PREDICTION OF CORONARY HEART DISEASE USING LOGISTIC MODELING

A PROJECT REPORT
Presented to the Department of Mathematics and Statistics
California State University, Long Beach

In Partial Fulfillment of the Requirements of the Degree
Master of Science in Applied Statistics

Faculty Reviewer:

Kagba Suaray, Ph.D.

By Marian Lu
B.S., 2019, University of Minnesota, Twin Cities

May 2024

Abstract

The objective is to build a logistic regression model that can accurately identify cases of coronary heart disease in patients. The data is from a cardiovascular study on residents in Framingham, Maryland, containing variables describing the patients' demographics, behavior, medical history, and current health condition. The most remarkable thing about the data is the class imbalance in the response variable, CHD, since it will impact the model's ability to predict instances of the minority class. Logistic regression models were built and tested with a backwards selection approach. The full model violated assumptions of multicollinearity so variables were removed accordingly. The reduced model was fitted, and variables previously taken out were added back in as interaction terms. Since the interaction terms did not have a significant contribution, another iteration of the reduced model was fitted: the weighted reduced model. These four models were evaluated using model selection criteria such as AIC, F-1 score, and sensitivity. The weighted model outperformed the others due to its versatility and ability to identify true positives, so it was chosen as the final model. Further analysis was carried out, analyzing the parameter estimates and acknowledging the limits of the current model. Overall, the best model is one that has the highest rate of correctly identifying individuals at risk for CHD while still satisfying the assumptions of logistic regression.

Table of Contents

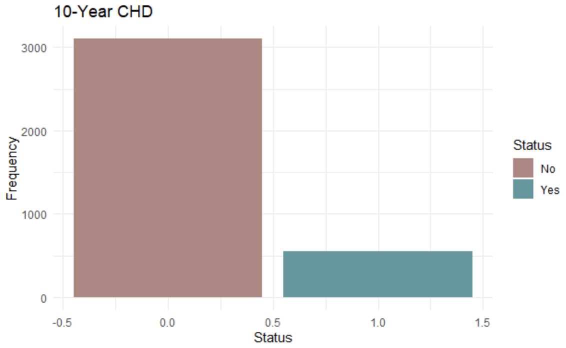
1 Introduction

Around 20.5 million adults in the United States suffer from coronary artery disease, a type of heart disease where the arteries of the heart fail to transfer enough oxygen in the form of blood to the heart. As the most common type of heart disease in the United States, identifying signs of the disease in its early stages can help with a more timely and effective treatment.

2 Data

The data is provided by a cardiovascular study on residents of the town of Framingham, Massachusetts. The variables include information about the patient's demographic, behavior, and risk factors.

Figure 2.1 Distribution of Response Variable



No Heart Disease	Heart Disease
84.76% (3099)	15.23% (557)

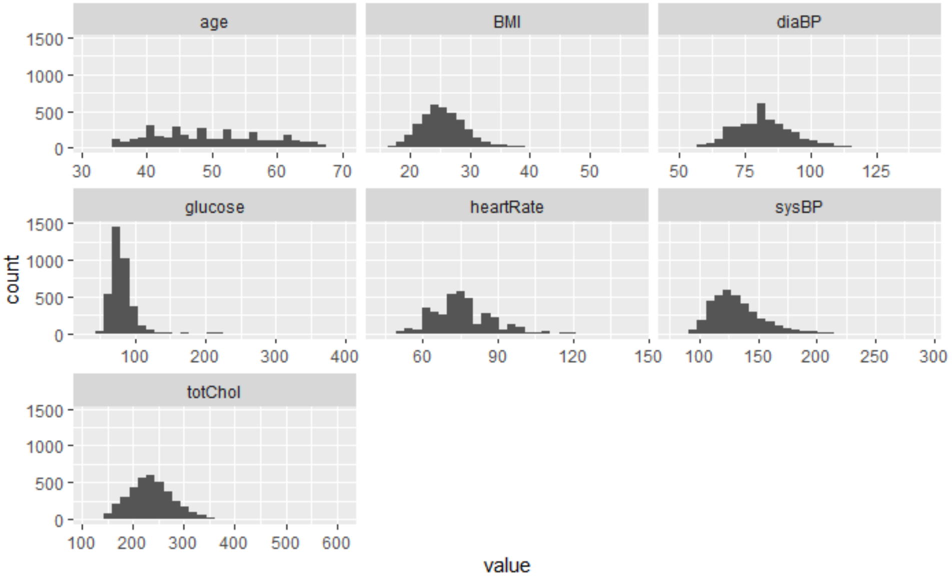
There is a large class imbalance in the response variable. This is to be expected, since despite the study participants were sampled from a residential population. However, this class imbalance would affect model performance, specifically its ability to predict instances of the minority class.

Table 2.1. Variable Description

Variable	Type	Description
Sex	Binomial	Male or female
Age	Binomial	Age of patient
Education	Discrete	
BPMeds	Binomial	Whether or not patient has history of using blood pressure medication
Current Smoker	Binomial	Whether patient currently smokes
Cigs Per Day	Discrete	
Prevalent Stroke	Binomial	Whether patient had previously had a stroke
Prevalent Hyp	Binomial	Whether or not patient has history of hypertension
Diabetes	Binomial	Whether or not the patient has history of diabetes
Totchol	Continuous	Total cholesterol level (mg/dL)
SysBP	Continuous	Systolic blood pressure (mmHg)
DiaBP	Continuous	Diastolic blood pressure (mmHg)
BMI	Continuous	Body Mass Index
HeartRate	Continuous	Patient heart rate (BPM)
Glucose	Continuous	Glucose levels (mmol/L)

Continuous Variables

Figure 2.2 Distribution of Medical Continuous Variables



These variables mostly follow a normal distribution. The variable for glucose level, systolic blood pressure, and total cholesterol has several outliers, which may also affect model fitting. Diastolic blood pressure, systolic blood pressure, and total cholesterol have relatively large medians compared to the other continuous predictors, so they may need to be scaled.

Categorical Variables

Figure 2.3 Distribution of Discrete Variables

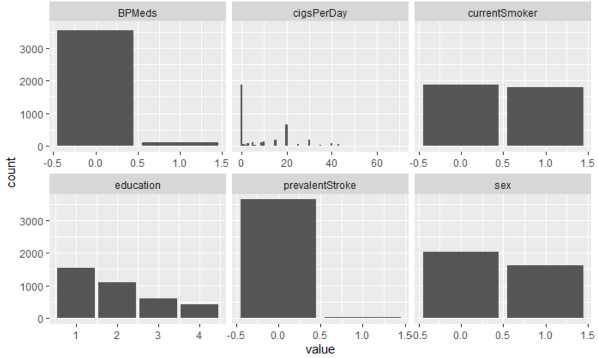


Table 2.2 Binary Variable Class Proportions

Variable	Yes	No
currentSmoker	48.91% (1868)	51.09% (1788)
BPMeds	3.04% (3545)	96.96% (111)
prevalentStroke	0.57% (3635)	99.43% (21)
prevalentHyp	31.15% (2517)	68.85% (1139)
diabetes	2.71% (3557)	97.29% (99)

There seems to be class imbalances in most of the binary variables other than currentSmoker and prevalentHyp. Cigarettes smoked per day almost follow a normal distribution except for the large number of non-smokers. Predictors with large class imbalances may not offer much contribution to the model since they are almost monotonous, but we will investigate the true effect in the following section.

