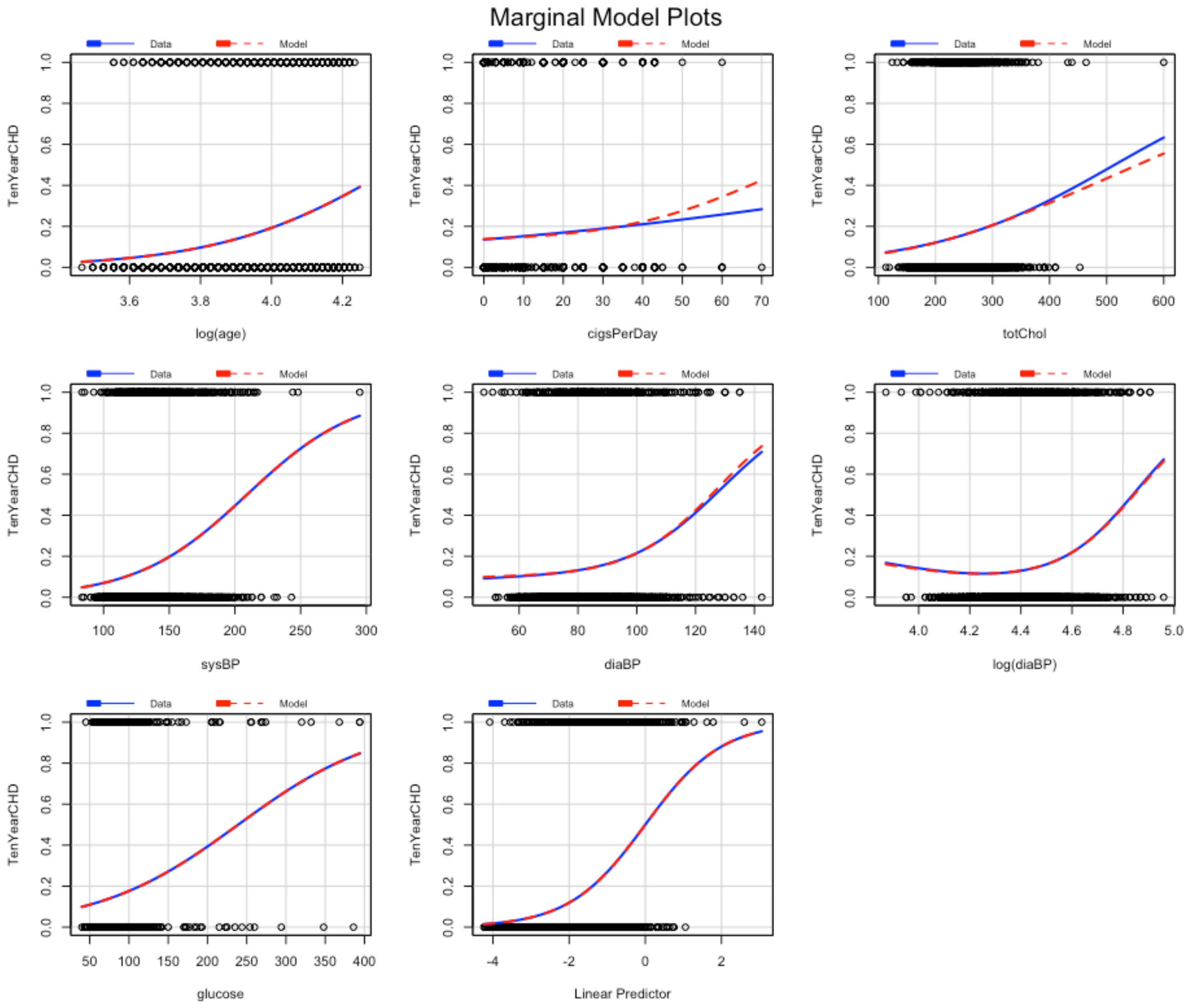


Figure 3: Marginal Model plots for Model 2

Each of the nonparametric estimates (marked as solid curves) matches (are very close to) the fitted values (marked as dashed curves). Model 2 is a valid model for the data.

4.3. Model 3

Model 3 is the logistic regression model which contains interaction terms.

$$\mathbf{log}\left(\frac{\widehat{\theta(x)}}{1-\widehat{\theta(x)}}\right) = 29.85 + 3.296 * \log(\mathrm{age}) + 0.5022 * I(\mathrm{sex} = \mathrm{Male}) + 0.02298 * \mathrm{cigsPerDay} + 4.012 * I(\mathrm{BPMeds} = 1) - 8.374 * I(\mathrm{prevalentStroke} = 1) + 1.123 * I(\mathrm{prevalentHyp} = 1) + 5.228 * I(\mathrm{diabetes} = 1) + 0.00351 * \mathrm{totChol} + 0.01537 * \mathrm{sysBP} + 0.1738 * \mathrm{diaBP} - 14.14 * \log(\mathrm{diaBP}) + 0.01487 * \mathrm{BMI} - 0.006713 * \mathrm{heartRate} + 0.0005497 * \mathrm{glucose} - 0.1351 * \mathrm{BMI} * \mathrm{BPMeds} + 0.01843 * \mathrm{glucose} * \mathrm{prevalentHYP} - 0.01537 * \mathrm{sysBP} * \mathrm{diabetes} + 0.1285 * \mathrm{heartRate} * \mathrm{prevalentStroke} - 0.02919 * \mathrm{diaBP} * \mathrm{prevalentHYP}$$

Even though variables like “BPMeds”, “diabetes”, “prevalentHyp”, “prevalentStroke”, “BMI”, “heartRate”, and “glucose” does not show any significance individually, they are added to the model as they show some significance with interaction with other variables.

- The odds of the TenYearCHD occurring will be multiplied by 27.0044 for every 1-year increase in log(age), with all other variables fixed.
- The odds of TenYearCHD occurring for Male patients are 1.6524 times the odds of TenYearCHD occurring for Female patients with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0232 for every additional 1 CigsPerDay, with all other variables fixed.
- The odds of TenYearCHD occurring for patients using BPMeds are 55.2572 times the odds of TenYearCHD occurring for patients without using BPMeds with all other variables fixed.
- The odds of TenYearCHD occurring for patients with prevalentStroke are 0.00023 times the odds of TenYearCHD occurring for patients without prevalentStroke with all other variables fixed.
- The odds of TenYearCHD occurring for patients with prevalentHyp are 3.074 times the odds of TenYearCHD occurring for patients without prevalentHyp with all other variables fixed.
- The odds of TenYearCHD occurring for patients with diabetes are 186.4195 times the odds of TenYearCHD occurring for patients without diabetes with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0035 for every 1 mg/dL increase in totChol, with all other variables fixed.

- The odds of the TenYearCHD occurring will be multiplied by 1.0154 for every 1 mmHg increase in sysBP, with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.1898 for every 1 mmHg increase in diaBP, with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0149 for every 1kg/m^2 increase in BMI, with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 0.9933 for every 1bpm increase in heartRate, with all other variables fixed.
- The odds of the TenYearCHD occurring will be multiplied by 1.0005 for every 1mg/dL increase in glucose, with all other variables fixed.
- The odds for the interaction term between BPMeds and BMI status indicate how the effect of BMI on the likelihood of TenYearCHD changes depending on whether an individual takes BPMeds or not while keeping all other factors constant.
- The odds for the interaction term between glucose and prevalentHyp status indicates how the effect of glucose on the likelihood of TenYearCHD changes depending on whether an individual has prevalentHyp or not while keeping all other factors constant.
- The odds for the interaction term between sysBP and diabetes status indicate how the effect of sysBP on the likelihood of TenYearCHD changes depending on whether an individual has diabetes or not while keeping all other factors constant.
- The odds for the interaction term between heartRate and prevalentStroke status indicate how the effect of heartRate on the likelihood of TenYearCHD changes depending on whether an individual has prevalentStroke or not while keeping all other factors constant.
- The odds for the interaction term between diaBP and prevalentHyp status indicates how the effect of diaBP on the likelihood of TenYearCHD changes depending on whether an individual has prevalentHyp or not while keeping all other factors constant.

