ITM 883 Project - Group 5

Addison Golo, Aibek Gubashov, Mounika Yallamandhala, Olayemi Adesina, Suhas Sundar

2023-04-16

Data Set

We chose the Cardiovascular Study Dataset which reflects cardiovascular study on residents of the town of Framingham, Massachusetts.

Data Source: Cardiovascular Study Dataset - Kaggle

Task

Overall goal of project is to apply various descriptive and predictive analytics methods in order to get predict whether the patient have 10-year risk of coronary heart disease (CHD) or not considering the potential risk factors for heart disease.

Purpose

Identifying cardiovascular diseases in their early stages can prove invaluable in advising high-risk patients to adopt healthy lifestyle changes, thereby mitigating the likelihood of complications.

Data description

The Cardiovascular Study dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD).The dataset provides the patients' information containing 4240 records and 15 attributes separated into 3390 train and 848 test data.

Below Data Description table provides more information about each attribute:

Variable	Type	Description
Demographic:		
sex	Categorical	male or female("M" or "F")
age	Continuous	age of the patient
education	Categorical	levels of education (1 to 4)
Behavioral:		
is_smoking	Categorical	whether or not the patient is a current smoker ("YES" or "NO")
cigsPerDay	Continuous	the number of cigarettes that the person smoked on average in one day.

Variable	Type	Description
BPMeds	Categorical (Binary)	whether or not the patient was on blood pressure medication (Nominal).
prevalentStroke	Categorical (Binary)	whether or not the patient had previously had a stroke.
prevalentHyp	Categorical (Binary)	whether or not the patient was hypertensive.
diabetes	Categorical (Binary)	whether or not the patient had diabetes.
Medical:		
totChol	Continuous	total cholesterol level
sysBP	Continuous	systolic blood pressure
diaBP	Continuous	diastolic blood pressure
BMI	Continuous	Body Mass Index
heartRate	Continuous	heart rate
glucose	Continuous	glucose level

Predict variable (TenYearCHD): 10 year risk of coronary heart disease CHD(binary: "1", means "Yes", "0" means "No")

Exploratory Data Analysis

In our data set we have a total of 15 independent variables out of which 7 are categorical and 8 are quantitative.

```
#train and test datasets
train = read.csv("~/Documents/MS-BDSA 2023/ITM 883/Group Project/train.csv")
test = read.csv("~/Documents/MS-BDSA 2023/ITM 883/Group Project/test.csv")

#Summary of dataset
summary(train)
```

```
##        id              age          education         sex
##  Min.   :   0.0   Min.   :32.00   Min.   :1.000   Length:3390
##  1st Qu.: 847.2   1st Qu.:42.00   1st Qu.:1.000   Class :character
##  Median :1694.5   Median :49.00   Median :2.000   Mode  :character
##  Mean   :1694.5   Mean   :49.54   Mean   :1.971
##  3rd Qu.:2541.8   3rd Qu.:56.00   3rd Qu.:3.000
##  Max.   :3389.0   Max.   :70.00   Max.   :4.000
##                                   NA's   :87
##   is_smoking          cigsPerDay        BPMeds        prevalentStroke
##  Length:3390        Min.   : 0.000   Min.   :0.00000   Min.   :0.00000
##  Class :character   1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.00000
##  Mode  :character   Median : 0.000   Median :0.00000   Median :0.00000
##                     Mean   : 9.069   Mean   :0.02989   Mean   :0.00649
##                     3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.00000
##                     Max.   :70.000   Max.   :1.00000   Max.   :1.00000
##                     NA's   :22       NA's   :44
##   prevalentHyp       diabetes          totChol          sysBP
##  Min.   :0.0000   Min.   :0.00000   Min.   :107.0   Min.   : 83.5
```

```
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0
##  Median :0.0000   Median :0.00000   Median :234.0   Median :128.5
##  Mean   :0.3153   Mean   :0.02566   Mean   :237.1   Mean   :132.6
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:264.0   3rd Qu.:144.0
##  Max.   :1.0000   Max.   :1.00000   Max.   :696.0   Max.   :295.0
##                                     NA's   :38
##      diaBP             BMI           heartRate         glucose
##  Min.   : 48.00   Min.   :15.96   Min.   : 45.00   Min.   : 40.00
##  1st Qu.: 74.50   1st Qu.:23.02   1st Qu.: 68.00   1st Qu.: 71.00
##  Median : 82.00   Median :25.38   Median : 75.00   Median : 78.00
##  Mean   : 82.88   Mean   :25.79   Mean   : 75.98   Mean   : 82.09
##  3rd Qu.: 90.00   3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00
##  Max.   :142.50   Max.   :56.80   Max.   :143.00   Max.   :394.00
##                   NA's   :14      NA's   :1        NA's   :304
##    TenYearCHD
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1507
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

Here we get information about common statistical methods applied to our data in order to get overall picture about data set

We can observe that variable Glucose has highest number of missing values followed by Education in the overall dataset but both of them are less than 10% of the total records.

Data Cleaning

We have some missing values in our train and test data set which is shown as below

```
colSums(is.na(train))
```

```
##             id            age      education            sex     is_smoking
##              0              0             87              0              0
##     cigsPerDay         BPMeds prevalentStroke   prevalentHyp       diabetes
##             22             44              0              0              0
##        totChol          sysBP          diaBP            BMI      heartRate
##             38              0              0             14              1
##        glucose     TenYearCHD
##            304              0
```

```
train = na.omit(train)
sum(is.na(train))
```

```
## [1] 0
```

```
colSums(is.na(test))
```

```
##              id             age       education             sex      is_smoking
##               0               0              18               0               0
##      cigsPerDay          BPMeds prevalentStroke    prevalentHyp        diabetes
##               7               9               0               0               0
##         totChol           sysBP           diaBP             BMI       heartRate
##              12               0               0               5               0
##         glucose
##              84
```

```
test = na.omit(test)
sum(is.na(test))
```

```
## [1] 0
```

```
# Removing id column from both train and test data set
train = train[,-1]
test = test[,-1]
dim(train)
```