3 Logistic Regression

Logistic Regression models are nonlinear regression models that are widely used for predicting qualitative variables. The simple logistic regression model follows the general form:

$$Y_i = \frac{exp(\beta_0 + \beta_1)}{1 + exp(\beta_0 + \beta_1)} + \varepsilon_i$$

Where x_i is deterministic and $y_i \sim Bernoulli(\pi_i)$. The Bernoulli distribution is the discrete distribution for the probability of a random variable. It is a good fit for the data set because the response variable is a binary variable. Since the variance of the Bernoulli distribution is $\pi_i(1 - \pi_i)$, which is dependent on x_i logistic regression uses the maximum likelihood approach for parameter estimation.

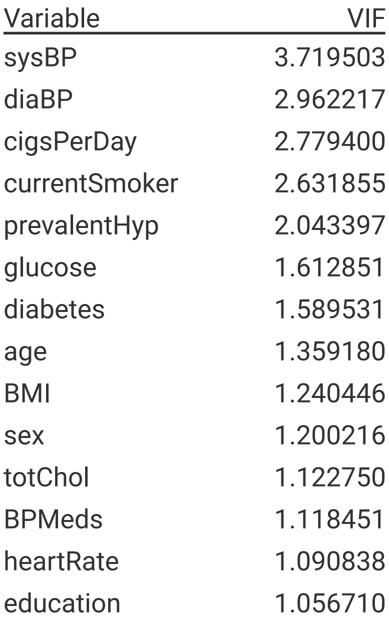
$$L(\beta_0, \beta_1) = \prod_{o=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

Using these methods, a logistic regression model will be fitted on a training data set to predict whether or not a patient will be affected by coronary heart disease in the next ten years. 80% of the original dataset was allocated for the training set and 20% for the testing set.

Model 1: Full Model

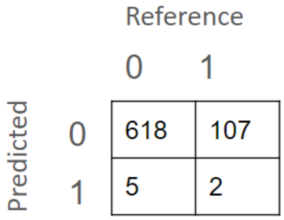
A full model was fitted with all 15 variables.

Table 3.1 VIFs for Full Model

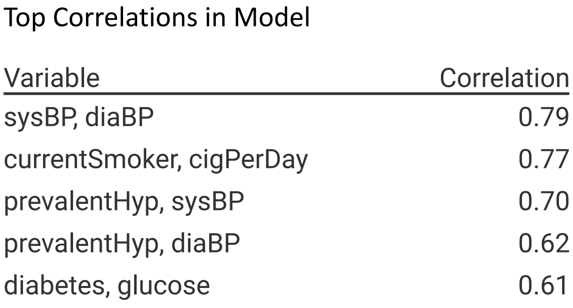


Variable	VIF
sysBP	3.719503
diaBP	2.962217
cigsPerDay	2.779400
currentSmoker	2.631855
prevalentHyp	2.043397
glucose	1.612851
diabetes	1.589531
age	1.359180
BMI	1.240446
sex	1.200216
totChol	1.122750
BPMeds	1.118451
heartRate	1.090838
education	1.056710

Table 3.2 Confusion Matrix for Full Model



Predicted \ Reference	0	1
0	618	107
1	5	2



Top Correlations in Model

Variable	Correlation
sysBP, diaBP	0.79
currentSmoker, cigPerDay	0.77
prevalentHyp, sysBP	0.70
prevalentHyp, diaBP	0.62
diabetes, glucose	0.61

Most of the VIFs seem reasonably low, with only a few greater than 2. To further assess the multicollinearity of the predictor variables, a table of the variable pairs with the highest correlations was generated (Appendix 1A).

Model 2: Reduced Model

The variables that were responsible for multicollinearity were removed (correlation>0.5), and only variables with low p-values remain (Appendix 2A).

Table 3.3 VIFs for Reduced Model

Variable	VIF
sysBP	1.284785
age	1.279233
cigsPerDay	1.167422
sex	1.128895
totChol	1.111162
glucose	1.026976

Table 3.4 Confusion Matrix for Reduced Model

		Reference	
		0	1
Predicted	0	620	107
	1	3	2

Now all the VIFs are reasonably low, but the confusion matrix looks similar to the one in the full model, which doesn't show any improvement in the model's predictive ability.

Model 3: Reduced Model with Interaction Terms

To account for the variables taken out of the model, they were added back as interaction terms. Variables that had a low p-value from the full model but taken out due to their high correlation with other variables were added back as an interaction term with the relevant variable that it is highly correlated with.

Table 3.5 VIFs for Reduced Model with Interaction Terms

Variable	VIF
age	1.198144
cigsPerDay	1.162950
sysBP*diaBP	1.134199
sex	1.118787
glucose*totCholesterol	1.065951

Table 3.6 Confusion Matrix for Reduced Model with Interaction Terms

		Reference	
		0	1
Predicted	0	621	107
	1	2	2

Model 4: Reduced Model with Weights

The previous models had good performance when predicting true negatives, or cases of no risk of coronary heart disease. The objective of the study is to observe the prediction of the disease, so a weight was added to the response variable in an effort to improve the prediction of true positives

For generalized linear models in R, the default weight is 1. To attempt to increase the model's predictions of positive cases, a weight of 4 was assigned to the minority class. Weights are usually a ratio of the majority class to minority class, in which this case would be approximately 5 (0.85:0.15), but I decided to use a more conservative weight of the ratio - 1.

VIFs for Reduced Model with Weights

Variable	VIF
age	1.219370
sysBP	1.192155
cigsPerDay	1.164384
sex	1.120754
glucose	1.025083

Confusion Matrix for Reduced Model with Weights

		Reference	
		0	1
Predicted	0	530	89
	1	93	20

The weight successfully increased the number of predictions for class 1, at the expense of more false positives. This tradeoff will be further evaluated in the model selection portion of the report.

4 Residual Diagnostics

Residual plots are used for analyzing model inadequacy, nonconstant variance, and response outliers. Since binary logistic regression all have nonconstant variance, we will only investigate model inadequacy and response outliers.

