Predicting Coronary Heart Disease Using A Generalized Logistic Linear Regression Model

Dongcheng Lu, Srinanda Kishore Yallapragada, Vansh Garg, Wacil Voltaire, Ziqian Zhang

April 7 2024

Introduction

The logistic regression model is used when we want to predict the particular class of some object based on some combination of predictor variables. Through the use of a logistic regression model, we can model some binary outcome as either 0 or 1, which represent outcomes like yes or no, success or failure, fatal or non-fatal. This type of regression belongs to a larger family of models known as Generalized Linear Regression Models, which further apply and develop the properties of linear regression models. Unlike linear regression models which return direct outcome variables, logistic regression models instead return the probability of the variable belonging to some category. We can pick and choose our desired level of tolerance for the probability to change categories. Our group is trying to use a logistic regression model to examine the outcome of coronary heart disease based on some combination of predictor variables. Through this analysis, we aim to establish a predictive relationship between a patient's characteristics and the probability of them developing coronary heart disease in the next ten years, which can be used by both doctors to flag incoming patients with the possibility of developing it as well as the general public for monitoring their health metrics under a certain standard to prevent developing it.

Formulas and Basics

The standard logistic regression model has the following formula: $p = \exp(y) / (1 + \exp(y))$ or simply $p = 1/(1 + \exp(-y))$. In this equation y is equal to $b_0 + b_1 * x$, exp() stands for the e exponential, and p represents the probability of the event occurring given some predictor variable x. Graphically the function has an "S" shaped curve. Through the use of logarithmic manipulation the formula can be written as $\log(p/(1-p)) = b_0 + b_1 * x1$ for a single predictor variable or as $\log(p/(1-p)) = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$ for multiple predictor variables. In these equations b_0 represents the intercept coefficient while $b_1 \ldots b_n$ represent increasing/decreasing x values associated with p depending on the sign of the beta coefficient. We call $\log(p/(1-p))$ the logarithm of the odd, or logit, function. In the function the odds represent the likelihood of some event occurring, and this can be interpreted as the ratio of "success" to "failure". From the odds we can

calculate probability as p = odds/(1 + odds). Our group is interested in creating a logistic regression model which examines the odds of coronary heart disease based on some predictor variables.

Loading Required R Packages

We begin by loading the required R packages for the creation of the generalized logistic linear regression model

```
library(ggplot2)
library(dplyr)
library(tidyverse)
library(broom)
library(caret)
library(bestglm)
```

Data Description

The dataset comprises different features and forms of expenditures as variables related to the weekly sales of a certain product at an unnamed company. It consists of seven input variables that represent these forms of weekly expenditures. And one output variable that follows the variation in sales for that week based on the other variables.

1. TenYearCHD (Coronary Heart Disease in 10 years) (binary variable - 1=yes, 0=no)
2. Male (Gender) (binary variable - 1=yes, 0=no)
3. Education (Level of education) (categorical variable - 1 to 4 represent level of education)
4. CurrentSmoker (binary variable - 1=yes, 0=no)
5. CigsPerDay (Cigarettes consumed daily) (continuous variable)
6. BPMeds (Whether or not BP medication prescribed) (binary variable - 1=yes, 0=no)
7. PrevalentStroke (Whether stroke is prevalent in family) (binary variable - 1=yes, 0=no)
8. PrevalentHyp (Whether hypertension is prevalent in family) (binary variable - 1=yes, 0=no)
9. Diabetes (Whether diabetes is prevalent in family) (binary variable - 1=yes, 0=no)
10. TotChol (Total cholesterol level) (continuous variable)
11. SysBP (Systolic blood pressure) (continuous variable)
12. DiaBP (Diastolic blood pressure) (continuous variable)
13. BMI (Body mass index) (continuous variable)

14. HeartRate (continuous variable)
15. Glucose (continuous variable)

The primary focus of our analysis lies in exploring the relationship between TenYearCHD and all other variables. TenYearCHD is the dependent variable (y) in our regression model, while the other variables are the independent variables (x's). Our aim is to create a model to accurately predict the probability of having Coronary Heart Disease in the next 10 years using these parameters,

```
#Loading data
data <- read.csv("heart_disease.csv")
```