```
## [1] 2927   16
```

```
dim(test)
```

```
## [1] 729  15
```

Visualizations

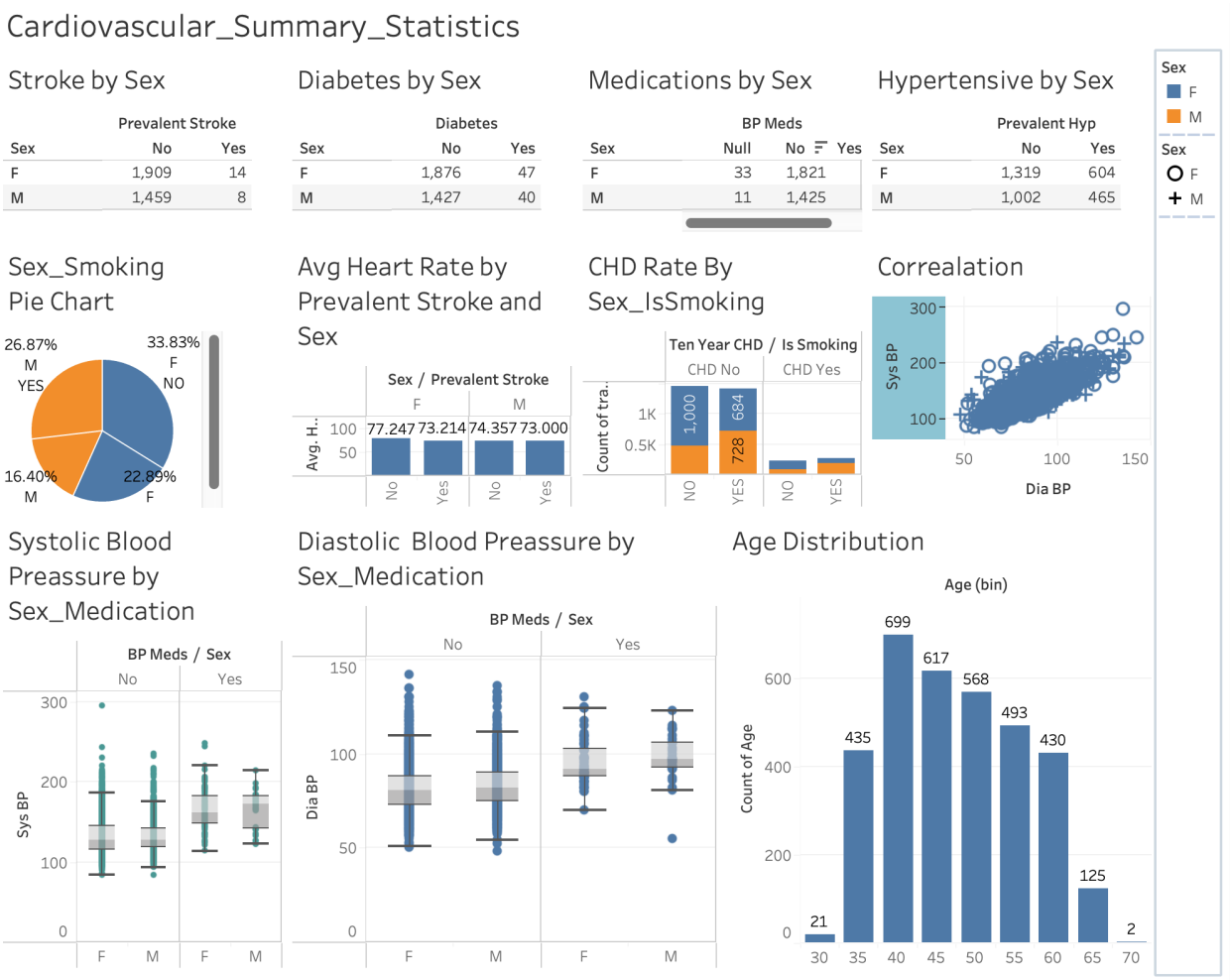


Figure 1: Data Overview (Tableau)

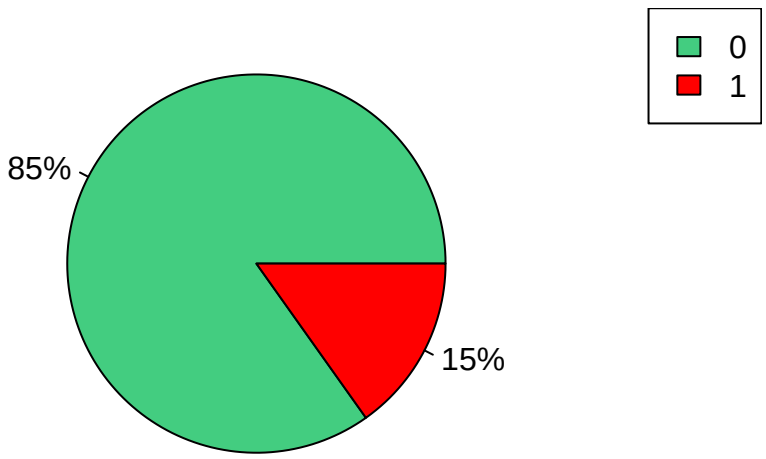
Useful outcomes from descriptive analysis which we take into consideration during our predictive and interpretative analysis:

1. *Age:* Quantitative variable with people age group of 32 to 70 years.
2. *Education:* Categorical variable with number of years of education ranging from "1" to "4" and has 87 missing values.
3. *Cigs Per Day:* Quantitative variable which has values ranging from "0" to "70". Value "0" tells us that these are the people who didn't smoke. Has 22 missing values however these people are indicated as smokers by the is_smoking variable.
4. *BP Meds:* Categorical variable having 44 missing values.
5. *Tot Chol:* Quantitative variable containing 38 missing values.