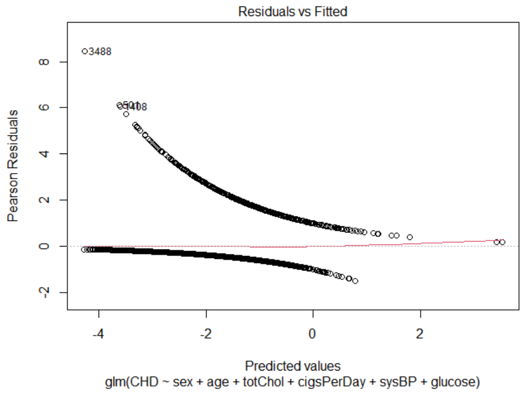
Model Adequacy: Lowess Smooth

A correct logistic regression model should follow

$$E\{Y_i - \widehat{\pi_i}\} = E\{e_i\} = 0$$

Assuming that $E\{Y_i = \pi_i\}$, a lowess smooth of the plot of residuals against the fitted values from the model should resemble a horizontal line with an intercept of 0.

Figure 4.1 Residual Plot with Lowess Smooth



The residuals vs. fitted plot confirms the adequacy of the model under the assumptions of the lowess smoothing method. The red line has an intercept of 0 and has a slope close to 0.

Response Outliers

The residuals vs. fitted plot can also help identify outliers. When observing Fig., there are three points that emerge as outliers.

Table 4.1 Medians of Predictors and Outlier Values

Variable	Median	Outlier 1	Outlier 2	Outlier 3
sex	0.00	0.00	0.0	0.00
age	49.00	35.00	42.0	58.00
edu	2.00	2.00	3.0	1.00
currSmk	0.00	1.00	0.0	0.00
cigs/day	0.00	20.00	0.0	0.00
BPMeds	0.00	0.00	0.0	0.00
Prev Stroke	0.00	0.00	0.0	0.00
Prev Hyp	0.00	0.00	0.0	0.00
dia	0.00	0.00	0.0	1.00
totChol	234.00	168.00	464.0	260.00
sysBP	128.00	83.50	128.0	85.50
diaBP	82.00	55.00	87.0	51.00
BMI	25.38	16.71	22.9	20.76
Heart Rate	75.00	79.00	72.0	87.00
glucose	78.00	63.00	72.0	206.00
CHD	0.00	1.00	1.00	1.00

The information for the three outliers are obtained. The variables where they are one standard deviation or more above the median are highlighted. Most of these high-deviance values seem reasonable, except for Outlier 1 smoking 20 cigarettes a day, Outlier 2 having a total cholesterol level of 464 mg/dL, and Outlier 3 having a glucose level of 206 mmol/L, which may be a result of an error in data entry.

	Standard Deviation
totChol	44.096
glucose	23.91

While smoking 20 cigarettes is uncommon for the average person, a decent portion of patients smoke 20 or more cigarettes according to the distribution plot. Healthy cholesterol levels range from 125 to 200 mm/dL, but the median for patients in the study is slightly higher. Although Outlier 2 and 3 have values 5 standard deviations above the mean (total cholesterol and glucose level respectively), this is still within the range of possible levels for these metrics. Additionally, they have all tested positive for heart disease, which makes their data valuable since it can help understand the disease better. Despite their contribution to the prediction, we will still investigate the effect on model performance by removing the outliers.

Figure 4.2 Residuals vs Fitted Plot (Post-Outlier Removal)

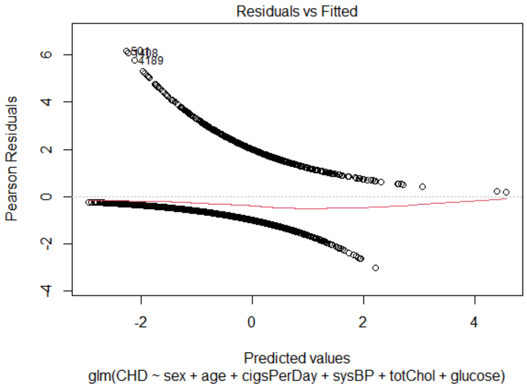
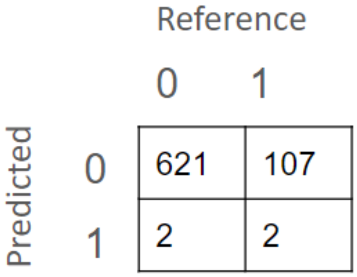


Table 4.2 Confusion Matrix for Weighted Model (Post-Outlier Removal)



		Reference	
		0	1
Predicted	0	621	107
	1	2	2

After removing the outlier, the residuals vs. fitted plot looks very similar to the original data. Additionally, the lowess smooth curve shows better performance in the previous plot prior to removing the outliers. The weighted model was fitted on the new training data, and its performance was similar to the performance on the old data, with only a small decrease in AIC. This change was not impactful enough for me to proceed with the training data. All the outliers were classified as positive cases for CHD, which makes them valuable to the dataset. Additionally, the distributions of the variables suggest the presence of multiple outliers other than the one discussed in this section. Removing all of them can improve model performance, but it would also affect the model's ability to predict positive cases of CHD so I did not decide to proceed with the new training data.

5 Model Selection

R-Squared, or the coefficient of determination, is a measure of the proportion of variance in the dependent variable that can be explained. An R-squared value close to 1 is an indicator of a good model. Akaike information criterion (AIC) is a score of which model is the best model for a common dataset based on goodness of fit and number of parameters used. The lower the AIC score, the better the model. Bayesian Information Criterion (BIC) is similar to AIC. While AIC asymptotically selects the model that minimizes mean square error, BIC will favor the true model regardless of sample size. F-1 score is a criterion for classification models that balances precision and recall. Accuracy is the percentage of correct classifications that the model is able to predict. Lastly, sensitivity is a measure of how well the model can identify positive instances. A positive instance in this study is the event of heart disease, which makes sensitivity an important metric in selecting the best performing model. Consider the table below.

Table 5.1 Comparison of Models

	Model 1	Model 2	Model 3	Model 4
	Full Model	Reduced Model	w/ Interactions	w/ Weights
AIC	2218.5	2207.7	2217.6	5022.4
BIC	2314.2	2249.6	2253.5	5058.3
F-1 Score	0.0345	0.0358	0.0354	0.164
Accuracy	0.847	0.848	0.851	0.749
Sensitivity	0.018	0.018	0.018	0.165

Based on all the model selection criteria, Model 4, the reduced model with weights, emerges as the top performer. Both Model 2 and Model 4 have the most ideal scores in two categories (AIC+BIC and F-1 Score+Sensitivity respectively). However, in the context of the study, sensitivity is more relevant because the objective is to correctly identify positive cases of CHD. Despite high AIC and BIC scores and a lower accuracy, Model 4 is the most versatile and best fit for the objective of the data.