Data Preparation

The Heart Disease dataset we have has 4238 entries. We will split this dataset into training data and testing data. The first thing we do is omit the NA entries from our dataset to streamline it

```
#remove NAs
data <- na.omit(data)
```

We see that our dataset now has 3656 entries, which we will further split into a training set and a testing set with an 80% to 20% ratio.

```
#splitting data into training and testing sets
set.seed(456)
training.samples <- data$TenYearCHD %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data <- data[training.samples, ]
test.data <- data[-training.samples, ]
```

Our training dataset has 2925 entries, and our testing dataset has 731 entries. These are the samples we will use to train and test all our models.

Preliminary Model

We start by creating a preliminary model using all 15 of our independent variables.

```
#Fit a model with all predictors
model.full <- glm( TenYearCHD ~., data = train.data, family = binomial)
```

```
summary(model.full)

Call:
glm(formula = TenYearCHD ~ ., family = binomial, data = train.data)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.849130   0.806116  -9.737  < 2e-16 ***
male             0.691735   0.124265   5.567 2.60e-08 ***
age              0.060600   0.007578   7.997 1.28e-15 ***
education       -0.040350   0.055882  -0.722   0.4703
currentSmoker    0.010293   0.177074   0.058   0.9536
cigsPerDay       0.017198   0.007099   2.423   0.0154 *
BPMeds           0.352775   0.253834   1.390   0.1646
prevalentStroke  1.149583   0.538366   2.135   0.0327 *
prevalentHyp     0.233021   0.156903   1.485   0.1375
diabetes        -0.059049   0.370051  -0.160   0.8732
totChol          0.003189   0.001274   2.503   0.0123 *
sysBP            0.017295   0.004324   4.000 6.35e-05 ***
diaBP           -0.011328   0.007288  -1.554   0.1201
BMI              0.002369   0.014330   0.165   0.8687
heartRate       -0.003768   0.004833  -0.780   0.4356
glucose          0.005448   0.002648   2.058   0.0396 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2438.8  on 2924  degrees of freedom
Residual deviance: 2154.9  on 2909  degrees of freedom
AIC: 2186.9
```

Best GLM Model using Subset Selection

We will now create a new model using subset selection from the bestglm() library. Note that since we have 15 parameters, choosing the best set of parameters can take about 2-3 minutes to compute.

```
# do best subset selection
bestglm(train.data,family=binomial,IC="AIC",method="exhaustive")

Morgan-Tatar search since family is non-gaussian.
AIC
```

```
BICq equivalent for q in (0.948423679956362, 0.952731263793734)
Best Model:
                    Estimate  Std. Error    z value     Pr(>|z|)
(Intercept)     -8.214011055 0.661168399 -12.423478 1.949004e-35
male             0.693822947 0.122934377   5.643848 1.662908e-08
age              0.061670607 0.007484794   8.239453 1.729998e-16
cigsPerDay       0.017045887 0.004768259   3.574866 3.504077e-04
prevalentStroke  1.229941101 0.531413944   2.314469 2.064201e-02
prevalentHyp     0.246227872 0.155443975   1.584030 1.131870e-01
totChol          0.003128693 0.001266486   2.470373 1.349722e-02
sysBP            0.018033780 0.004273018   4.220385 2.438859e-05
diaBP           -0.011743610 0.007126299  -1.647926 9.936796e-02
glucose          0.005088832 0.001994629   2.551267 1.073320e-02
```