6. *Sys BP:* There are some potential outliers in Sys BP variable, and we can identify their id in whisker plot. Overall people who are using blood pressure medication (BP Meds) have in average greater systolic blood pressure (Sys BP) than people who are not using blood pressure medication.
7. *Dia BP:* There are some potential outliers in Dia BP variable, and we can identify their id in whisker plot. Overall people who are using blood pressure medication (BP Meds) have in average greater diastolic blood pressure (Dia BP) than people who are not using blood pressure medication

8. *BMI:* This variable contains 14 missing values. Can be converted to categorical variable by binning into groups like
    - Underweight: <= 18.5
    - Normal weight: 18.5–24.9
    - Overweight: 25–29.9
    - Obesity: >= 30
9. *Heartrate:* This variable contains only 1 missing value.
10. *Glucose:* This variable contains 304 missing values.

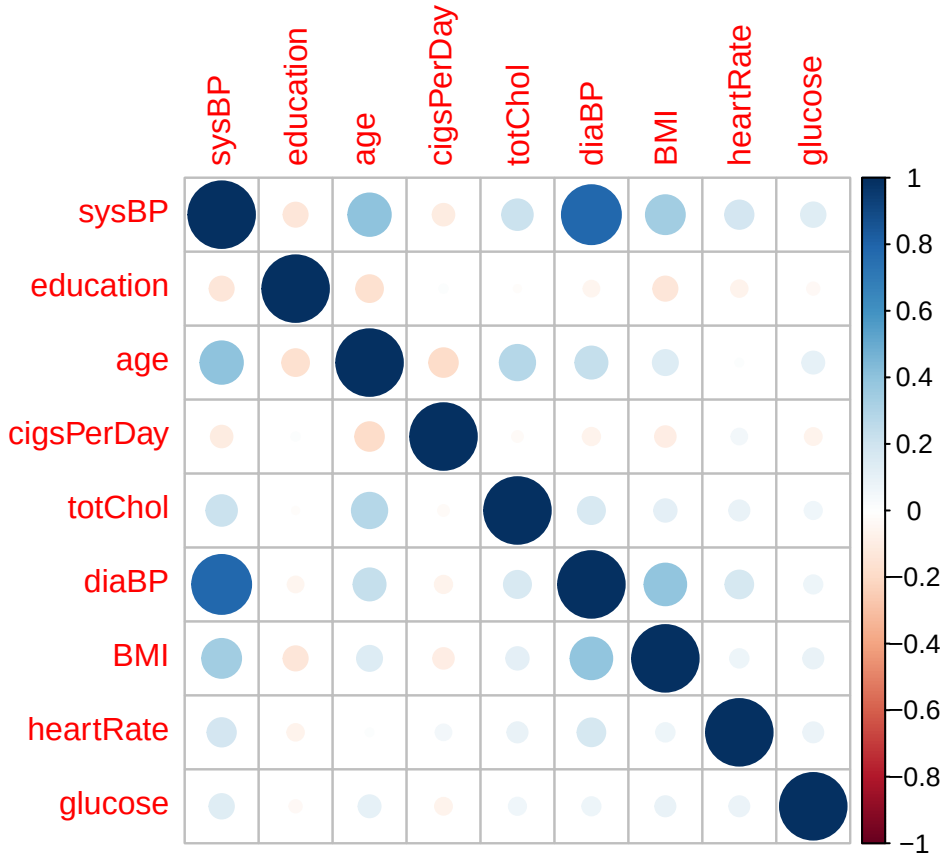
Target variable distribution in train data set