6 Ridge Regression

Ridge regression is often used to adjust the coefficients for overfitting or multicollinearity. The ridge standardize estimators are calculated by incorporating a biasing constant c into the least squares normal equations and shrinking the coefficients.

As the bias constant gets larger, the MSE of the estimator b^R increases while the variance component decreases. The MSE has a value of

$$E\{b^R - \beta\}^2 = \sigma^2\{b^R\} + (E\{b^R\} - \beta)^2$$

where b^R is a biased estimator, β is the true parameter, and σ is the standard deviation.

This method was applied to the full model, since the multicollinearity within the predictors will be addressed by ridge regression. A weight was also assigned to the response variable to improve the model’s ability to predict true positives.

After fitting the ridge model, we obtained an optimized bias constant value of c = 0.013. This is a relatively low value, meaning that there is a low amount of bias in the estimators

Table 6.1 Ridge Model Summary

	Model 5
	Ridge Model
AIC	1198.347
Accuracy	0.851
Sensitivity	0.018

The ridge model did not outperform the previous models. Its AIC is lower and its accuracy is around the same as the non-weighted models. Despite a weight being added, the ridge model behaves similarly to the non-weighted models when it comes to its prediction rate (Appendix 4B). While its AIC proves it to be a contender to the non-weighted models, it is subpar to the weighted model by the criterion of the study. This may be due to my inexperience with this concept, but as of now, I decided to not move forward with the model.

7 Final Model

Based on the model selection criterion, the reduced model with weights is the best model.

$$Y_i = \frac{exp(-7.72\ +.503x_1 + .0685x_2 + .0021x_3 + .019x_4 + .019x_5 + .010x_6)}{1 + exp(-7.72\ +.503x_1 + .0685x_2 + .0021x_3 + .019x_4 + .019x_5 + .010x_6)} + \varepsilon_i$$

The reduced model had a respectable R-square value relative to the other models, the lowest AIC, and accuracy and sensitivity similar to most of the other models. Since logistic regression is a form of nonlinear regression, the coefficient interpretation is not as straightforward. A unit increase in X is not constant for the logistic regression model because of its nonlinear characteristics. Thus, the effect of the variables can be interpreted by their odds ratio. The odds ratio is calculated by exponentiating the parameter estimate.

Odds Ratio Equation:

$$\widehat{OR} = \frac{odds_2}{odds_1} = exp(b_1)$$

Table 7.1 Estimated Odds Ratios

Variable	Odds
sex	1.6917
age	1.0693
totChol	1.0021
cigsPerDay	1.0214
sysBP	1.0190
glucose	1.0092

Based on the table, sex had the largest impact on the odds of coronary heart disease. An odds ratio of 1.69 for sex means that the odds of a male having CHD is 69% higher than the odds of a female having it (female=0, male=1). A unit in age increases the odds of having heart disease by 6.9%. Aging is associated with many health issues, such chronic obstructive pulmonary disease and diabetes. The decrease in overall health condition can make the patient more susceptible to heart disease. A unit increase in total cholesterol only results in a 0.2% increase in odds. However, the values for cholesterol are generally higher because a single unit of cholesterol (mm/dL) is not very impactful. Typically, cholesterol levels are categorized by increments of 50 mm/dL. For example, the healthy range is from 125 to 200 mm/dL, an unhealthy range is from 200 to 250 mm/dL, and an extreme range above 250 mm/dL. With this interpretation, an increase of 50 mm/dL would result in a 10% increase in the odds of having CHD. A more comprehensive approach could be to reassign ordinal categories to the cholesterol variable to see how that would affect the model. A unit increase in cigarettes per day results in a 2.14% increase in the odds of having CHD. Smoking has always been responsible for many health issues, so it is expected that it would also put individuals at risk for heart disease. Similar to total cholesterol, systolic blood pressure and glucose show only a small increase in odds ratio, but these are usually measured in large quantities so a multiple unit increase is more effective than a single unit increase for interpreting the odds ratio..

Further evaluation of the model can be conducted by observing the ANOVA (or ANODA in this case) table.

Table 7.2 Analysis of Deviance Table for Final Model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
	NA	NA	2923	5806	NA
x1 sex	1	59.86	2922	5746	1.02e-14
x2 age	1	463.1	2921	5283	1.04e-102
x3 totChol	1	21.01	2920	5262	4.58e-06
x4 cigs PerDay	1	38.76	2919	5223	4.80e-10
x5 sysBP	1	168.63	2918	5055	1.47e-38
x6 glucose	1	51.88	2917	5003	5.887e-13

Deviance, also represented by G^2, is the analog of the F-statistic. The likelihood ratio test can be performed by comparing the G^2 value of a ratio between two models to the critical value, which follows the Chi-Square distribution.

$$- 2log(\frac{L_R}{L_F}) \sim \chi^2_{df=1}$$

Based on the ANOVA table, the F-statistics are significant with the addition of each value when tested against the critical value of $\chi^2_{0.5,\, df=1}$= 3.84. Additionally, the significance of each variable can also be evaluated by analyzing the p-values in the Pr(>Chi) column. Most of the p-values < 0.05, meaning that their corresponding model is significant.

Accuracy

The ability of the model to predict new observations was evaluated by testing the model with five different samples of data. Each sample is 20% of the original data and randomly sampled.

Table 7.3 Accuracy of Final Model Across Random Samples

	Accuracy
Test 1 (seed=100)	0.7423
Test 2 (seed=200)	0.7510
Test 3 (seed=300)	0.7456
Test 4 (seed=400)	0.7456
Test 5 (seed=500)	0.7360
Average	0.8420

The accuracy of the model is decent. It is able to consistently perform well, scoring within a range of (0.73, 0.76). While these values are lower than the accuracies of the non-weighted models, their high accuracies may also be inflated as a result of class imbalance in the testing data. The majority class accounts for 84.76% of the response variable values, so the model could predict 0 for every single prediction and still have a decent accuracy score. The weighted model having lower accuracies is a trade-off for its flexibility, which is more valuable in the context of the data.

Limitations

While the final model is the best out of all the models tested, it is still objectively very weak. For example, the final model has the highest F-1 score of 0.164 but it is still below the benchmark of an F-1 score that is considered adequate. The weighted model predicts CHD at a similar rate to its true proportion (15% predicted, 15% true proportion), the majority of these predictions are wrong. Despite this resulting in a high false positive rate, the false negative rate is still higher than ideal. Some approaches to this issue can be oversampling and undersampling, nonparametric regression methods, or other advanced machine learning methods like neural networks.

8 Conclusion

Patient sex, age, total cholesterol, cigarettes smoked per day, systolic blood pressure, and glucose levels have a significant impact on an individual's risk of being affected by coronary heart disease.