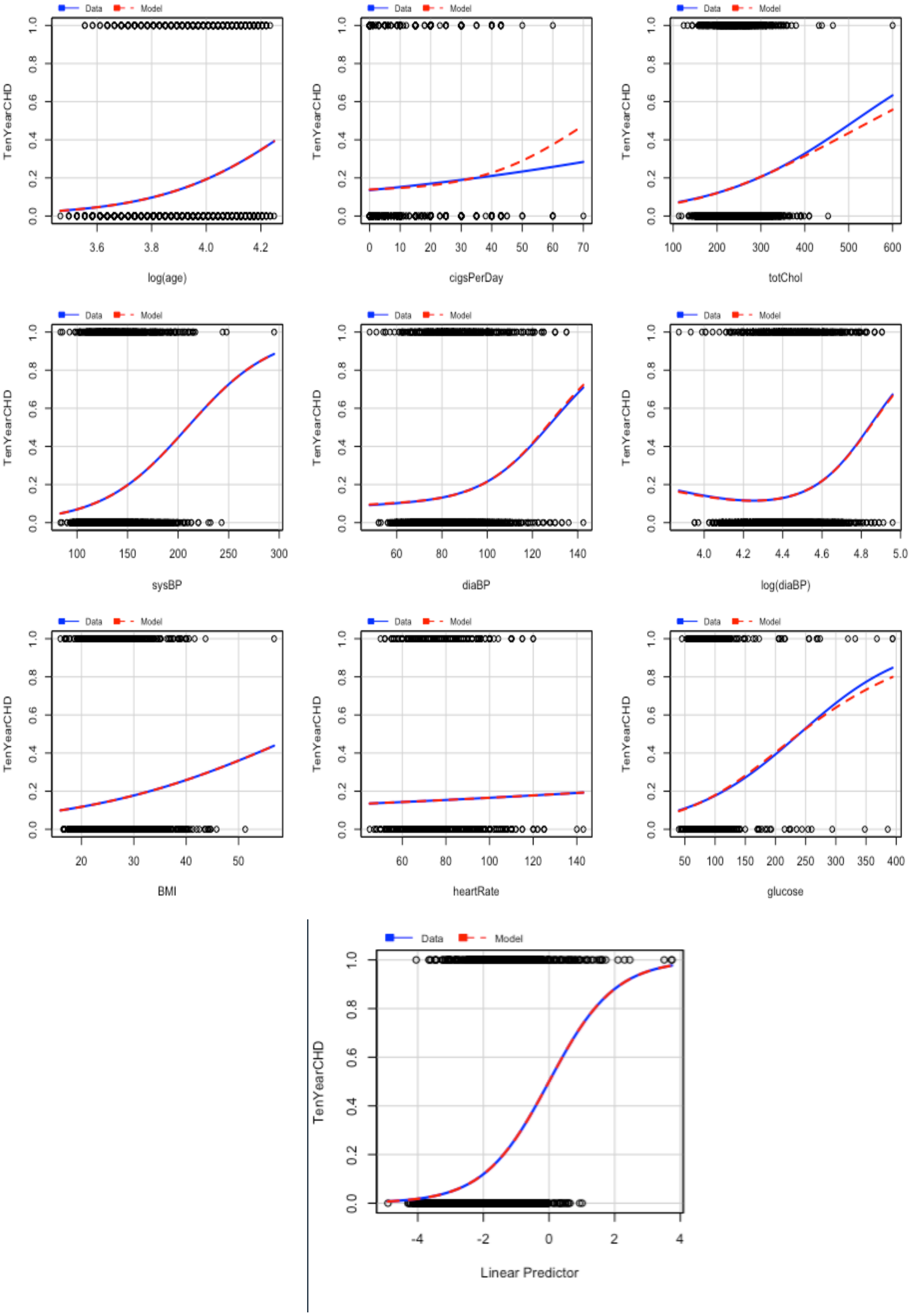


Figure 3: Marginal Model Plot for Model 3

Even though there is some deviation between the two fits for CigsPerDay over all these marginal model plots show reasonable agreement across the two sets of fits indicating that model 2 is valid.

5. Model Comparison

5.1. AIC

model	df	AIC
Model1	8	2218.004
Model2	10	2214.338
Model3	20	2197.494

The Model 3 has the lowest AIC value as compared to Model 1 and Model 2.

5.2. Hosmer-Lemeshow Test

H_0: the logistic model1 provides an adequate fit to the data.
H_a: the logistic model1 does not adequately fit the data.

Test statistics = χ^2= 6.2592, df = 8, p-value = 0.6182
Decision: 0.6182 > 0.10, we fail to reject H_0 at a 10 % significance level

Conclusion: We have insufficient evidence to state that model 1 does not adequately fit the data at a 10% significance level.

H_0: the logistic model 2 provides an adequate fit to the data.
H_a: the logistic model 2 does not adequately fit the data.

Test statistics = χ^2= 9.7497, df = 8, p-value = 0.283
Decision: 0.283> 0.10, we fail to reject H_0 at a 10 % significance level

Conclusion: We have insufficient evidence to state that model 2 does not adequately fit the data at a 10% significance level.

H_0: the logistic model 3 provides an adequate fit to the data.
H_a: the logistic model 3 does not adequately fit the data.

Test statistics = χ^2= 3.3633, df = 8, p-value = 0.9095
Decision: 0.9095> 0.10, we fail to reject H_0 at a 10 % significance level

Conclusion: We do not have sufficient evidence to state that model 3 does not adequately fit the data at a 10% significance level.

5.3. The difference in deviance between models:

H_0: $log\left(\frac{\widehat{\theta(x)}}{1-\widehat{\theta(x)}}\right) = \beta_0 + \beta_1 * log(age) + \beta_2 * I(sex = Male) + \beta_3 * cigsPerDay + \beta_4 * I(prevalentStroke = 1) + \beta_5 * totChol + \beta_6 * sysBP + \beta_7 * diaBP + \beta_8 * log(diaBP) + \beta_9 * glucose$ → model 2

H_a: $\log\left(\frac{\widehat{\theta(x)}}{1-\widehat{\theta(x)}}\right) = \beta_0 + \beta_1 * \log(age) + \beta_2 * I(sex = Male) + \beta_3 * cigsPerDay + \beta_4 * I(BPMeds = 1) + \beta_5 * I(prevalentStroke = 1) + \beta_6 * I(prevalentHyp = 1) + \beta_7 * I(diabetes = 1) + \beta_8 * totChol + \beta_9 * sysBP + \beta_{10} * diaBP + \beta_{11} * \log(diaBP) + \beta_{12} * BMI + \beta_{13} * heartRate + \beta_{14} * glucose + \beta_{15} * BMI * BPMeds + \beta_{16} * glucose * prevalentHYP + \beta_{17} * sysBP * diabetes + \beta_{18} * heartrate * prevalentStroked + \beta_{19} * diaBP * prevalentHyp$ → model 3

$G^2_{H_0} - G^2_{H_a}$ = 33.663 and p-value = 0.0002106

Decision: 0.0002106 < α = 0.1, hence we reject H_0 at a 10% significance level

Conclusion: There is sufficient evidence to conclude that model 3 provides a reasonable fit to the data at a 10% significance level.

As Model 3 is suggested as the best model by all 3 model comparisons, we consider Model 3 as the best-fitted logistic regression model for the data.

6. Wald's Test at 10% significance level

The Wald's test is the test of significance for individual regression coefficients in logistic regression.

$$H_0: \beta_i = 0 \text{ vs. } H_a: \beta_i \neq 0 \qquad \text{where } 1 \leq i \leq 19$$

For maximum likelihood estimates, the Wald's test statistic is:

$$Z = \frac{\widehat{\beta_\iota}}{se(\widehat{\beta_\iota})}$$

Note: If p-value < 0.1, We reject H_0, else we fail to reject H_0

HYPOTHESIS	Z_{calc}	P-VALUE	DECISION	CONCLUSION
$H_0: \beta_{\log{(age)}} = 0$ $H_a: \beta_{\log{(age)}} \neq 0$	8.557	$2 * 10^{-16}$	Reject H_0	sufficient evidence to state age^2 provides significant effect on TenYearCHD.
$H_0: \beta_{sex} = 0$ $H_a: \beta_{sex} \neq 0$	4.042	$5.29 * 10^{-5}$	Reject H_0	sufficient evidence to state sex provides significant effect on TenYearCHD.
$H_0: \beta_{cigsPerDay} = 0$ $H_a: \beta_{cigsPerDay} \neq 0$	4.754	$2 * 10^{-6}$	Reject H_0	sufficient evidence to state $cigsPerDay$ provides significant effect on TenYearCHD.
$H_0: \beta_{BPMeds} = 0$ $H_a: \beta_{BPMeds} \neq 0$	2.626	0.008651	Reject H_0	sufficient evidence to state $BPMeds$ provides significant effect on TenYearCHD.
$H_0: \beta_{prevalentStroke} = 0$ $H_a: \beta_{prevalentStroke} \neq 0$	-1.579	0.114345	Fail to reject H_0	insufficient evidence to state *prevalentStroke* provides significant effect on TenYearCHD
$H_0: \beta_{prevalentHyp} = 0$ $H_a: \beta_{prevalentHyp} \neq 0$	0.833	0.404624	Fail to reject H_0	insufficient evidence to state *prevalentHyp* provides a significant effect on TenYearCHD.
$H_0: \beta_{diabetes} = 0$ $H_a: \beta_{diabetes} \neq 0$	2.877	0.004020	Reject H_0	sufficient evidence to state $diabetes$ provides a significant effect on TenYearCHD.
$H_0: \beta_{totCHol} = 0$ $H_a: \beta_{totCHol} \neq 0$	2.787	0.005312	Reject H_0	sufficient evidence to state that *totChol* provides a significant effect on TenYearCHD.
$H_0: \beta_{sysBP} = 0$ $H_a: \beta_{sysBP} \neq 0$	3.443	0.000574	Reject H_0	sufficient evidence to state *sysBP* provides significant effect on TenYearCHD