```
#Target variable distribution
freq = table(train$TenYearCHD)
perc = prop.table(freq)
my_colors <- c("seagreen3", "red")
my_labels <- c("0", "1")
pie(perc,labels = paste0(round(perc*100), "%"), col = my_colors, main = "TenYearCHD Binary Variable")
legend("topright", legend = my_labels, fill = my_colors)
```



Correlogram and Correlation Matrix

```r
#Select only quantitative variables for further analysis
train_quan_vars <- subset(train, select=c(sysBP, education, age, cigsPerDay, totChol,
                                          diaBP, BMI, heartRate, glucose))

# Correlogram
corrplot(cor(train_quan_vars, use="pairwise.complete.obs"))
```




```r
# Correlation matrix
round(cor(train_quan_vars, use="pairwise.complete.obs"), 2)
```

```
##             sysBP education   age cigsPerDay totChol diaBP   BMI heartRate
## sysBP        1.00     -0.13  0.41      -0.11    0.21  0.78  0.34      0.18
## education   -0.13      1.00 -0.16       0.02   -0.01 -0.06 -0.13     -0.06
## age          0.41     -0.16  1.00      -0.18    0.28  0.23  0.14      0.01
## cigsPerDay  -0.11      0.02 -0.18       1.00   -0.03 -0.07 -0.10      0.06
## totChol      0.21     -0.01  0.28      -0.03    1.00  0.17  0.11      0.09
## diaBP        0.78     -0.06  0.23      -0.07    0.17  1.00  0.39      0.17
## BMI          0.34     -0.13  0.14      -0.10    0.11  0.39  1.00      0.08
## heartRate    0.18     -0.06  0.01       0.06    0.09  0.17  0.08      1.00
## glucose      0.13     -0.03  0.11      -0.06    0.07  0.07  0.09      0.09
##             glucose
## sysBP          0.13
## education     -0.03
## age            0.11
```

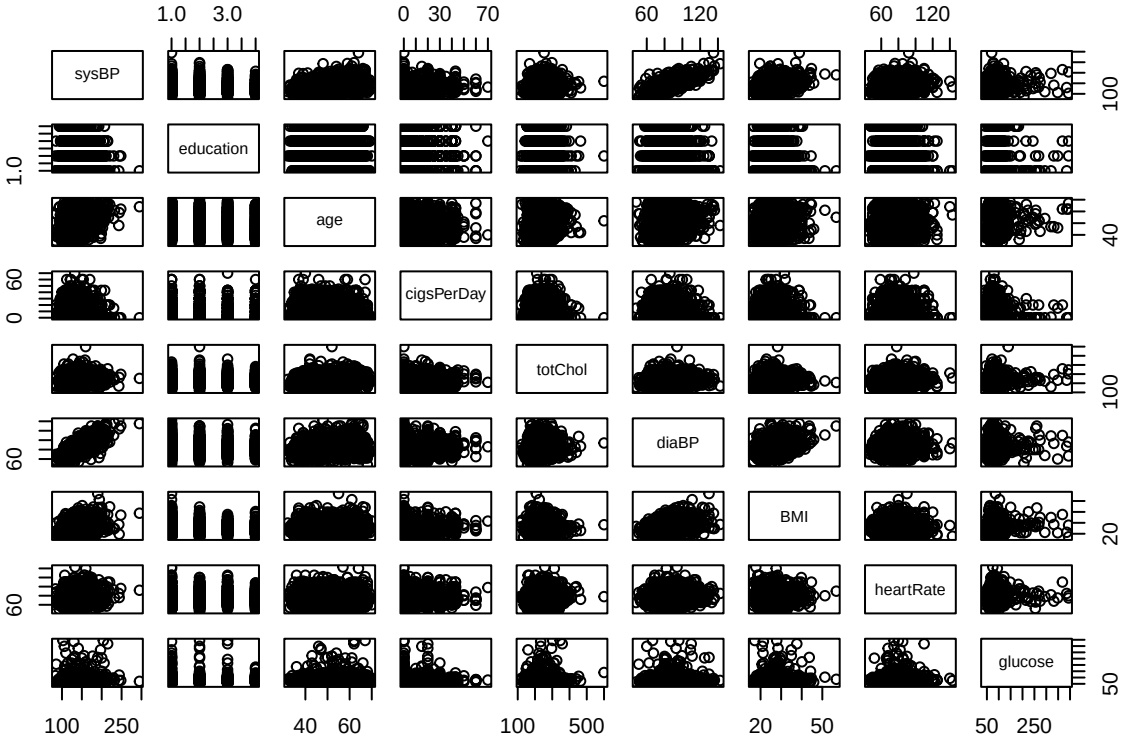
```
## cigsPerDay   -0.06
## totChol       0.07
## diaBP         0.07
## BMI           0.09
## heartRate     0.09
## glucose       1.00
```

Correlogram Interpretation

The correlogram represents the correlations for all pairs of variables. Positive correlations are displayed in blue and negative correlations in red. The intensity of the color is proportional to the correlation coefficient so the stronger the correlation (i.e., the closer to -1 or 1), the darker the boxes. So according to our correlogram we can see that highest positive correlation is between sysBP and diaBP variables. There are some medium positive correlation between sysBP-Age and sysBP-BMI. There is no negative strong correleation that we can take into account.

Scatter Plot Matrix

```
pairs(train_quan_vars)
```

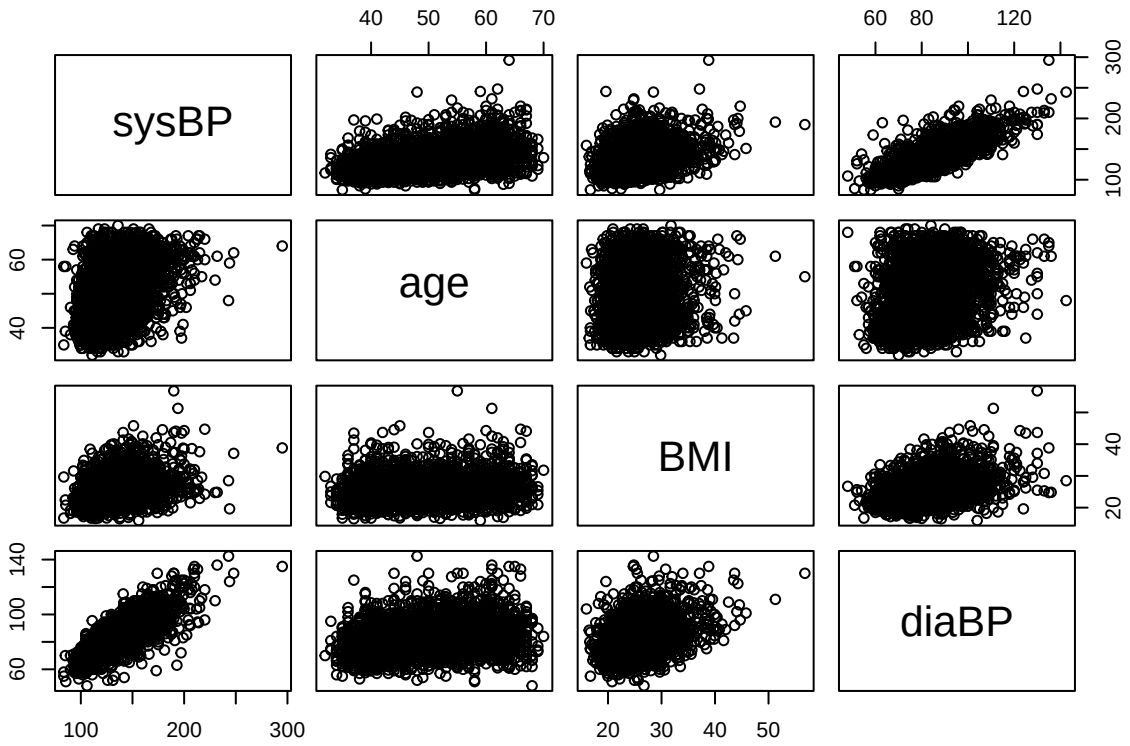


Scatter Plot Matrix Interpretation:

Scatterplot matrices are a great way to roughly determine if you have a linear correlation between multiple variables. In this case, we get a very busy scatterplot matrix. However, using information in previous correlogram we can see that there is some linear relationship between sysBP and diaBP variables

Scatter Plot Matrix with subset of data

```
pairs(~ sysBP+age+BMI+diaBP, data = train_quan_vars)
```



Here we used subset of data in order to get more clear Scatter Plot Matrix and we definitely see some linear relationship between sysBP and diaBP

Comparing proportions between two populations