Adding weights to the model allowed it to correctly identify more patients that are at risk for CHD. Despite its tendency to have a high false positive rate, patients falsely classified with CHD can still be diagnosed with a different kind of heart disease, such as arrhythmia or peripheral arterial disease. Some things that may improve model performance is a study population that may be more at risk. Including lab information, such as platelets and creatinine levels, can also help us better understand CHD.

Ultimately, even a perfectly optimized predictive model cannot replace the medical knowledge and service of doctors and nurses. It should mostly be used to help patients be more aware of their own health condition and help hospitals better identify at-risk groups.

9 Bibliography

"Aging and Health." *World Health Organization*, World Health Organization, Accessed 14 May 2024

Bias-Variance Decomposition in Ridge Linear Regression, www.cs.cornell.edu/courses/cs4780/2023fa/slides/Bias_Var_Ridge.pdf. Accessed 14 May 2024.

Dileep. "Logistic Regression to Predict Heart Disease." *Kaggle*, 7 June 2019, www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression/data.

Karthik Kumar, MBBS, and MD Shaziya Allarakha. "What Does It Mean When the Bottom Number of Your Blood Pressure Is over 100?" *MedicineNet*, MedicineNet, 23 Jan. 2023, www.medicinenet.com/bottom_number_of_your_blood_pressure_is_over_100/article.htm.\

Kutner, Michael H. "Applied Linear Regression Models Fourth Edition" McGraw-Hill Irwin, 2004

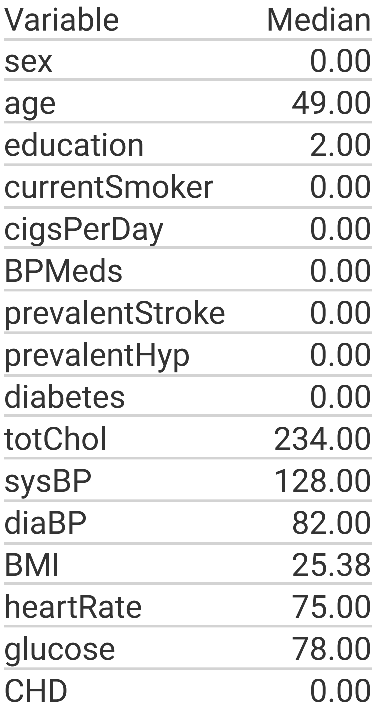
"Linear, Lasso, and Ridge Regression with R", *Plural Sight,* 12 Nov, 2019

Views, R. "Survival Analysis with R." · *R Views*, 25 Sept. 2017, rviews.rstudio.com/2017/09/25/survival-analysis-with-r/.

10 Appendix

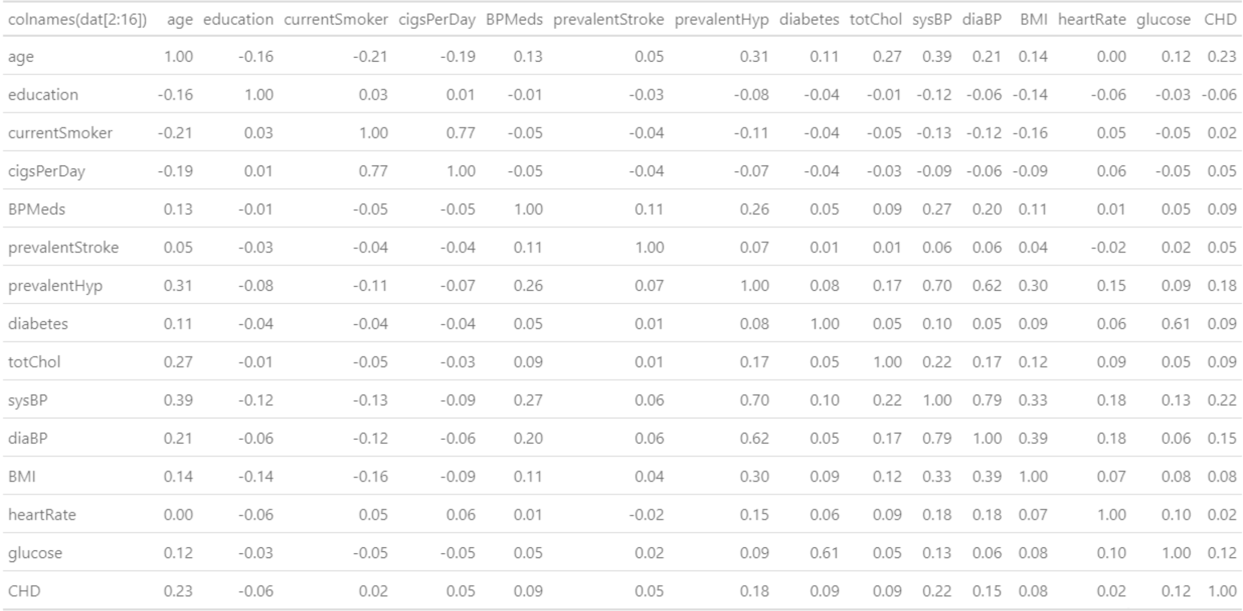
1 Data Analysis

A. Medians of Predictor Variables



Variable	Median
sex	0.00
age	49.00
education	2.00
currentSmoker	0.00
cigsPerDay	0.00
BPMeds	0.00
prevalentStroke	0.00
prevalentHyp	0.00
diabetes	0.00
totChol	234.00
sysBP	128.00
diaBP	82.00
BMI	25.38
heartRate	75.00
glucose	78.00
CHD	0.00