$H_0: \beta_{diaBP} = 0$ $H_a: \beta_{diaBP} \neq 0$	2.885	0.003919	Reject H_0	sufficient evidence to state *diaBP* provides significant effect on TenYearCHD.
$H_0: \beta_{log(diaBP)} = 0$ $H_a: \beta_{log(diaBP)} \neq 0$	-3.114	0.001848	Reject H_0	sufficient evidence to state *log(diaBP)* provides a significant effect on TenYearCHD.
$H_0: \beta_{BMI} = 0$ $H_a: \beta_{BMI} \neq 0$	1.010	0.312595	Fail to reject H_0	insufficient evidence to state *BMI* provides significant effect on TenYearCHD.
$H_0: \beta_{\mathrm{Heart}Rate} = 0$ $H_a: \beta_{\mathrm{Heart}Rate} \neq 0$	-1.397	0.162489	Fail to reject H_0	insufficient evidence to state *heartRate* provides a significant effect on TenYearCHD.
$H_0: \beta_{glucose} = 0$ $H_a: \beta_{glucose} \neq 0$	0.165	0.869221	Fail to reject H_0	insufficient evidence to state *glucose* provides significant effect on TenYearCHD.
$H_0: \beta_{BMI:BPMeds} = 0$ $H_a: \beta_{BMI:BPMeds} \neq 0$	-2.453	0.014156	Reject H_0	sufficient evidence to state interaction between *BMI: BPMeds* provides significant effect on TenYearCHD.
$H_0: \beta_{glucose:prevalentHyp} = 0$ $H_a: \beta_{glucose:prevalentHyp} \neq 0$	3.646	0.000266	Reject H_0	sufficient evidence to state interaction between *glucose: prevalentHyp* provides significant effect on TenYearCHD.
$H_0: \beta_{sysBP:diabetes} = 0$ $H_a: \beta_{sysBP:diabetes} \neq 0$	-2.944	0.003239	Reject H_0	sufficient evidence to state interaction between *sysBP: diabetes* provides significant effect on TenYearCHD.
$H_0: \beta_{heartRate:prevalentStroke} = 0$ $H_a: \beta_{heartRate:prevalentStroke} \neq 0$	1.815	0.069569	Reject H_0	sufficient evidence to state interaction between *heartRate: prevalentStroke* provides

				significant effect on TenYearCHD.
$H_0: \beta_{diaBP:prevalentHyp} = 0$ $H_a: \beta_{diaBP:prevalentHyp} \neq 0$	-1.943	0.051970	Reject H_0	sufficient evidence to state interaction between *diaBP: prevalentHyp* provides a significant effect on TenYearCHD.

Table 2: Wald's Test

7. Confidence Interval for Odds Ratio

Confidence intervals based on the Wald's statistic are of the form:

$$\boldsymbol{\widehat{\beta_\iota} \pm Z_{1-\frac{\alpha}{2}}\, se(\widehat{\beta_\iota}),} \qquad \text{where } 1 \leq i \leq 19$$

Predictor Variable	**90% CI for odds ratio**
Intercept	$(106.35,\ 7.922 * 10^{23})$
log(age)	(14.32678, 50.86678)
sex	(1.347005, 2.027108)
CigsPerDay	(1.015144, 1.031420)
BPMeds	(4.476131, 682.803326)
prevalentStroke	$(3.755182\times10^{-8}$, 1.418298)
prevalentHyp	(0.3349994, 28.2188557)
diabetes	(9.378821, 3703.327019)
totChol	(1.001440, 1.005597)
sysBP	(1.008058, 1.022965)
diaBP	(1.077550, 1.313745)
log(diaBP)	$(4.150179\times10^{-10}$, $1.270976\times10^{-3})$
BMI	(0.9906933, 1.0398593)
heartRate	(0.9854881, 1.0011931)

glucose	(0.99507, 1.00606)
BMI:BPMeds	(0.7980071, 0.9564652)
glucose: prevalentHyp	(1.010167, 1.027101)
sysBP: diabetes	(0.9442255, 0.9838831)
heartRate: prevalentStroke	(1.012096, 1.277491)
diaBP: prevalentHyp	(0.9475321, 0.9955262)

- As the 1 is not included in the Confidence Interval of Odds for log(age), sex, CigsPerDay, BPMeds, diabetes, totChol, sysBP, diaBP, log(diaBP), BPMeds, prevalentHyp: glucose, diabetes: sysBP, prevalentStroke: heartRate, prevalentHyp: diaBP we can conclude that the odds of TenYearCHD depend on these terms.

- But the terms BMI, prevalentHyp, prevalentStroke, glucose, and heartRate which have 1 included in the Confidence Interval of Odds also included in the model because they show some significance with interaction with other terms.

8. Conclusion

- In conclusion, the logistic regression analysis reveals valuable insights into the factors influencing the likelihood of TenYearCHD occurrence.
- Model 3, incorporating interaction terms, emerges as the most suitable, as indicated by various statistical tests and comparisons.
- While variables like age, sex, CigsPerDay, BPMeds, diabetes, totChol, sysBP, diaBP, log(diaBP), and their interactions show significant effects, others exhibit varying degrees of significance.
- Notably, the odds of TenYearCHD are influenced by factors such as age, sex, smoking habits, medication use, and underlying health conditions.
- However, further investigation is warranted for variables with confidence intervals containing 1.
- Overall, the analysis underscores the complex interplay of multiple factors in predicting TenYearCHD risk, emphasizing the importance of comprehensive models for accurate risk assessment.

9. Reference

[1] Anshori, M., & Haris, M. S. (2022). Predicting heart disease using logistic regression. *Knowledge Engineering and Data Science*, *5*(2), 188. https://doi.org/10.17977/um018v5i22022p188-196

[2] Azhar, M., & Gladence, Dr. L. (2022). A machine learning approach for predicting disease in heart using logistic regression. *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*. https://doi.org/10.1109/ic3i56241.2022.10073332

[3] Bansodesandeep. (2022, November 28). *Cardiovascular risk prediction*. Kaggle. https://www.kaggle.com/code/bansodesandeep/cardiovascular-risk-prediction

[4] Blackboard learn. (n.d.). https://myclass.ufv.ca/ultra/courses/_66406_1/cl/outline

[5] camir93. (2021, February 11). *Cardiovascular study dataset - analysis*. Kaggle. https://www.kaggle.com/code/camir93/cardiovascular-study-dataset-analysis

[6] Ciu, T., & Oetama, R. S. (2020). Logistic regression prediction model for cardiovascular disease. *IJNMT (International Journal of New Media Technology)*, *7*(1), 33–38. https://doi.org/10.31937/ijnmt.v7i1.1340

[7] G, A., Ganesh, B., Ganesh, A., Srinivas, C., Dhanraj, & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings*, *3*(1), 127–130. https://doi.org/10.1016/j.gltp.2022.04.008

[8] Ganteng, C. (2020, September 22). *Cardiovascular study dataset*. Kaggle. https://www.kaggle.com/datasets/christofel04/cardiovascular-study-dataset-predict-heart-disea

[9] Li, R., Yang, S., & Xie, W. (2021). Cardiovascular disease prediction model based on logistic regression and euclidean distance. *2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*. https://doi.org/10.1109/aemcse51986.2021.00147

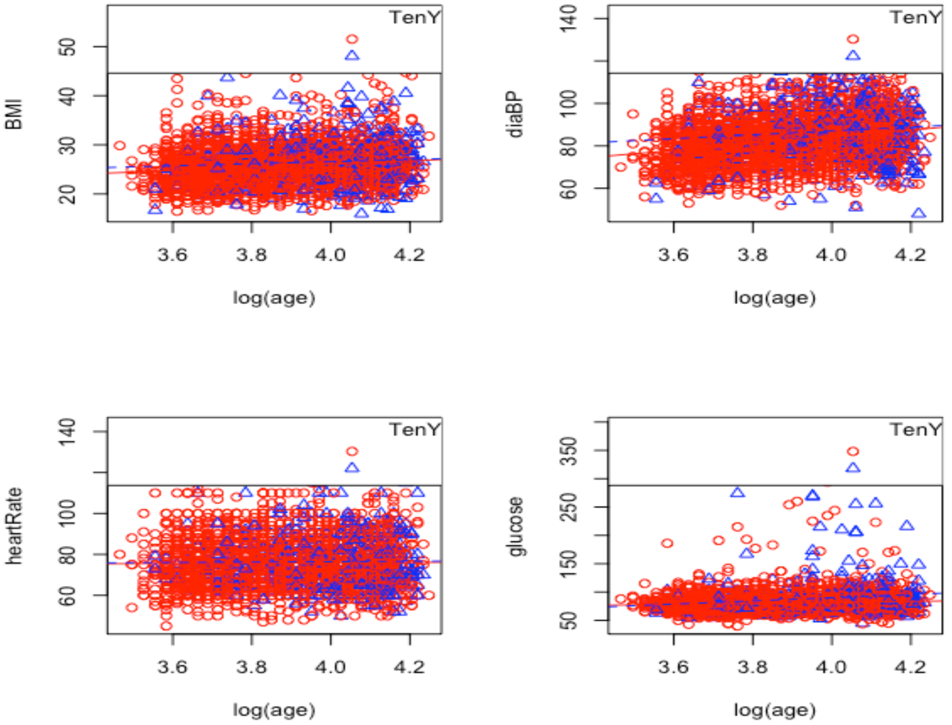
[10] Rani, K. S., Manoj, M. S., & Mani, G. S. (2018). A heart disease prediction model using logistic regression. *International Journal of Trend in Scientific Research and Development*, *Volume-2*(Issue-3), 1463–1466. https://doi.org/10.31142/ijtsrd11401