```
xtabs( ~ is_smoking+TenYearCHD, data=train)
```

```
##           TenYearCHD
## is_smoking    0    1
##        NO  1273  207
##        YES 1210  237
```

```
X <- c(236, 275) #is_smoking No, Yes
TOTAL <- c(1467+236, 1412+275)

prop.test(X, TOTAL)
```

```
##
```

```
##  2-sample test for equality of proportions with continuity correction
##
## data:  X out of TOTAL
## X-squared = 3.7633, df = 1, p-value = 0.05239
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.0491045759  0.0002400073
## sample estimates:
##    prop 1    prop 2
## 0.1385790 0.1630113
```

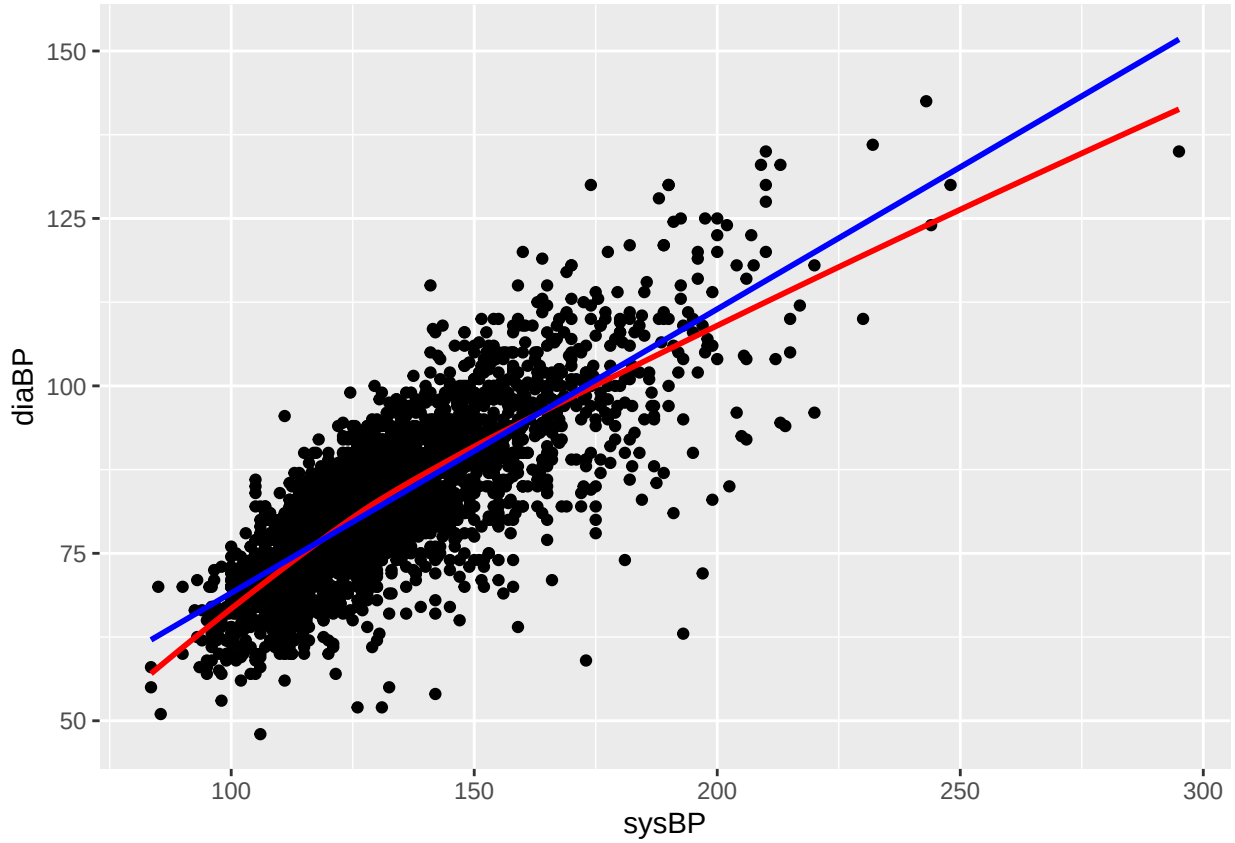
Interpretation of prop.test result

Our p-value (0.05239) is greater than 0.05 so we can't conclude that the difference in proportions between smokers and people who had heart diereses is statistically significant. Consequently, we can't move further and do inferential statistical analysis

Linear Model1

```
#Drawing LOESS curves
ggplot(data = train_quan_vars) +
  geom_point(mapping = aes(x=sysBP, y=diaBP)) +
  geom_smooth(mapping = aes(x=sysBP, y=diaBP), method = "loess", se = FALSE, color = "red") +
  geom_smooth(mapping = aes(x=sysBP, y=diaBP), method = "lm", se = FALSE, color="blue")
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```




```
lm_Model1 <- lm(diaBP~sysBP, data = train_quan_vars)
summary(lm_Model1)
```

```
##
## Call:
## lm(formula = diaBP ~ sysBP, data = train_quan_vars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.501  -4.374   0.327   4.816  29.554
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.681335   0.835861   31.92   <2e-16 ***
## sysBP        0.423934   0.006215   68.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.506 on 2925 degrees of freedom
## Multiple R-squared:  0.614,  Adjusted R-squared:  0.6139
## F-statistic:  4653 on 1 and 2925 DF,  p-value: < 2.2e-16
```

Linear Model1 Interpretation:

The linear curve is very close and almost perfectly follows LOESS curve. Only concern can be the one potential outlier which pulls LOESS curve down when sysBP value is around 300.