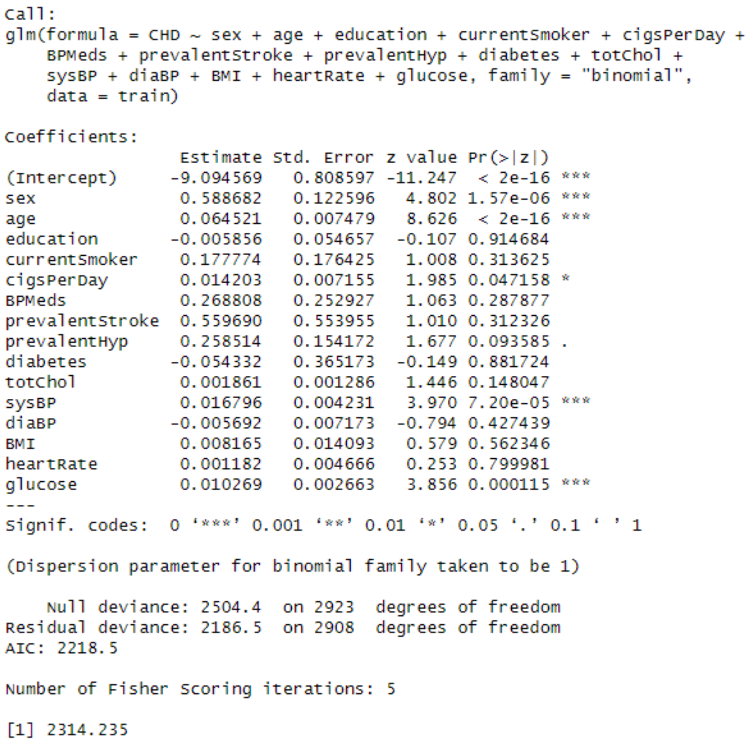
B. Correlation Matrix for full dataset



colnames(dat[2:16])	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	CHD
age	1.00	-0.16	-0.21	-0.19	0.13	0.05	0.31	0.11	0.27	0.39	0.21	0.14	0.00	0.12	0.23
education	-0.16	1.00	0.03	0.01	-0.01	-0.03	-0.08	-0.04	-0.01	-0.12	-0.06	-0.14	-0.06	-0.03	-0.06
currentSmoker	-0.21	0.03	1.00	0.77	-0.05	-0.04	-0.11	-0.04	-0.05	-0.13	-0.12	-0.16	0.05	-0.05	0.02
cigsPerDay	-0.19	0.01	0.77	1.00	-0.05	-0.04	-0.07	-0.04	-0.03	-0.09	-0.06	-0.09	0.06	-0.05	0.05
BPMeds	0.13	-0.01	-0.05	-0.05	1.00	0.11	0.26	0.05	0.09	0.27	0.20	0.11	0.01	0.05	0.09
prevalentStroke	0.05	-0.03	-0.04	-0.04	0.11	1.00	0.07	0.01	0.01	0.06	0.06	0.04	-0.02	0.02	0.05
prevalentHyp	0.31	-0.08	-0.11	-0.07	0.26	0.07	1.00	0.08	0.17	0.70	0.62	0.30	0.15	0.09	0.18
diabetes	0.11	-0.04	-0.04	-0.04	0.05	0.01	0.08	1.00	0.05	0.10	0.05	0.09	0.06	0.61	0.09
totChol	0.27	-0.01	-0.05	-0.03	0.09	0.01	0.17	0.05	1.00	0.22	0.17	0.12	0.09	0.05	0.09
sysBP	0.39	-0.12	-0.13	-0.09	0.27	0.06	0.70	0.10	0.22	1.00	0.79	0.33	0.18	0.13	0.22
diaBP	0.21	-0.06	-0.12	-0.06	0.20	0.06	0.62	0.05	0.17	0.79	1.00	0.39	0.18	0.06	0.15
BMI	0.14	-0.14	-0.16	-0.09	0.11	0.04	0.30	0.09	0.12	0.33	0.39	1.00	0.07	0.08	0.08
heartRate	0.00	-0.06	0.05	0.06	0.01	-0.02	0.15	0.06	0.09	0.18	0.18	0.07	1.00	0.10	0.02
glucose	0.12	-0.03	-0.05	-0.05	0.05	0.02	0.09	0.61	0.05	0.13	0.06	0.08	0.10	1.00	0.12
CHD	0.23	-0.06	0.02	0.05	0.09	0.05	0.18	0.09	0.09	0.22	0.15	0.08	0.02	0.12	1.00

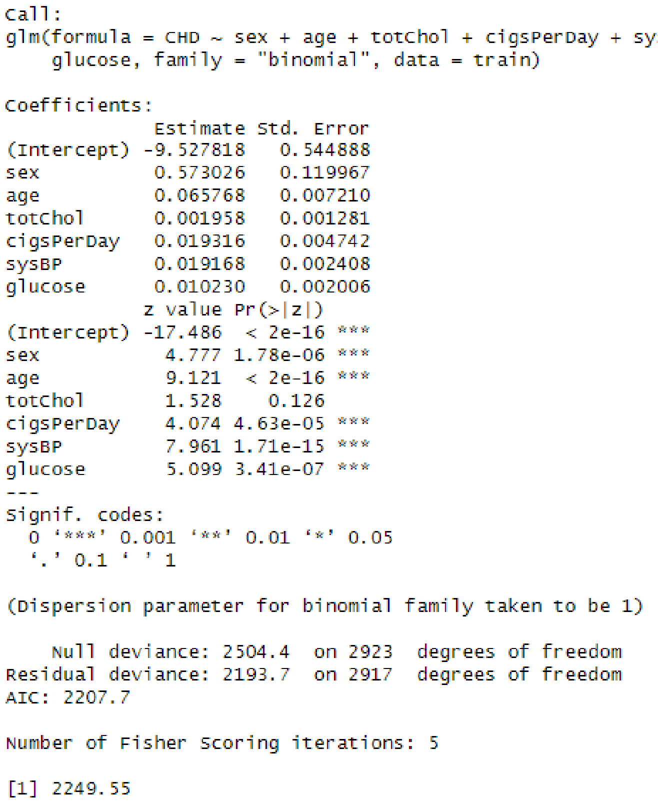
2 Model

A. Summary of full model



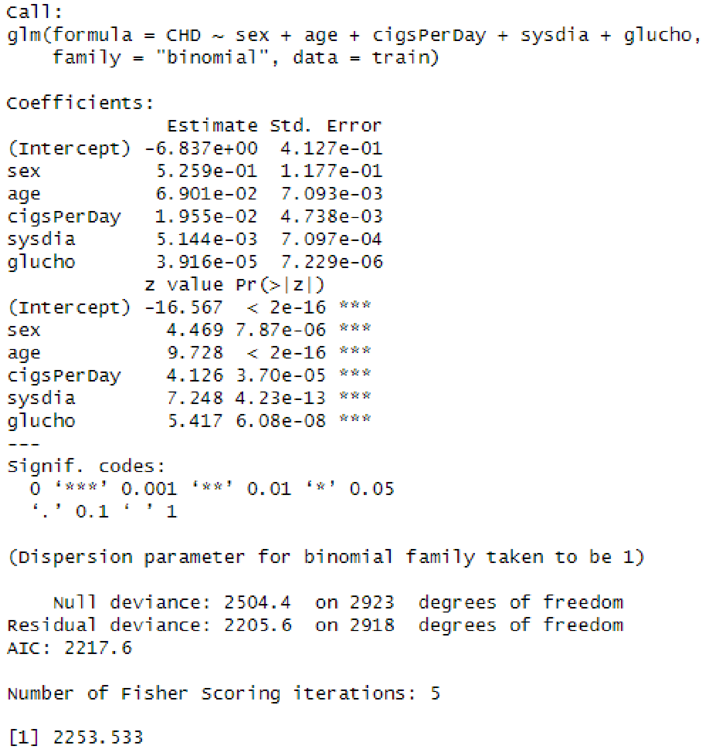
```
Call:
glm(formula = CHD ~ sex + age + education + currentSmoker + cigsPerDay + 
    BPMeds + prevalentStroke + prevalentHyp + diabetes + totChol + 
    sysBP + diaBP + BMI + heartRate + glucose, family = "binomial", 
    data = train)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -9.094569   0.808597 -11.247  < 2e-16 ***
sex              0.588682   0.122596   4.802 1.57e-06 ***
age              0.064521   0.007479   8.626  < 2e-16 ***
education       -0.005856   0.054657  -0.107 0.914684    
currentSmoker    0.177774   0.176425   1.008 0.313625    
cigsPerDay       0.014203   0.007155   1.985 0.047158 *  
BPMeds           0.268808   0.252927   1.063 0.287877    
prevalentStroke  0.559690   0.553955   1.010 0.312326    
prevalentHyp     0.258514   0.154172   1.677 0.093585 .  
diabetes        -0.054332   0.365173  -0.149 0.881724    
totChol          0.001861   0.001286   1.446 0.148047    
sysBP            0.016796   0.004231   3.970 7.20e-05 ***
diaBP           -0.005692   0.007173  -0.794 0.427439    
BMI              0.008165   0.014093   0.579 0.562346    
heartRate        0.001182   0.004666   0.253 0.799981    
glucose          0.010269   0.002663   3.856 0.000115 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2504.4  on 2923  degrees of freedom
Residual deviance: 2186.5  on 2908  degrees of freedom
AIC: 2218.5

Number of Fisher Scoring iterations: 5

[1] 2314.235
```