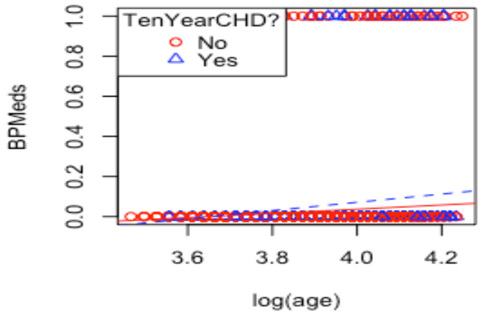
[11] Salau, A. O., Assegie, T. A., Markus, E. D., Eneh, J. N., & Ozue, T. I. (2024). Prediction of the risk of developing heart disease using logistic regression. *International Journal of Electrical and Computer Engineering (IJECE)*, *14*(2), 1809. https://doi.org/10.11591/ijece.v14i2.pp1809-1815

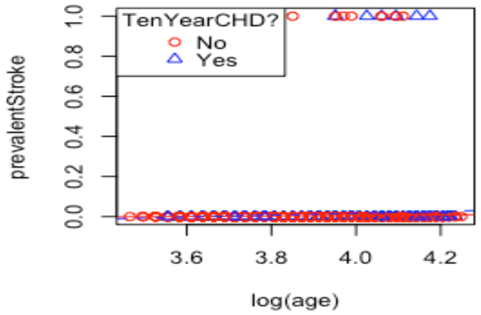
[12] Zhang, Y., Diao, L., & Ma, L. (2021). Logistic regression models in predicting heart disease. *Journal of Physics: Conference Series*, *1769*(1), 012024. https://doi.org/10.1088/1742-6596/1769/1/012024

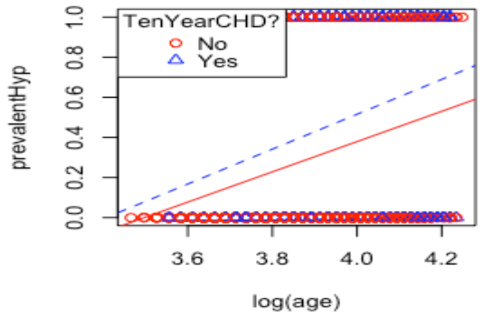
I. Appendix:

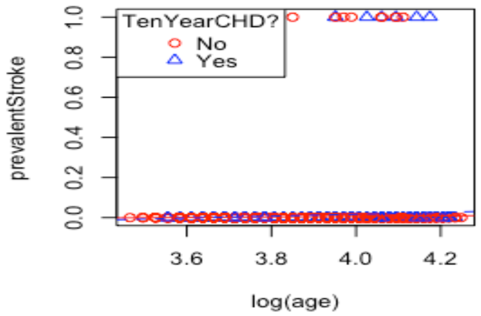
Plots of predictor interaction terms with different slopes for each value of TenYearCHD.

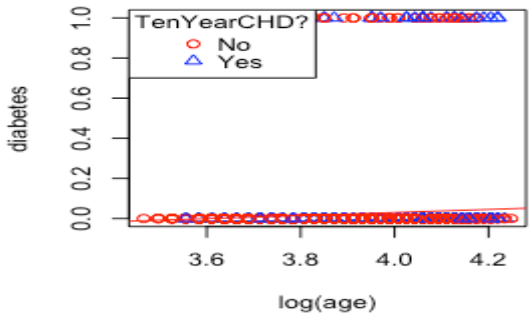


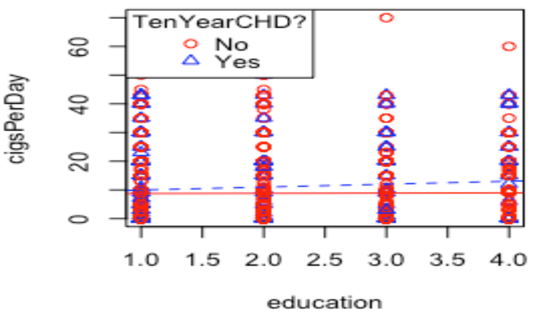

TenYearCHD?
No
Yes
BPMeds
0.0
0.2
0.4
0.6
0.8
1.0
3.6
3.8
4.0
4.2
log(age)

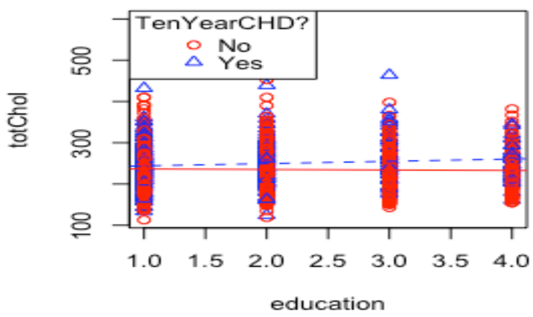

TenYearCHD?
No
Yes
prevalentStroke
0.0
0.2
0.4
0.6
0.8
1.0
3.6
3.8
4.0
4.2
log(age)

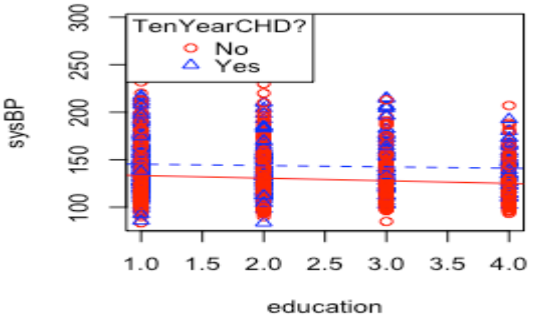

TenYearCHD?
No
Yes
prevalentHyp
0.0
0.2
0.4
0.6
0.8
1.0
3.6
3.8
4.0
4.2
log(age)

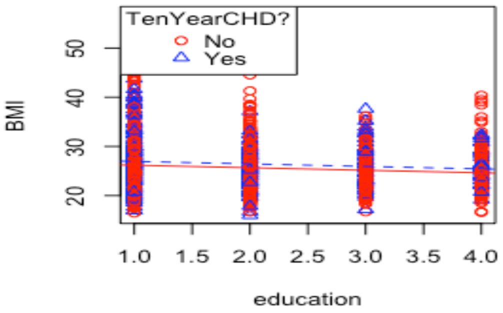

TenYearCHD?
No
Yes
prevalentStroke
0.0
0.2
0.4
0.6
0.8
1.0
3.6
3.8
4.0
4.2
log(age)

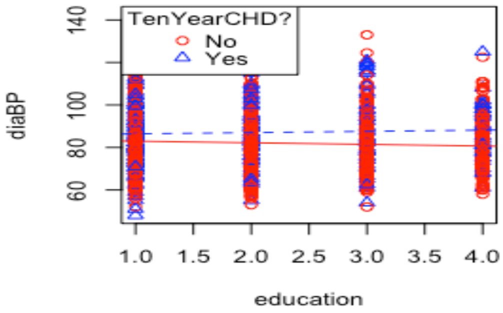

TenYearCHD?
No
Yes
diabetes
0.0
0.2
0.4
0.6
0.8
1.0
3.6
3.8
4.0
4.2
log(age)

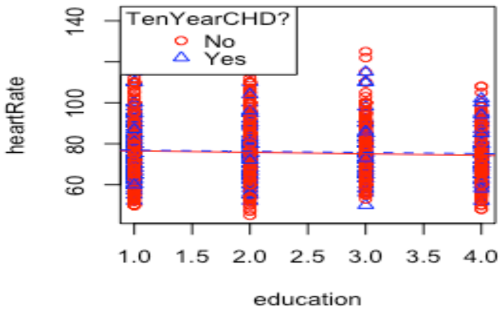

TenYearCHD?
No
Yes
cigsPerDay
0
20
40
60
1.0
1.5
2.0
2.5
3.0
3.5
4.0
education

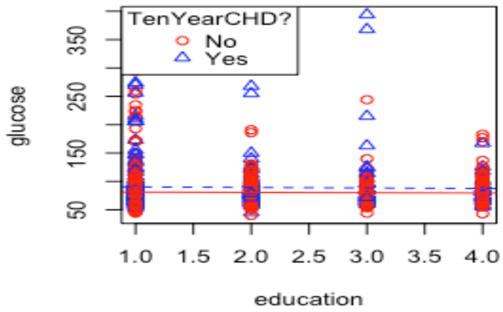

TenYearCHD?
No
Yes
totChol
100
300
500
1.0
1.5
2.0
2.5
3.0
3.5
4.0
education

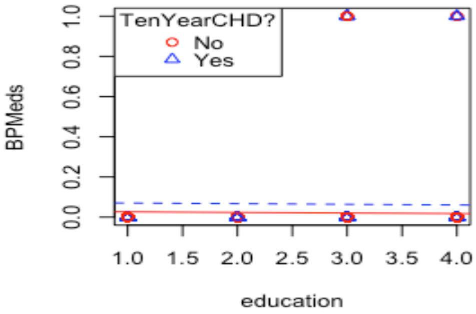

TenYearCHD?
No
Yes
sysBP
100
150
200
250
300
1.0
1.5
2.0
2.5
3.0
3.5
4.0
education