Based on the best selected parameters, we create our new "best" model.

```
#Fit the best fit model using male, age, cigsPerDay, prevalentHyp,
totChol, sysBP, glucose
model.best <- glm( TenYearCHD ~
male+age+cigsPerDay+prevalentHyp+totChol+sysBP+glucose, data=train.data,
family=binomial)
summary(model.best)

Call:
glm(formula = TenYearCHD ~ male + age + cigsPerDay + prevalentHyp +
    totChol + sysBP + glucose, family = binomial, data = train.data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.747538   0.588804 -14.856  < 2e-16 ***
male          0.664693   0.121371   5.477 4.34e-08 ***
age           0.064845   0.007283   8.904  < 2e-16 ***
cigsPerDay    0.017134   0.004758   3.601 0.000317 ***
prevalentHyp  0.219993   0.152955   1.438 0.150354
totChol       0.003106   0.001264   2.458 0.013987 *
sysBP         0.013620   0.003244   4.199 2.68e-05 ***
glucose       0.005331   0.001980   2.692 0.007094 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 2438.8  on 2924  degrees of freedom
Residual deviance: 2165.7  on 2917  degrees of freedom
AIC: 2181.7
```

Simple GLM Model

Finally, for a clean comparison, we create a simple GLM model using only 1 parameter. We select the best single parameter using the same method.

```
# finding a variable for simple glm model
bestglm(train.data,family=binomial,IC="AIC",method="exhaustive",nvmax = 1)

Morgan-Tatar search since family is non-gaussian.
AIC
Best Model:
               Estimate  Std. Error   z value     Pr(>|z|)
(Intercept) -5.73563492 0.350921524 -16.34449 4.761028e-60
age          0.07702624 0.006487732  11.87260 1.642836e-32
```

Using this, we create our simple GLM model

```
#fitting the simple glm model
model.simple <- glm( TenYearCHD ~ male+age, data=train.data,
family=binomial)
summary(model.simple)

Call:
glm(formula = TenYearCHD ~ male + age, family = binomial, data =
train.data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.121727   0.361566 -16.931  < 2e-16 ***
male         0.647299   0.108862   5.946 2.75e-09 ***
age          0.078316   0.006536  11.982  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2438.8  on 2924  degrees of freedom
Residual deviance: 2250.9  on 2922  degrees of freedom
AIC: 2256.9
```