B. Summary of reduced model



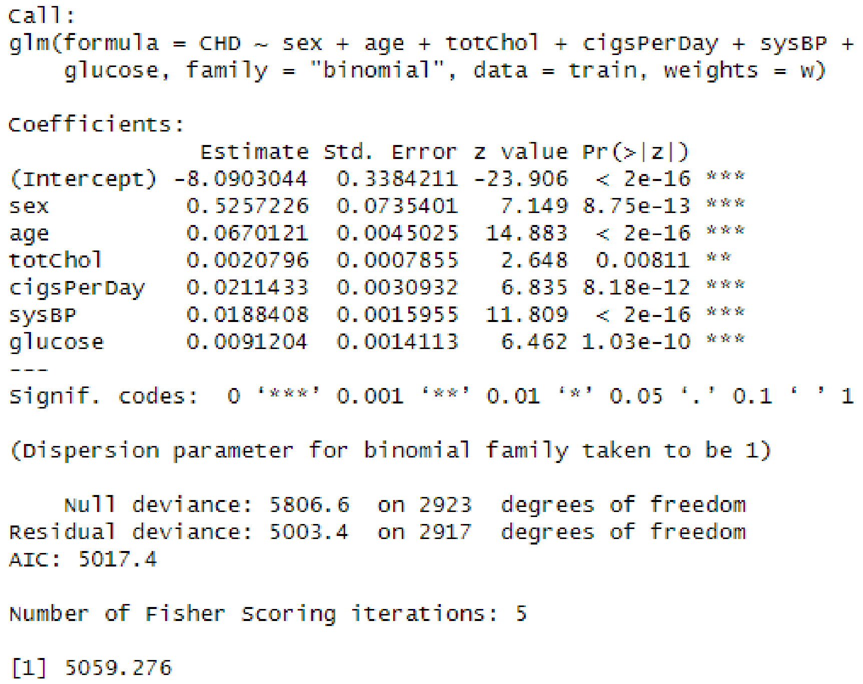
```
Call:
glm(formula = CHD ~ sex + age + totChol + cigsPerDay + sysBP + 
    glucose, family = "binomial", data = train)

Coefficients:
             Estimate Std. Error
(Intercept) -9.527818   0.544888
sex          0.573026   0.119967
age          0.065768   0.007210
totChol      0.001958   0.001281
cigsPerDay   0.019316   0.004742
sysBP        0.019168   0.002408
glucose      0.010230   0.002006
            z value Pr(>|z|)    
(Intercept) -17.486  < 2e-16 ***
sex           4.777 1.78e-06 ***
age           9.121  < 2e-16 ***
totChol       1.528    0.126    
cigsPerDay    4.074 4.63e-05 ***
sysBP         7.961 1.71e-15 ***
glucose       5.099 3.41e-07 ***
---
Signif. codes:  
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2504.4  on 2923  degrees of freedom
Residual deviance: 2193.7  on 2917  degrees of freedom
AIC: 2207.7

Number of Fisher Scoring iterations: 5

[1] 2249.55
```

C. Summary of reduced model with interaction



```
Call:
glm(formula = CHD ~ sex + age + cigsPerDay + sysdia + glucho,
    family = "binomial", data = train)

Coefficients:
              Estimate Std. Error
(Intercept) -6.837e+00  4.127e-01
sex          5.259e-01  1.177e-01
age          6.901e-02  7.093e-03
cigsPerDay   1.955e-02  4.738e-03
sysdia       5.144e-03  7.097e-04
glucho       3.916e-05  7.229e-06
            z value Pr(>|z|)
(Intercept) -16.567  < 2e-16 ***
sex           4.469 7.87e-06 ***
age           9.728  < 2e-16 ***
cigsPerDay    4.126 3.70e-05 ***
sysdia        7.248 4.23e-13 ***
glucho        5.417 6.08e-08 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05
  '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2504.4  on 2923  degrees of freedom
Residual deviance: 2205.6  on 2918  degrees of freedom
AIC: 2217.6

Number of Fisher Scoring iterations: 5

[1] 2253.533
```