1. Estimate linear equation which can be used to estimate average value of diastolic blood pressure is: $diaBP = 26.960484 + 0.421735(sysBP)$
2. Intercept: Average value of diastolic blood pressure when systolic blood pressure is equal to 0. So in our case we are statistically confident enough to say that (p-value <2e-16 *** is very small) diastolic blood pressure will be equal to 26.96 when systolic blood pressure equals 0. However it doesn't provide us meaningful interpretation because according to current context we can't assume that systolic blood pressure of alive person will be 0.
3. Slope: Change of diastolic blood pressure over systolic blood pressure. So in our case it means, whenever systolic blood pressure increases by 10 unit, we are statistically confident enough to say that (p-value <2e-16 *** is very small) the diastolic blood pressure will increase by 4.2 units.
4. R-Squared: 61.14%. In our case 61.14% of the diastolic blood pressure can be explained by the systolic blood pressure. We can assume that we got relevantly small R-Squared because there is natural variability in the data if we look to LOESS and Linear curves plot

Linear Model2

```
#Dependent var: TenYearCHD, Predictors: age, prevalentStroke, sysBP, and glucose

#Check for NA values
age_isna <- sum(is.na(train$age))
prevalentStroke_isna <- sum(is.na(train$prevalentStroke))
sysBP_isna <- sum(is.na(train$sysBP))
glucose_isna <- sum(is.na(train$glucose))
c(age_isna, prevalentStroke_isna, sysBP_isna, glucose_isna)
```

```
## [1]   0   0   0 304
```

```
#remove records with missing glucose values
data = train[!is.na(train$glucose),]

#Linear probability model
lm_model2 = lm(TenYearCHD ~ age+prevalentStroke+sysBP+glucose, data=data)
summary(lm_model2)
```

```
##
## Call:
## lm(formula = TenYearCHD ~ age + prevalentStroke + sysBP + glucose,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67571 -0.18673 -0.10474 -0.03413  1.08345
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.5968930  0.0460913 -12.950  < 2e-16 ***
## age              0.0068832  0.0007894   8.720  < 2e-16 ***
## prevalentStroke  0.2295734  0.0756630   3.034  0.00243 **
## sysBP            0.0021592  0.0003054   7.070 1.91e-12 ***
## glucose          0.0014641  0.0002590   5.653 1.72e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3445 on 3081 degrees of freedom
## Multiple R-squared:  0.08556,    Adjusted R-squared:  0.08437
## F-statistic: 72.07 on 4 and 3081 DF,  p-value: < 2.2e-16
```

Linear Model2 Interpretation

1. Predicted probability predict whether a patient has a 10-year risk of future coronary heart disease (TenYearCHD) $= -0.5969 + 0.0069(age) + 0.2296(prevalentStroke) + 0.0022(sysBP) + 0.0014(glucose)$
2. The four predictors have very small p-values which inidicates that they are all statistically significant to the probability of a patient having a 10-year risk of future coronary heart disease.
3. Age: When we control for if a patient has previously had stroke, the systolic blood pressure, and glucose level, the predicted probability that a patient would have a 10-year risk of future coronary heart disease will increase by 0.69% when the age increases by 1.
4. Prevalent Stroke (prevalentStroke): Considering patient's with the same age, systolic blood pressure, and glucose level, the predicted probability that a patient would have a 10-year risk of future coronary heart disease is 22.9% higher in patients who have previously had stroke comapred to patients who have not.
5. Systolic blood pressure (sysBP): After controlling for the patient's age, glucose level and whether or not the patient has previously had stroke, the predicted probability that a patient would have a 10-year risk of future coronary heart disease will increase by 0.2% when the systolic blood pressure increases by 1 unit.
6. Glucose: Considering patients with the same age, prevalent stroke status and systolic blood pressure,the predicted probability that a patient would have a 10-year risk of future coronary heart disease will increase by 0.1% when the glucose level increases by 1 unit.

Linear Model3

```
traindata <- read.csv("~/Documents/MS-BDSA 2023/ITM 883/Group Project/train.csv")
testdata <- read.csv("~/Documents/MS-BDSA 2023/ITM 883/Group Project/test.csv")
traindata=na.omit(traindata)
testdata=na.omit(testdata)

lm_model3 <- lm(BPMeds ~ age + BMI + totChol + is_smoking, data = traindata)
summary(lm_model3)
```

```
##
## Call:
## lm(formula = BPMeds ~ age + BMI + totChol + is_smoking, data = traindata)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12542 -0.04951 -0.02783 -0.00557  1.00942
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.024e-01  2.869e-02  -7.056 2.14e-12 ***
## age            2.216e-03  3.868e-04   5.729 1.11e-08 ***
## BMI            2.713e-03  7.722e-04   3.514 0.000448 ***
## totChol        2.265e-04  7.302e-05   3.101 0.001945 **
## is_smokingYES -1.903e-03  6.431e-03  -0.296 0.767313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1686 on 2922 degrees of freedom
## Multiple R-squared:  0.02734,    Adjusted R-squared:  0.02601
## F-statistic: 20.53 on 4 and 2922 DF,  p-value: < 2.2e-16
```

Linear Model 3 Interpretation

Coefficients:

1. The intercept of -0.20240120 is small. It represents the expected change in the dependent variable "BPMeds" when all other independent variables are held constant. It indicates the predicted probabilities are not zero(0) for each of the independent variables. This also translates into a probability of 44.9%, the chance that one will be one BPMeds (BP Medication) when all the other factors are controlled to zero or constant. In this case, when age, BMI, totChol, and is_smokingYES are held constant (0), "BPMeds" is expected to decrease by approxi -0.202 units.

2. is_smokingYES: has an estimated coefficient of - 0.00022646. This represents a 49.9% probability that respondents who answered YES to smoking would be on BPMeds in the future compared to when "NO" to smoking. This is a small number given a higher p-value of (0.767313) which is greater than the commonly used benchmark of 0.05.

3. *age:* with a coefficient of 0.00221604 representing a 50% probability, this means there is an expected change in "BPMeds" for a one-unit increase in "age", while holding all other variables/factors constant. As low as this probability may be indicating unlikely probability, this is not the case for p-value of 0.00000001114939. A small p-value indicates a strong relationship between dependent variable and the independent variable, meaning with increased in age, a respondent is likely to be on BPMeds.