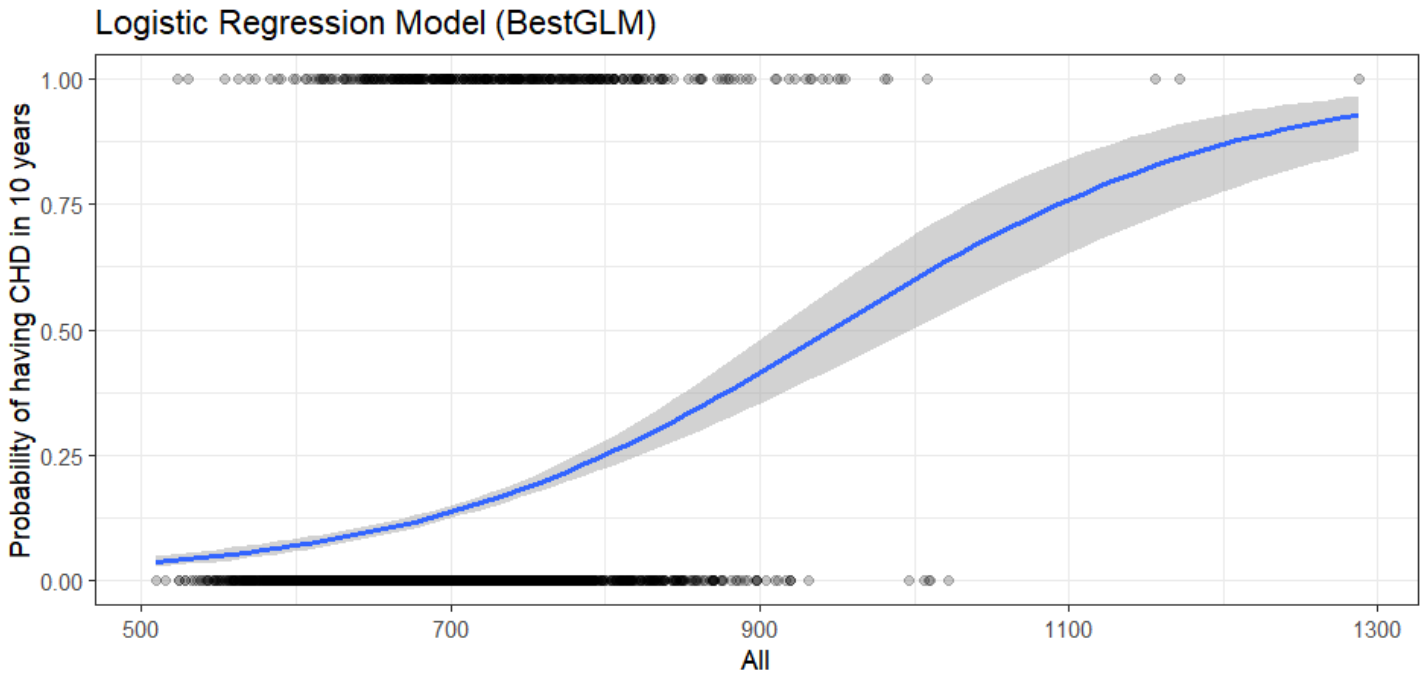
Analysis

Pictures and Tables

Let us visualize the distributions of all three of our models and see if they form the characteristic “S” shaped curve associated with the generalized logistic linear regression function.

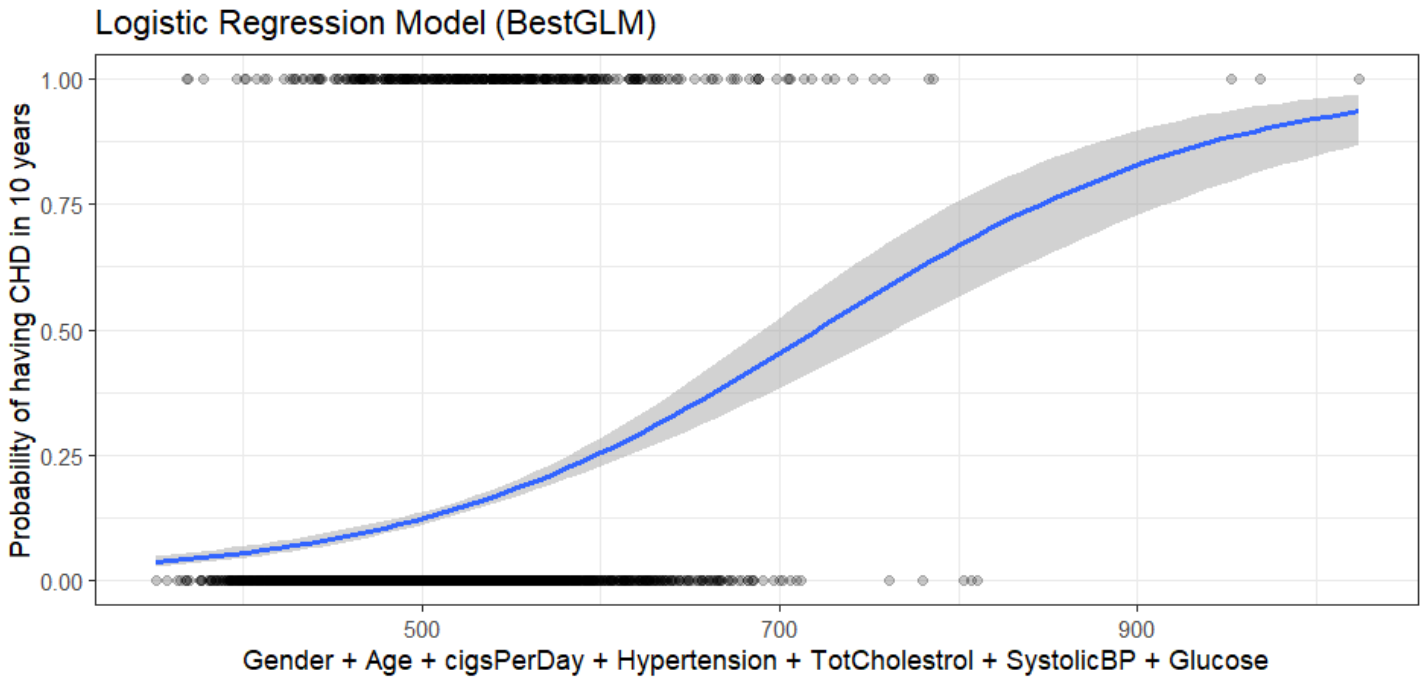
For the model with all parameters:

```
train.data %>%
  mutate(prob = ifelse(TenYearCHD == 1, 1, 0)) %>%
ggplot(aes(male+age+cigsPerDay+prevalentHyp+totChol+sysBP+glucose+educatio
n+currentSmoker+BPMeds+prevalentStroke+diabetes+diaBP+BMI+heartRate,
prob)) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +
  labs(
    title = "Logistic Regression Model (BestGLM)",
    x = "All",
    y = "Probability of having CHD in 10 years"
  )
```