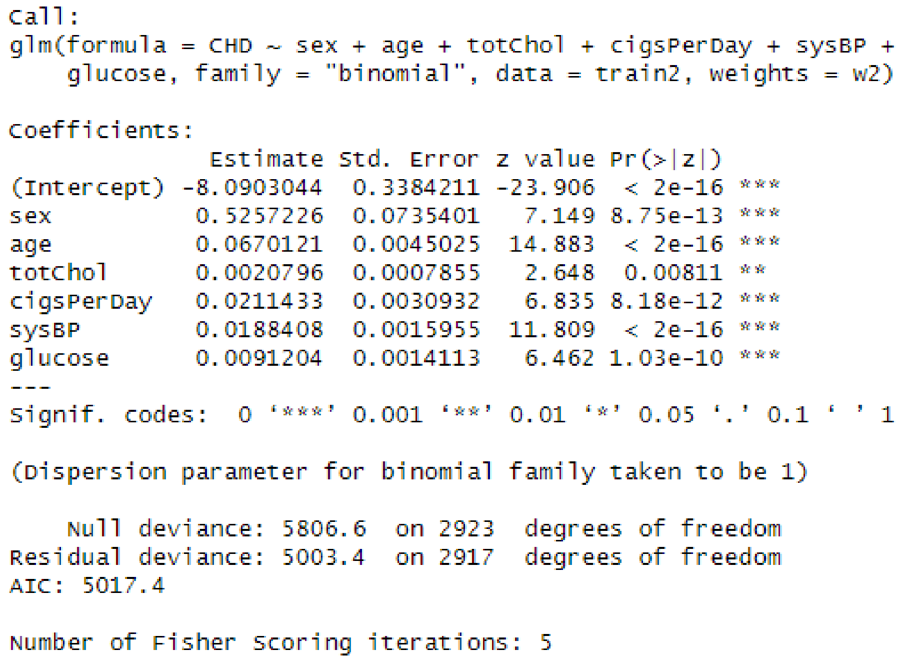
D. Summary of reduced model with weights



```
Call:
glm(formula = CHD ~ sex + age + totChol + cigsPerDay + sysBP +
    glucose, family = "binomial", data = train, weights = w)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.0903044  0.3384211 -23.906  < 2e-16 ***
sex          0.5257226  0.0735401   7.149 8.75e-13 ***
age          0.0670121  0.0045025  14.883  < 2e-16 ***
totChol      0.0020796  0.0007855   2.648  0.00811 **
cigsPerDay   0.0211433  0.0030932   6.835 8.18e-12 ***
sysBP        0.0188408  0.0015955  11.809  < 2e-16 ***
glucose      0.0091204  0.0014113   6.462 1.03e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5806.6  on 2923  degrees of freedom
Residual deviance: 5003.4  on 2917  degrees of freedom
AIC: 5017.4

Number of Fisher Scoring iterations: 5

[1] 5059.276
```

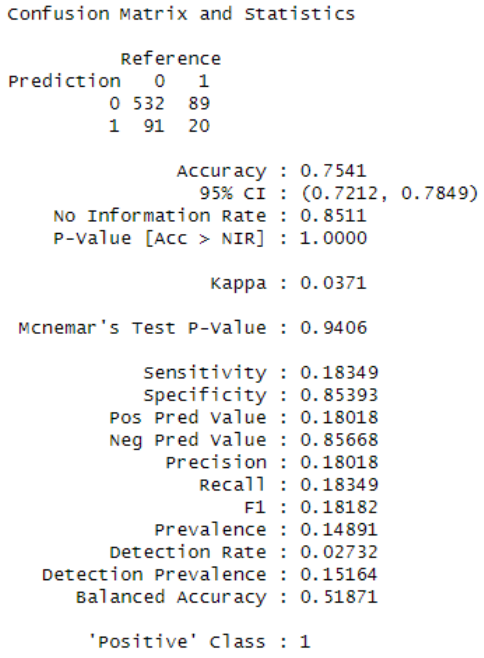
3 Residual Diagnostics

A. Summary of weighted model on training data with outliers removed



```
Call:
glm(formula = CHD ~ sex + age + totChol + cigsPerDay + sysBP + 
    glucose, family = "binomial", data = train2, weights = w2)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -8.0903044  0.3384211 -23.906  < 2e-16 ***
sex          0.5257226  0.0735401   7.149 8.75e-13 ***
age          0.0670121  0.0045025  14.883  < 2e-16 ***
totChol      0.0020796  0.0007855   2.648  0.00811 ** 
cigsPerDay   0.0211433  0.0030932   6.835 8.18e-12 ***
sysBP        0.0188408  0.0015955  11.809  < 2e-16 ***
glucose      0.0091204  0.0014113   6.462 1.03e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5806.6  on 2923  degrees of freedom
Residual deviance: 5003.4  on 2917  degrees of freedom
AIC: 5017.4

Number of Fisher Scoring iterations: 5
```

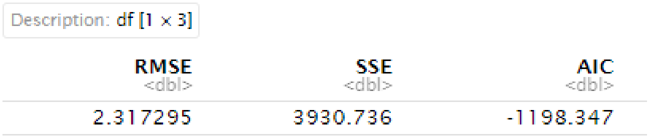
B. Confusion matrix and evaluation metrics for training data with outliers removed



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 532  89
         1  91  20
                                          
               Accuracy : 0.7541          
                 95% CI : (0.7212, 0.7849)
    No Information Rate : 0.8511          
    P-Value [Acc > NIR] : 1.0000          
                                          
                  Kappa : 0.0371          
                                          
 Mcnemar's Test P-Value : 0.9406          
                                          
            Sensitivity : 0.18349         
            Specificity : 0.85393         
         Pos Pred Value : 0.18018         
         Neg Pred Value : 0.85668         
              Precision : 0.18018         
                 Recall : 0.18349         
                     F1 : 0.18182         
             Prevalence : 0.14891         
         Detection Rate : 0.02732         
   Detection Prevalence : 0.15164         
      Balanced Accuracy : 0.51871         
                                          
       'Positive' Class : 1               
```

4 Ridge Regression

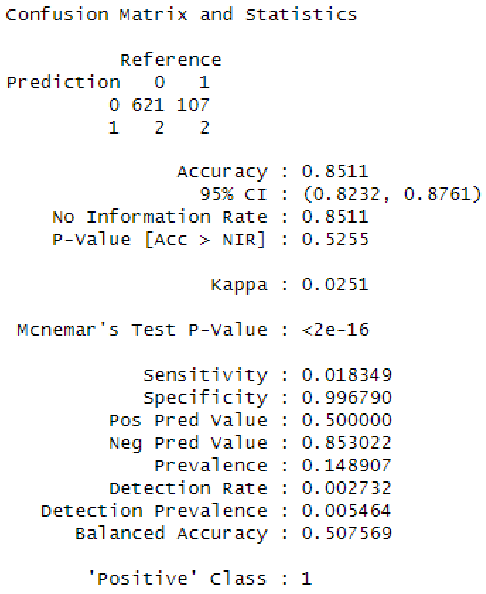
A Summary of ridge model



Description: df [1 × 3]

RMSE <dbl>	SSE <dbl>	AIC <dbl>
2.317295	3930.736	-1198.347

B. Confusion matrix and evaluation metrics for training data



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 621 107
         1   2   2

               Accuracy : 0.8511
                 95% CI : (0.8232, 0.8761)
    No Information Rate : 0.8511
    P-Value [Acc > NIR] : 0.5255

                  Kappa : 0.0251

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.018349
            Specificity : 0.996790
         Pos Pred Value : 0.500000
         Neg Pred Value : 0.853022
             Prevalence : 0.148907
         Detection Rate : 0.002732
   Detection Prevalence : 0.005464
      Balanced Accuracy : 0.507569

       'Positive' Class : 1
```