4. *BMI:* The estimated coefficient for "BMI" is 0.00271332 . It represents the expected change in "BPMeds" for a one-unit increase in "BMI", while holding all other variables constant. In this case, for each one-unit increase in "BMI", "BPMeds" is expected to increase by approximately 0.0028 units. This also represents a probability of 50%.

5. *totChol:* The estimated coefficient for "totChol" is 0.00022646. It represents the expected change in "BPMeds" for a one-unit increase in "totChol", while holding all other variables constant. In this case, for each one-unit increase in "totChol", "BPMeds" is expected to increase by approximately 0.0002 units.

Logistic Regression using all the variables in the data set

```r
## Complete model
train <- read.csv("~/Documents/MS-BDSA 2023/ITM 883/Group Project/train.csv")
train <- na.omit(train)

dim(train)
```

```
## [1] 2927   17
```

```r
lr_model_full <- glm(TenYearCHD ~ ., data = train, family = binomial(link = "logit"))
summary(lr_model_full)
```

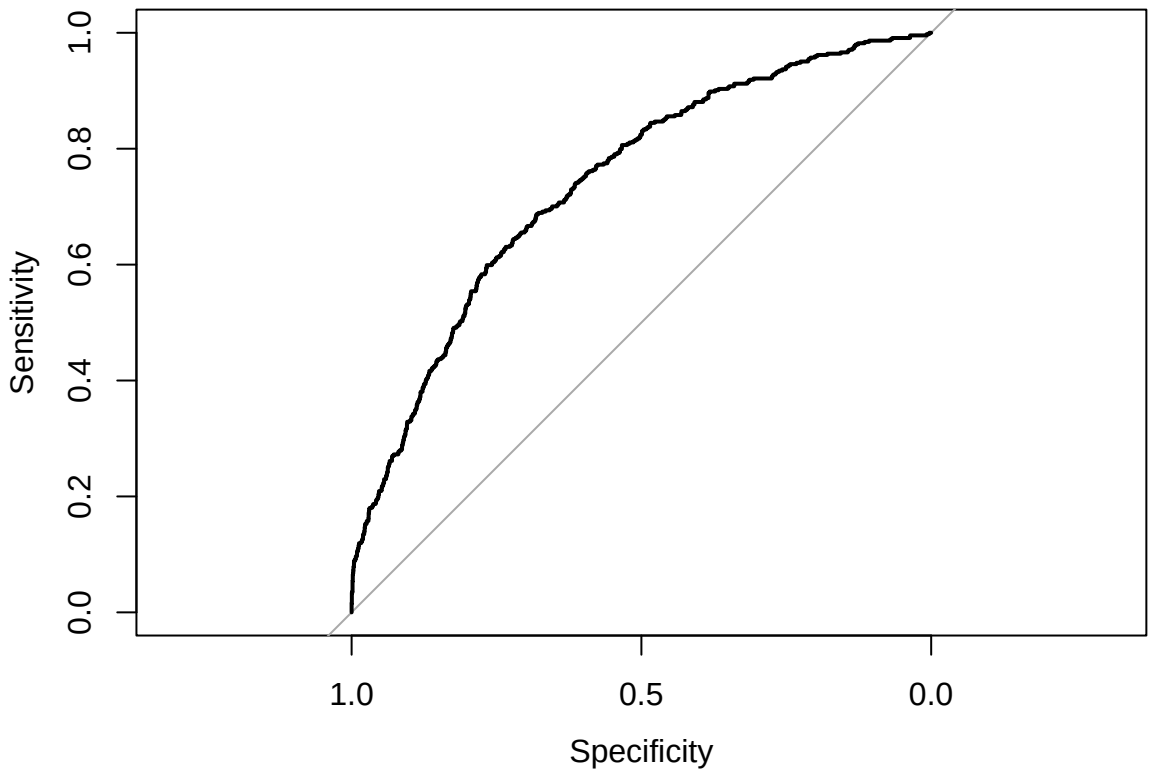
```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5488  -0.5918  -0.4238  -0.2807   2.8323
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.515e+00  7.950e-01 -10.711  < 2e-16 ***
## id               1.583e-05  5.559e-05   0.285 0.775776
## age              6.375e-02  7.459e-03   8.547  < 2e-16 ***
## education       -5.024e-02  5.569e-02  -0.902 0.366950
## sexM             4.937e-01  1.229e-01   4.016 5.91e-05 ***
## is_smokingYES    2.052e-01  1.749e-01   1.173 0.240627
## cigsPerDay       1.750e-02  6.997e-03   2.502 0.012367 *
## BPMeds           1.214e-01  2.654e-01   0.457 0.647437
## prevalentStroke  9.152e-01  5.273e-01   1.736 0.082611 .
## prevalentHyp     1.830e-01  1.551e-01   1.180 0.238129
## diabetes        -8.992e-02  3.580e-01  -0.251 0.801671
## totChol          3.307e-03  1.243e-03   2.660 0.007810 **
## sysBP            1.663e-02  4.240e-03   3.923 8.76e-05 ***
## diaBP           -8.172e-03  7.052e-03  -1.159 0.246562
## BMI              6.366e-03  1.416e-02   0.450 0.652902
## heartRate       -4.510e-03  4.698e-03  -0.960 0.337070
## glucose          9.122e-03  2.535e-03   3.599 0.000319 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2491.6  on 2926  degrees of freedom
## Residual deviance: 2192.9  on 2910  degrees of freedom
## AIC: 2226.9
##
## Number of Fisher Scoring iterations: 5
```

```r
# Prediction

PredictedProbability <- predict(lr_model_full, train, type = "response")

# ROC curves

library(pROC)
MyROC <- roc(train$TenYearCHD, PredictedProbability)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
plot(MyROC)
```




```r
treshold <- coords(MyROC, "best")[1]
coords(MyROC, "best")
```

```
##   threshold specificity sensitivity
## 1 0.1521778    0.681031   0.6869369
```

```
MyROC
```

```
##
## Call:
## roc.default(response = train$TenYearCHD, predictor = PredictedProbability)
##
## Data: PredictedProbability in 2483 controls (train$TenYearCHD 0) < 444 cases (train$TenYearCHD 1).
## Area under the curve: 0.7398
```

```
treshold <<- treshold$threshold


# Testing validation accuracy


library(boot)
# Defining cost functions
cost.error <- function(r, pi=0){
  # Using mean as values are 1 and 0
  # If |r-pi| > 0.5, then the prediction is wrong as 0.5 is used as the treshold
  mean(abs(r - pi)>treshold)
}

cost.accuracy <- function(r, pi=0){
  # |r-pi| < 0.5 gives correct predictions, and mean gives accuracy
  mean(abs(r-pi)<treshold)
}

cost.specificity <- function(r, pi=0){
  # Sum of cases where positive is predicted accurately

  TN = sum((pi<treshold)&(r==0))
  # Sum of Cases where predicted as 1 but actual is 0
  FP = sum((pi>treshold)&(r==0))

  return(TN/(TN+FP))
}
cost.sensitivity <- function(r, pi=0){
  # Sum of cases where positive is predicted accurately

  TP = sum((pi>treshold)&(r==1))
  # Sum of Cases where predicted as 1 but actual is 0
  FN = sum((pi<treshold)&(r==1))

  return(TP/(TP+FN))
}


# Cross-validation
set.seed(36)

cv.error <- cv.glm(data = train, glmfit = lr_model_full, cost = cost.error, K = 5)
cv.accuracy <- cv.glm(data = train, glmfit = lr_model_full, cost = cost.accuracy, K = 5)
```

```
cv.sensitivity <- cv.glm(data = train, glmfit = lr_model_full, cost = cost.sensitivity, K = 5)
cv.specificity <- cv.glm(data = train, glmfit = lr_model_full, cost = cost.specificity, K = 5)


print(list(error = cv.error$delta[1], accuracy = cv.accuracy$delta[1],
           sensitivity = cv.sensitivity$delta[1], specificity = cv.specificity$delta[1]))
```

```
## $error
## [1] 0.423642
##
## $accuracy
## [1] 0.5756748
##
## $sensitivity
## [1] 0.664812
##
## $specificity
## [1] 0.6751715
```