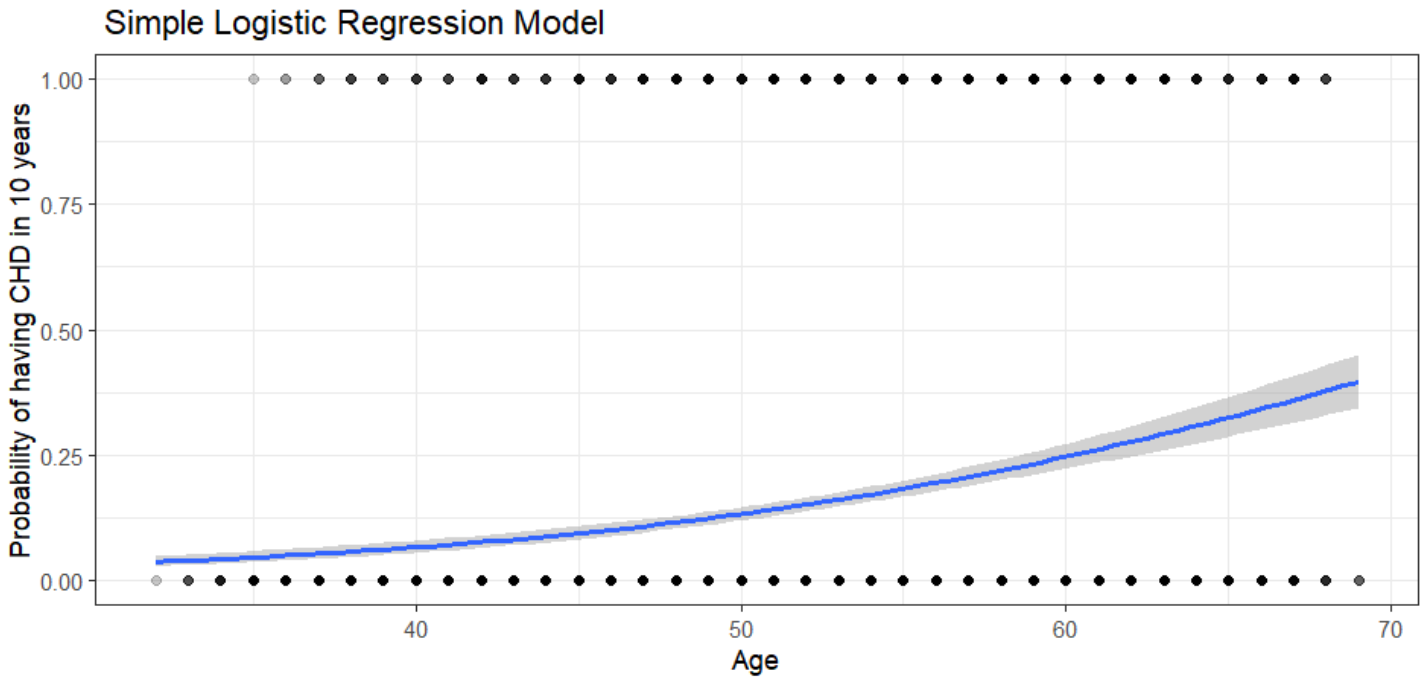
For the model with our best subset selected parameters:

```
train.data %>%
  mutate(prob = ifelse(TenYearCHD == 1, 1, 0)) %>%
  ggplot(aes(male+age+cigsPerDay+prevalentHyp+totChol+sysBP+glucose,
prob)) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +
  labs(
    title = "Logistic Regression Model (BestGLM)",
    x = "Gender + Age + cigsPerDay + Hypertension + TotCholestrol +
SystolicBP + Glucose",
    y = "Probability of having CHD in 10 years"
  )
```