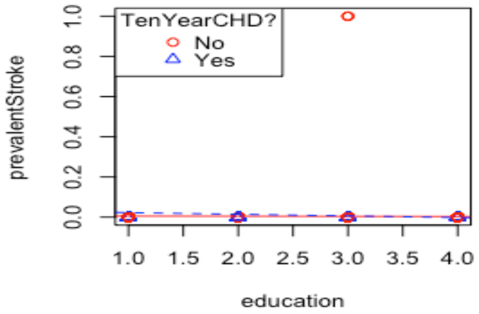

TenYearCHD?
○ No
△ Yes
BMI
20
30
40
50
1.0
1.5
2.0
2.5
3.0
3.5
4.0
education

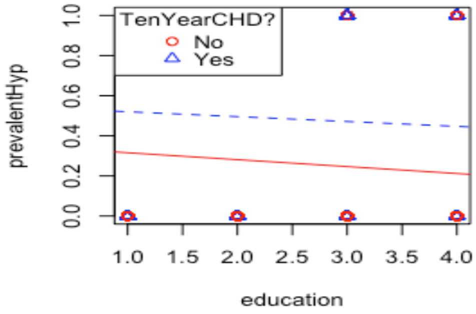

TenYearCHD?
○ No
△ Yes
diaBP
60
80
100
140
1.0
1.5
2.0
2.5
3.0
3.5
4.0
education

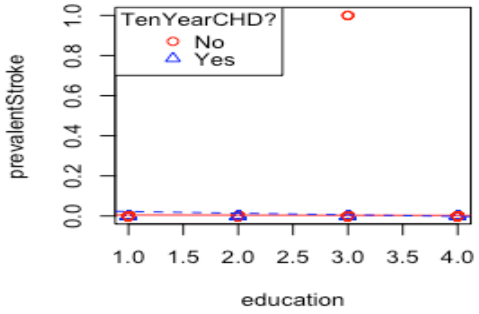

TenYearCHD?
○ No
△ Yes
heartRate
60
80
100
140
1.0
1.5
2.0
2.5
3.0
3.5
4.0
education

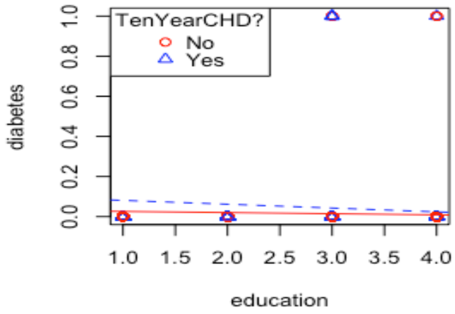

TenYearCHD?
○ No
△ Yes
glucose
50
150
250
350
1.0
1.5
2.0
2.5
3.0
3.5
4.0
education

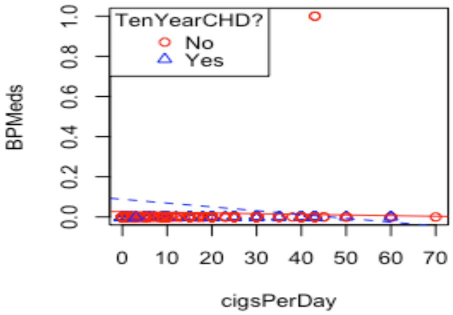

TenYearCHD?
○ No
△ Yes
BPMeds
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
education

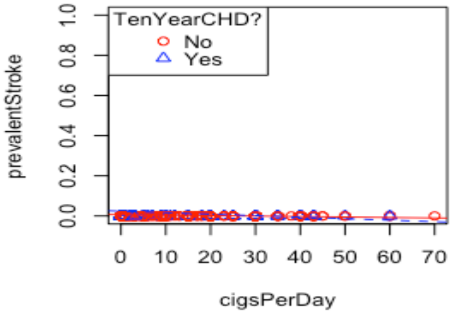

TenYearCHD?
○ No
△ Yes
prevalentStroke
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
education

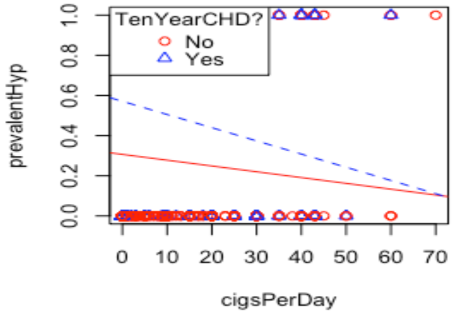

TenYearCHD?
○ No
△ Yes
prevalentHyp
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
education

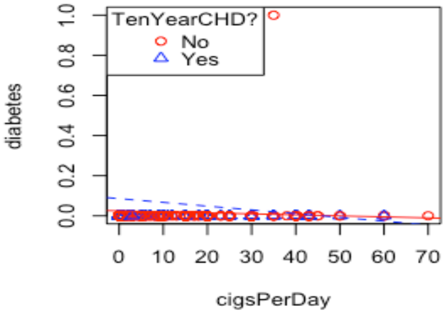

TenYearCHD?
○ No
△ Yes
prevalentStroke
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
education

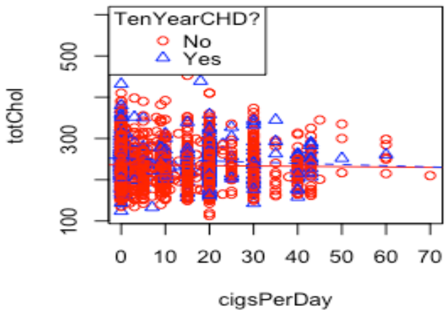

diabetes
0.0 0.2 0.4 0.6 0.8 1.0
TenYearCHD?
No
Yes
1.0 1.5 2.0 2.5 3.0 3.5 4.0
education

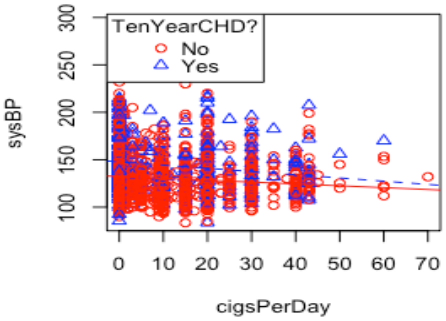

BPMeds
0.0 0.2 0.4 0.6 0.8 1.0
TenYearCHD?
No
Yes
0 10 20 30 40 50 60 70
cigsPerDay

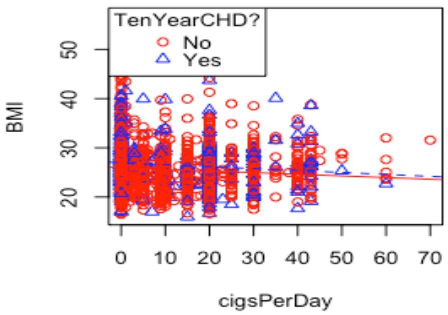

prevalentStroke
0.0 0.2 0.4 0.6 0.8 1.0
TenYearCHD?
No
Yes
0 10 20 30 40 50 60 70
cigsPerDay

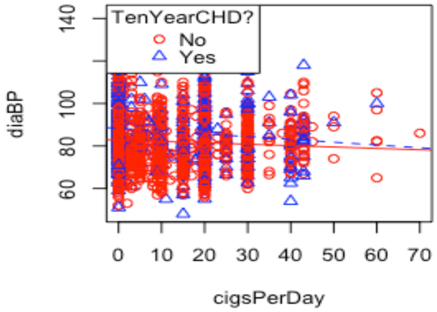

prevalentHyp
0.0 0.2 0.4 0.6 0.8 1.0
TenYearCHD?
No
Yes
0 10 20 30 40 50 60 70
cigsPerDay

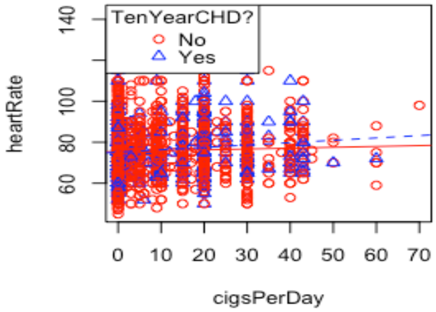

diabetes
0.0 0.2 0.4 0.6 0.8 1.0
TenYearCHD?
No
Yes
0 10 20 30 40 50 60 70
cigsPerDay

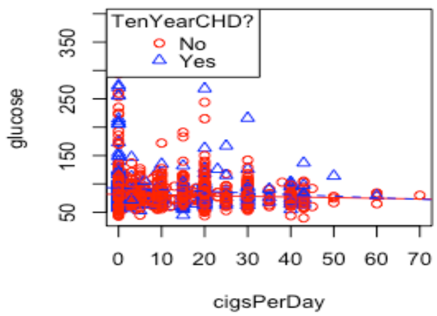

totChol
100 300 500
TenYearCHD?
No
Yes
0 10 20 30 40 50 60 70
cigsPerDay

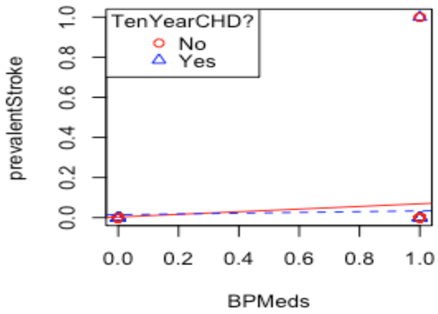

sysBP
100 150 200 250 300
TenYearCHD?
No
Yes
0 10 20 30 40 50 60 70
cigsPerDay

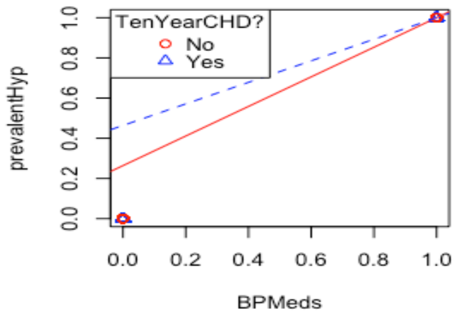

BMI
20 30 40 50
TenYearCHD?
No
Yes
0 10 20 30 40 50 60 70
cigsPerDay

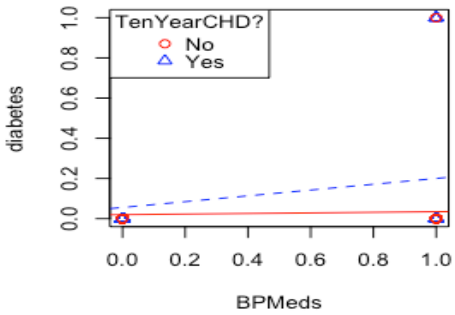

TenYearCHD?
No
Yes
diaBP
60
80
100
140
0
10
20
30
40
50
60
70
cigsPerDay

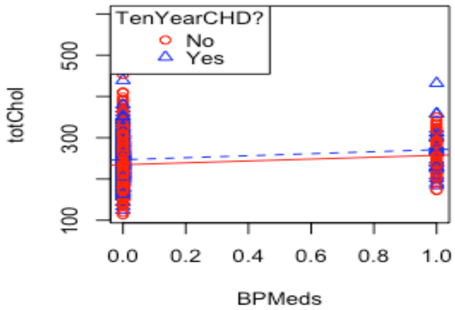

TenYearCHD?
No
Yes
heartRate
60
80
100
140
0
10
20
30
40
50
60
70
cigsPerDay

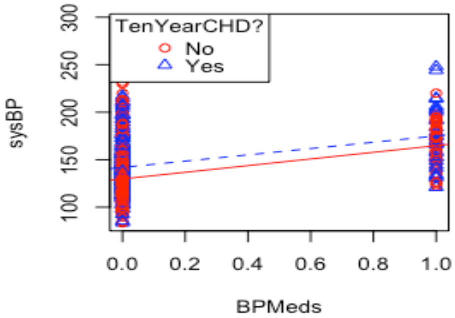

TenYearCHD?
No
Yes
glucose
50
150
250
350
0
10
20
30
40
50
60
70
cigsPerDay

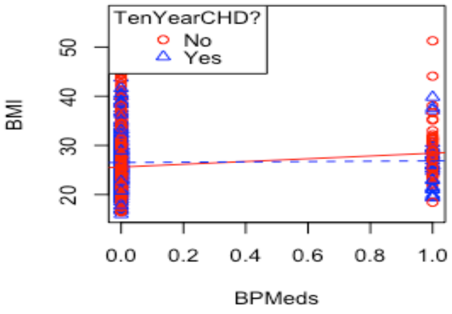

TenYearCHD?
No
Yes
prevalentStroke
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
BPMeds

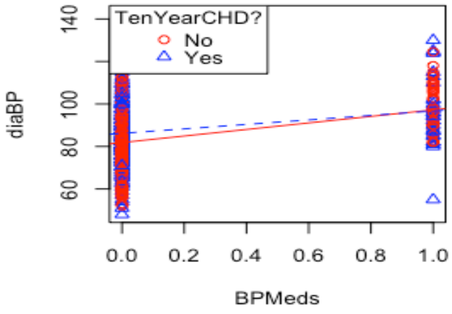

TenYearCHD?
No
Yes
prevalentHyp
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
BPMeds

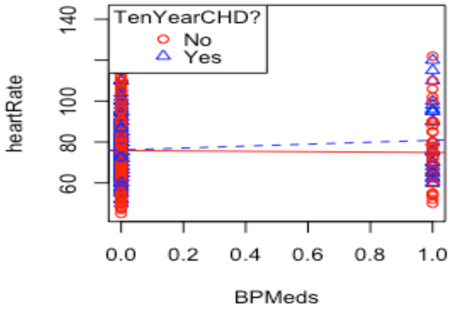

TenYearCHD?
No
Yes
diabetes
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
BPMeds

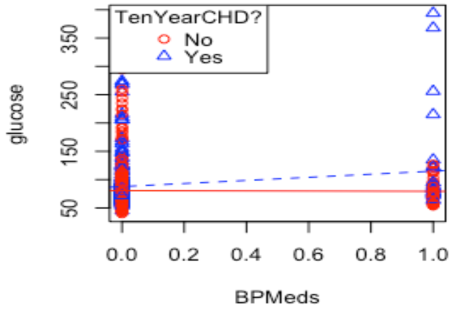

TenYearCHD?
No
Yes
totChol
100
300
500
0.0
0.2
0.4
0.6
0.8
1.0
BPMeds

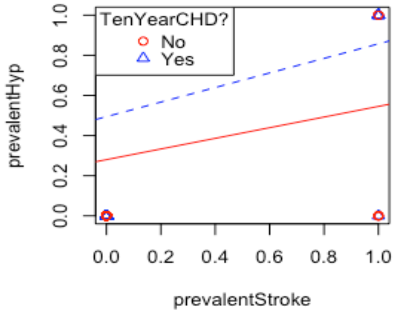

TenYearCHD?
No
Yes
sysBP
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0
BPMeds

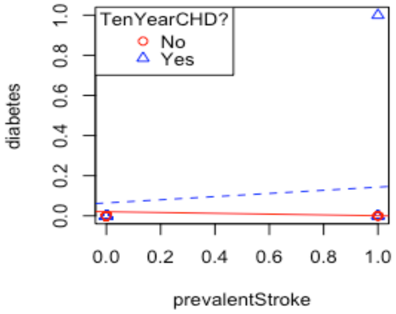

TenYearCHD?
No
Yes
BMI
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
BPMeds

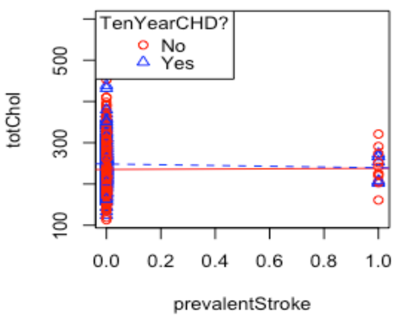

TenYearCHD?
No
Yes
diaBP
60
80
100
140
0.0
0.2
0.4
0.6
0.8
1.0
BPMeds

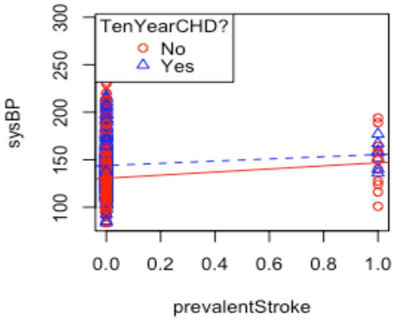

TenYearCHD?
No
Yes
heartRate
60
80
100
140
0.0
0.2
0.4
0.6
0.8
1.0
BPMeds

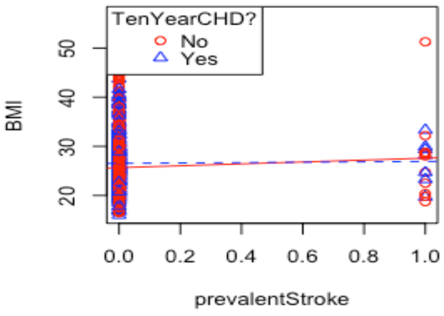

TenYearCHD?
No
Yes
glucose
50
150
250
350
0.0
0.2
0.4
0.6
0.8
1.0
BPMeds

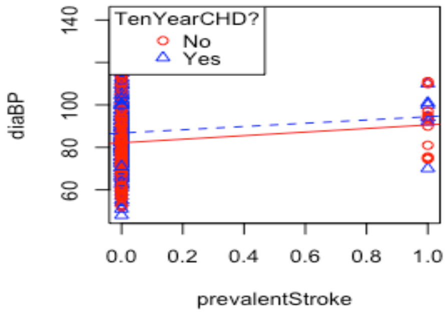

TenYearCHD?
No
Yes
prevalentHyp
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
prevalentStroke