For the simple GLM model using only 1 parameter:

```
train.data %>%
  mutate(prob = ifelse(TenYearCHD == 1, 1, 0)) %>%
  ggplot(aes(age, prob)) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +
  labs(
    title = " Simple Logistic Regression Model",
    x = " Age ",
    y = "Probability of having CHD in 10 years"
  )
```



Note that the simple GLM model does not clearly form the characteristic “S” shaped curve.

Computation

Now let us compare all three of our models to decide which is the best one by comparing their prediction accuracy on the test data.

```
#evaluate the full model
probabilities <- model.full %>% predict(test.data, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, 1, 0)
# Model accuracy
mean(predicted.classes == test.data$TenYearCHD)
```

```
[1] 0.8303694
# evaluate the best model
new.probabilities <- model.best %>% predict(test.data, type = "response")
new.predicted.classes <- ifelse(new.probabilities > 0.5, 1, 0)
# Model accuracy
mean(new.predicted.classes == test.data$TenYearCHD)
```

```
[1] 0.8331053
```

```
# evaluate the simple model
simple.probabilities <- model.simple %>% predict(test.data, type =
"response")
simple.predicted.classes <- ifelse(simple.probabilities > 0.5, 1, 0)
# Model accuracy
mean(simple.predicted.classes == test.data$TenYearCHD)

[1] 0.8248974
```

Thus, we see that our "best" model is in fact the best in terms of prediction accuracy, better than both the simple GLM model and the model using all parameters. Using this model, our logistic equation comes out to be:
p = exp(-8.747538 + 0.664693 * male + 0.064845 * age + 0.017134 * cigsPerDay + 0.219993 * prevalentHyp + 0.003106 * totChol + 0.013620 * sysBP + 0.005331 * glucose) / (1 + exp(-8.747538 + 0.664693 * male + 0.064845 * age + 0.017134 * cigsPerDay + 0.219993 * prevalentHyp + 0.003106 * totChol + 0.013620 * sysBP + 0.005331 * glucose)).

Interpretation

The multiple logistic regression model includes predictor variables such as age, systolic blood pressure, cigarettes per day, glucose level, and gender that can be used to predict the probability of 10 year risk of coronary heart disease. From our "best" multiple logistic regression model, we see that the coefficient estimate for: gender is 0.664693, age is 0.064845, cigsPerDay is 0.017134, prevalentHyp is 0.219993, totChol is 0.003106, sysBP is 0.013620, and glucose is 0.005331. All these beta coefficients have a positive value. This means that for an increase in the predictor variable there is an associated increase in the probability of having a ten year risk of coronary heart disease.

The odds ratio helps measure the association between some predictor variable (x) and the outcome variable (y). This ratio is the odds of an event occurring given some predictor variable is present.
For the multiple logistic regression model we can also interpret the odds ratio for each beta coefficient.

- For one unit change in gender there will be an increase in the odds of having a ten year risk of coronary heart disease by exp(0.664693) or 1.94 times
- For one unit increase in age there will be an increase in the odds of having a ten year risk of coronary heart disease by exp(0.064845) or 1.07 times.
- For one unit increase in cigsPerDay there will be an increase in the odds of having a ten year risk of coronary heart disease by exp(0.017134) or 1.02 times.