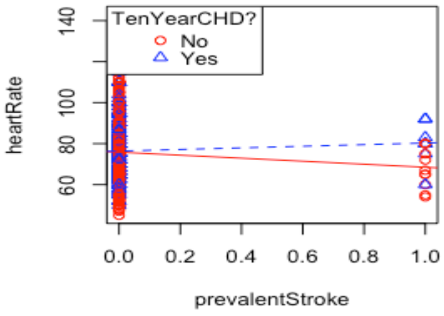

TenYearCHD?
No
Yes
diabetes
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
prevalentStroke

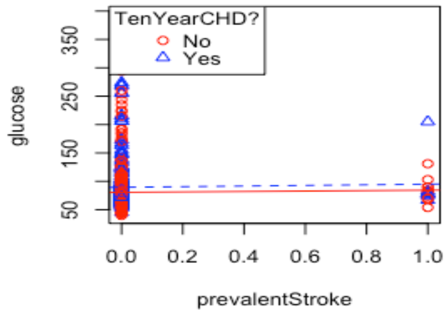

TenYearCHD?
No
Yes
totChol
100
300
500
0.0
0.2
0.4
0.6
0.8
1.0
prevalentStroke

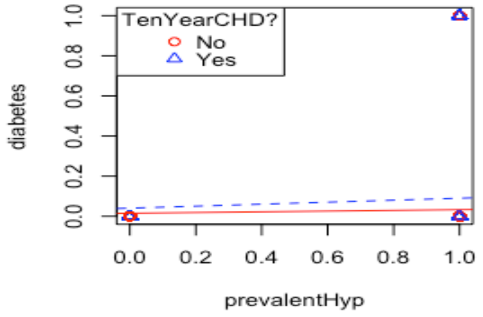

TenYearCHD?
No
Yes
sysBP
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0
prevalentStroke

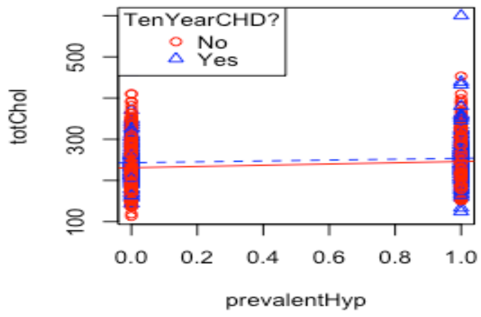

TenYearCHD?
No
Yes
BMI
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
prevalentStroke

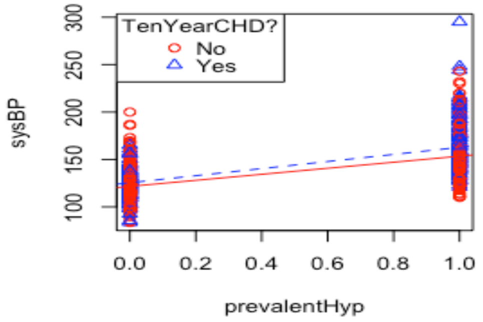

TenYearCHD?
No
Yes
diaBP
60
80
100
140
0.0
0.2
0.4
0.6
0.8
1.0
prevalentStroke

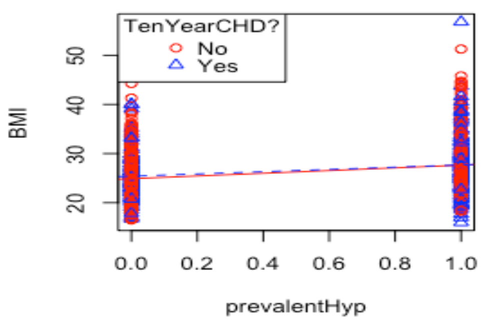

TenYearCHD?
No
Yes
heartRate
60
80
100
140
0.0
0.2
0.4
0.6
0.8
1.0
prevalentStroke

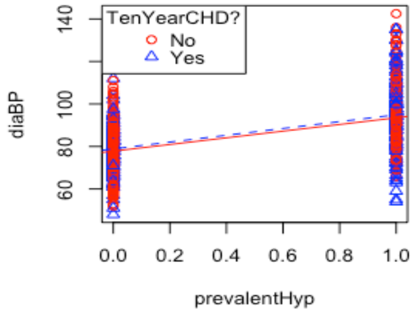

TenYearCHD?
No
Yes
glucose
50
150
250
350
0.0
0.2
0.4
0.6
0.8
1.0
prevalentStroke

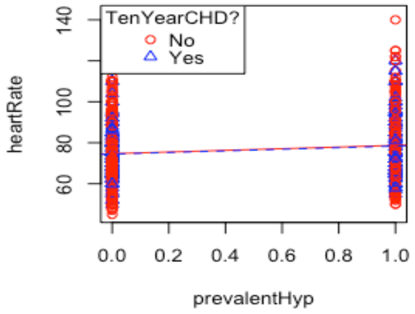

TenYearCHD?
No
Yes
diabetes
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
prevalentHyp

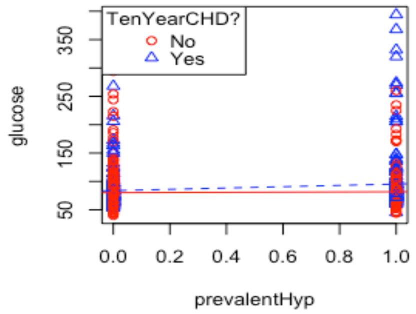

TenYearCHD?
No
Yes
totChol
100
300
500
0.0
0.2
0.4
0.6
0.8
1.0
prevalentHyp

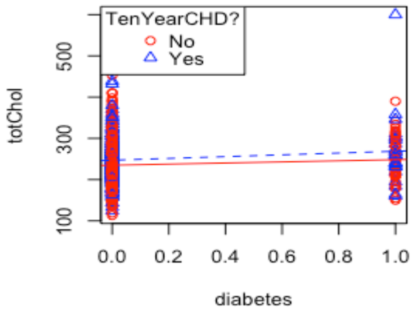

TenYearCHD?
No
Yes
sysBP
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0
prevalentHyp

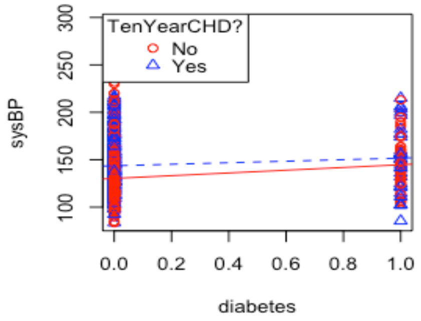

TenYearCHD?
No
Yes
BMI
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
prevalentHyp

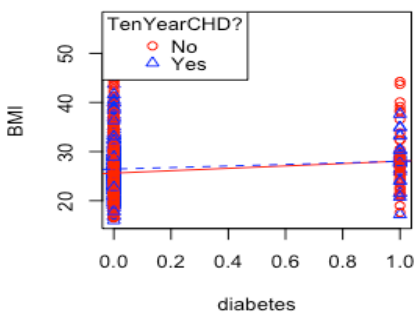

TenYearCHD?
No
Yes
diaBP
60
80
100
140
0.0
0.2
0.4
0.6
0.8
1.0
prevalentHyp

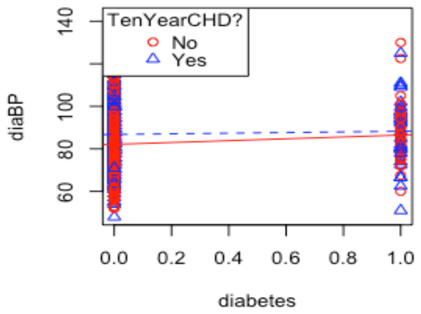

TenYearCHD?
No
Yes
heartRate
60
80
100
140
0.0
0.2
0.4
0.6
0.8
1.0
prevalentHyp

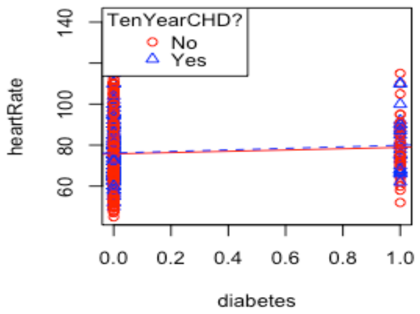

TenYearCHD?
No
Yes
glucose
50
150
250
350
0.0
0.2
0.4
0.6
0.8
1.0
prevalentHyp

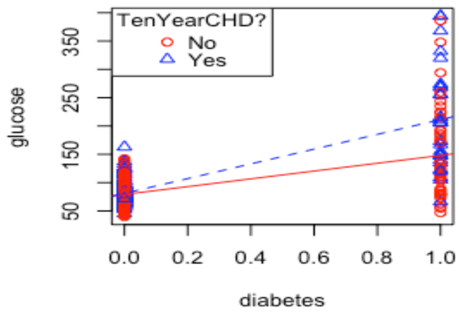

TenYearCHD?
No
Yes
totChol
100
300
500
0.0
0.2
0.4
0.6
0.8
1.0
diabetes

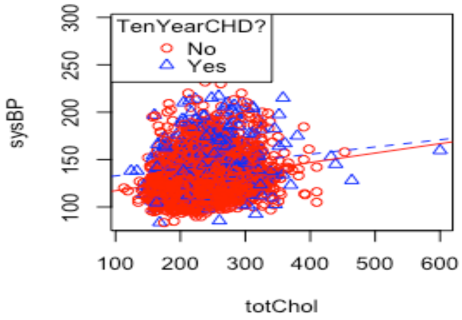

TenYearCHD?
No
Yes
sysBP
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0
diabetes

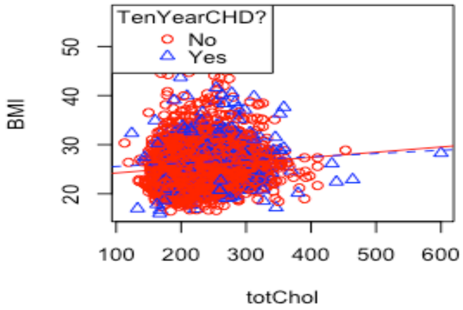

TenYearCHD?
No
Yes
BMI
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
diabetes

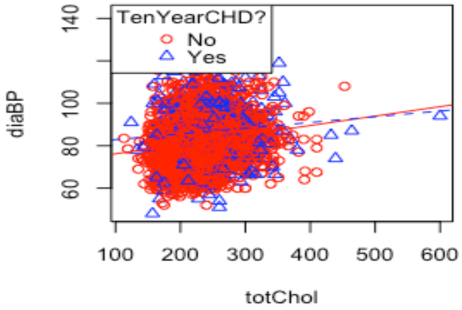

TenYearCHD?
No
Yes
diaBP
60
80
100
140
0.0
0.2
0.4
0.6
0.8
1.0
diabetes

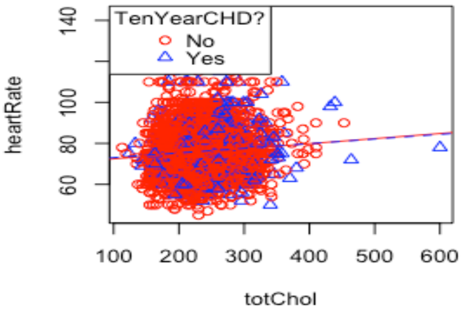

TenYearCHD?
No
Yes
heartRate
60
80
100
140
0.0
0.2
0.4
0.6
0.8
1.0
diabetes

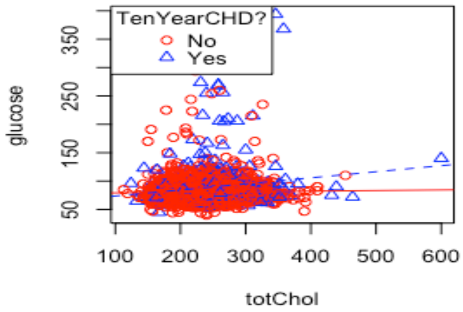

TenYearCHD?
No
Yes
glucose
50
150
250
350
0.0
0.2
0.4
0.6
0.8
1.0
diabetes

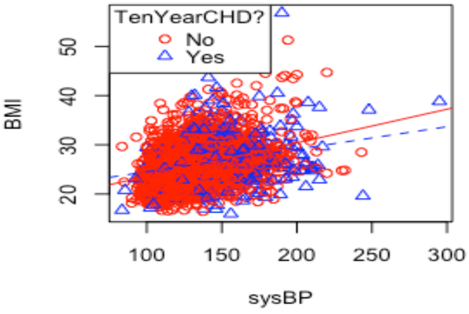

TenYearCHD?
No
Yes
sysBP
100
150
200
250
300
100
200
300
400
500
600
totChol

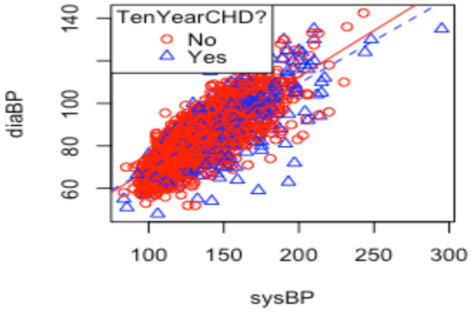

TenYearCHD?
No
Yes
BMI
20
30
40
50
100
200
300
400
500
600
totChol

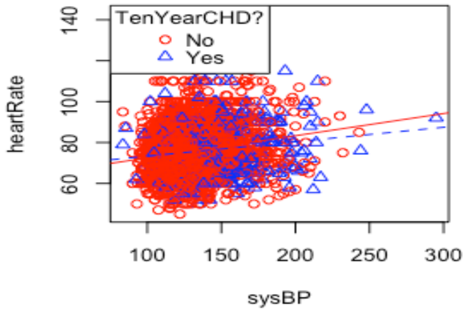

TenYearCHD?
No
Yes
diaBP
60
80
100
140
100
200
300
400
500
600
totChol

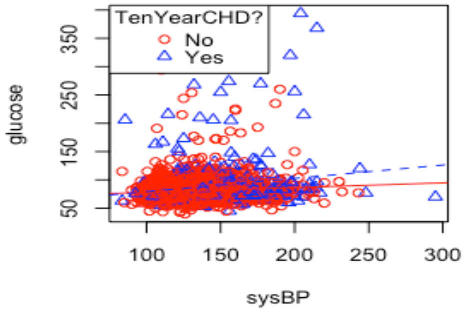

TenYearCHD?
No
Yes
heartRate
60
80
100
140
100
200
300
400
500
600
totChol

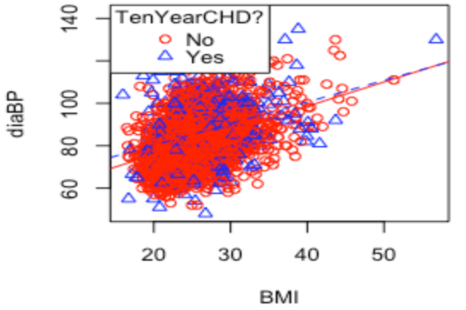

TenYearCHD?
No
Yes
glucose
50
150
250
350
100
200
300
400
500
600
totChol

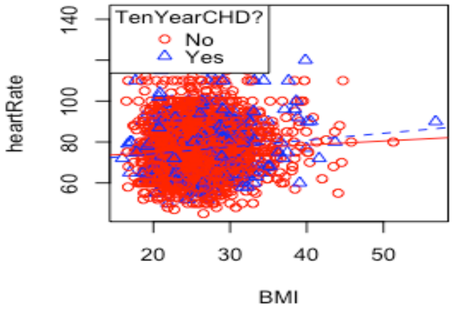

TenYearCHD?
No
Yes
BMI
20
30
40
50
100
150
200
250
300
sysBP

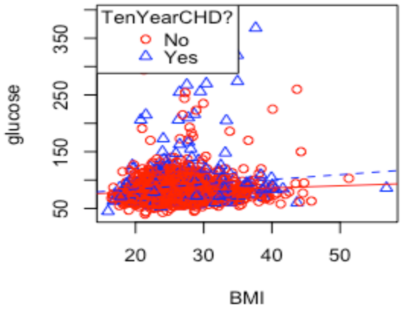

TenYearCHD?
No
Yes
diaBP
60
80
100
140
100
150
200
250
300
sysBP

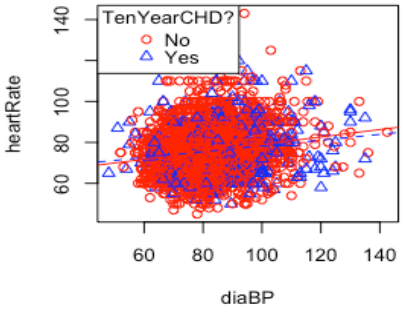

TenYearCHD?
No
Yes
heartRate
60
80
100
140
100
150
200
250
300
sysBP

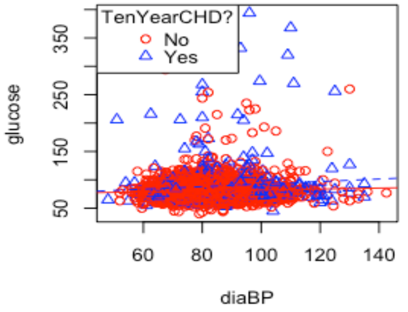

TenYearCHD?
No
Yes
glucose
50
150
250
350
100
150
200
250
300
sysBP

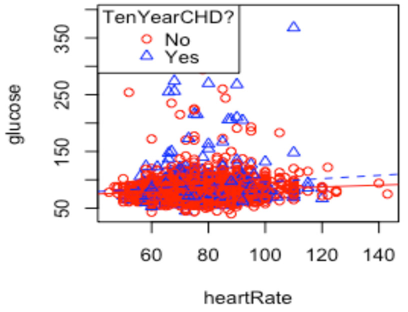

TenYearCHD?
No
Yes
diaBP
60
80
100
140
20
30
40
50
BMI


TenYearCHD?
No
Yes
heartRate
60
80
100
140
20
30
40
50
BMI


TenYearCHD?
No
Yes
glucose
50
150
250
350
20
30
40
50
BMI


TenYearCHD?
No
Yes
heartRate
60
80
100
140
60
80
100
120
140
diaBP


TenYearCHD?
No
Yes
glucose
50
150
250
350
60
80
100
120
140
diaBP


TenYearCHD?
No
Yes
glucose
50
150
250
350
60
80
100
120
140
heartRate