- For one unit increase in prevalentHyp there will be an increase in the odds of having a ten year risk of coronary heart disease by exp(0.219993) or 1.25 times.
- For one unit increase in totChol there will be an increase in the odds of having a ten year risk of coronary heart disease by exp(0.003106) or 1.00 times.
- For one unit increase in sysBP there will be an increase in the odds of having a ten year risk of coronary heart disease by exp(0.013620) or 1.01 times.
- For one unit increase in glucose there will be an increase in the odds of having a ten year risk of coronary heart disease by exp(0.005331) or 1.01 times.

Model Evaluation

Model Summary

We can now evaluate whether or not the multiple logistic regression model we have defined is actually statistically significant.

```
summary(model.best)

Call:
glm(formula = TenYearCHD ~ male + age + cigsPerDay + prevalentHyp +
    totChol + sysBP + glucose, family = binomial, data = train.data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.747538   0.588804 -14.856  < 2e-16 ***
male          0.664693   0.121371   5.477 4.34e-08 ***
age           0.064845   0.007283   8.904  < 2e-16 ***
cigsPerDay    0.017134   0.004758   3.601 0.000317 ***
prevalentHyp  0.219993   0.152955   1.438 0.150354
totChol       0.003106   0.001264   2.458 0.013987 *
sysBP         0.013620   0.003244   4.199 2.68e-05 ***
glucose       0.005331   0.001980   2.692 0.007094 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2438.8  on 2924  degrees of freedom
```

```
Residual deviance: 2165.7  on 2917  degrees of freedom
AIC: 2181.7
```

The statistics given include the beta coefficients, null deviance, residual deviance, and AIC, which is the Akaike information criterion. All of these measures can help us determine goodness of fit. We can compute the Chi-Square statistic, X^2, to determine the p-value and determine significance. X^2 is defined as below: X^2 = null deviance - residual deviance = 2438.8 - 2165.7 = 273.1 The degree of freedom is 7 since we have 7 predictor values. A Chi-Square value of 273.1 with df = 7 corresponds to a p-value < 0.00001, which is significant because it is less than 0.05. This means that our multiple logistic model fits our data and is statistically significant.

Coefficient Significance

From the above table we can see that the p values for each of our variables are very low and close to 0, indicating that all of our coefficients are statistically significant.

Prediction and Model Accuracy

To see how accurate our model actually is, we must make some predictions.

```
# evaluate the best model
new.probabilities <- model.best %>% predict(test.data, type = "response")
new.predicted.classes <- ifelse(new.probabilities > 0.5, 1, 0)
# Model accuracy
mean(new.predicted.classes == test.data$TenYearCHD)
```

```
[1] 0.8331053
```

The accuracy for predicting with the multiple GLM is 83.31053%. The accuracy for each model, stated as a percentage here, means that the classification prediction is correct at that rate. For the multiple GLM, it misclassified approximately 16.68% of the time. The accuracy rate is relatively high, which indicates a good model.

Conclusion and Summary

In conclusion, our group used different sets of personal characteristics as predictor variables to try and the probability of a person having a ten year risk of coronary heart disease. From those predictor variables we

performed stepwise regression to select the best possible single and multiple variable models and examined the corresponding logistic regression model. As the summary statistics such as Chi-square have shown, the single and multiple logistic regression models appear to be statistically significant. By examining the p-values, we found that at the 0.05 level of significance, each predictor variable in the single or multiple variable models appears to be significantly related to the probability of the outcome variable. From our health data, we found that at the 0.05 level of significance there exist significant relationships between the probability of developing coronary heart disease and the predictor variables age, cigarettes smoked, glucose level, systolic blood pressure, and gender.

References

- Hartmann, K., Krois, J., Waske, B. (2018): E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin.
- Kassambara. (2018, November 3) Stepwise Regression Essentials in R. STHDA. Retrieved April 5, 2023, from Stepwise Regression Essentials in R
- Kassambara. (2018, November 3) Logistic Regression Essentials in R. STHDA. Retrieved April 5, 2023, from http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regressionessentials-in-r/
- World Health Organization (2021). Cardiovascular Disease [Data Set]. https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression/download?datasetVersionNumber=1