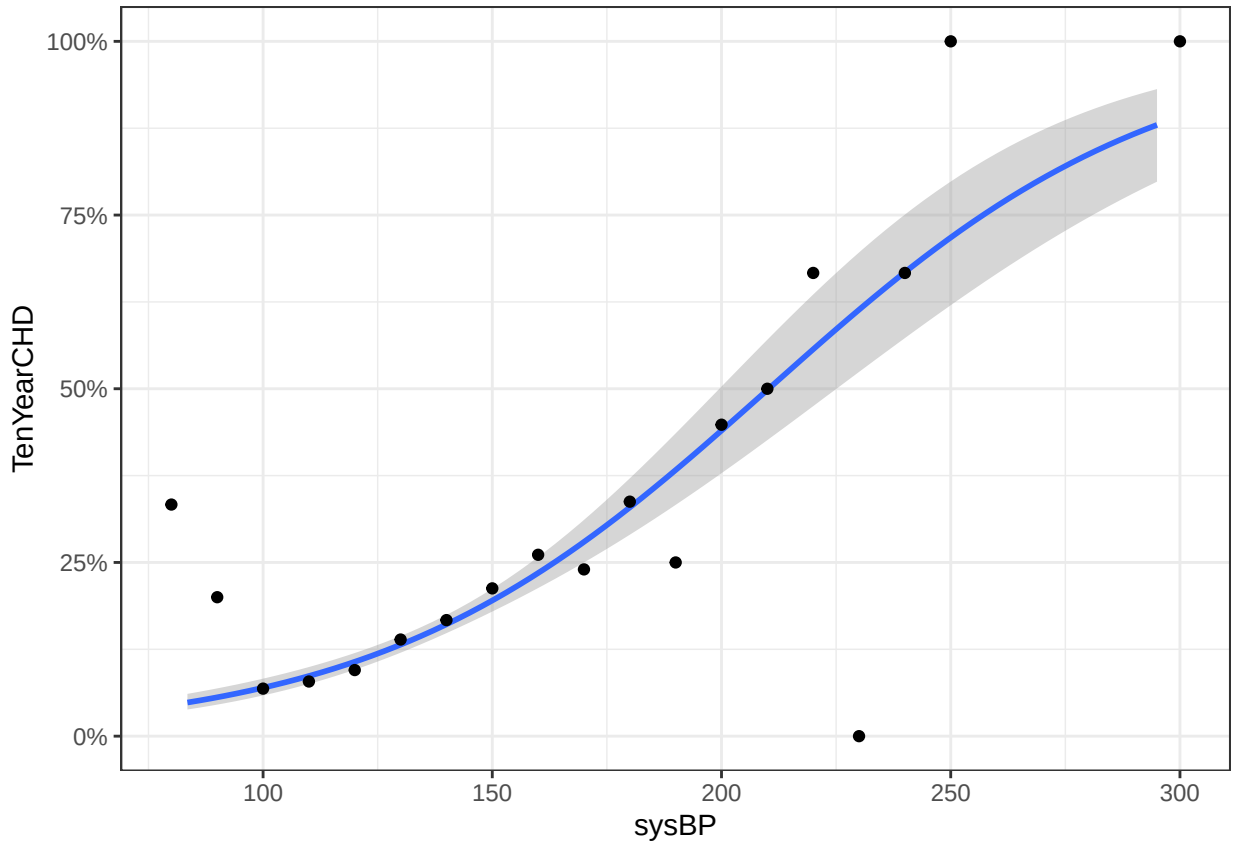
Interpretation

1. Age, Sex, cigsPerDay, totChol, sysBP, and glucose are the variables with low p-value, and have statistically significant impact on the Ten Year CHD
2. We observe a higher accuracy in the complete model as compared to picking only specific variables

Logistic Regression Model1

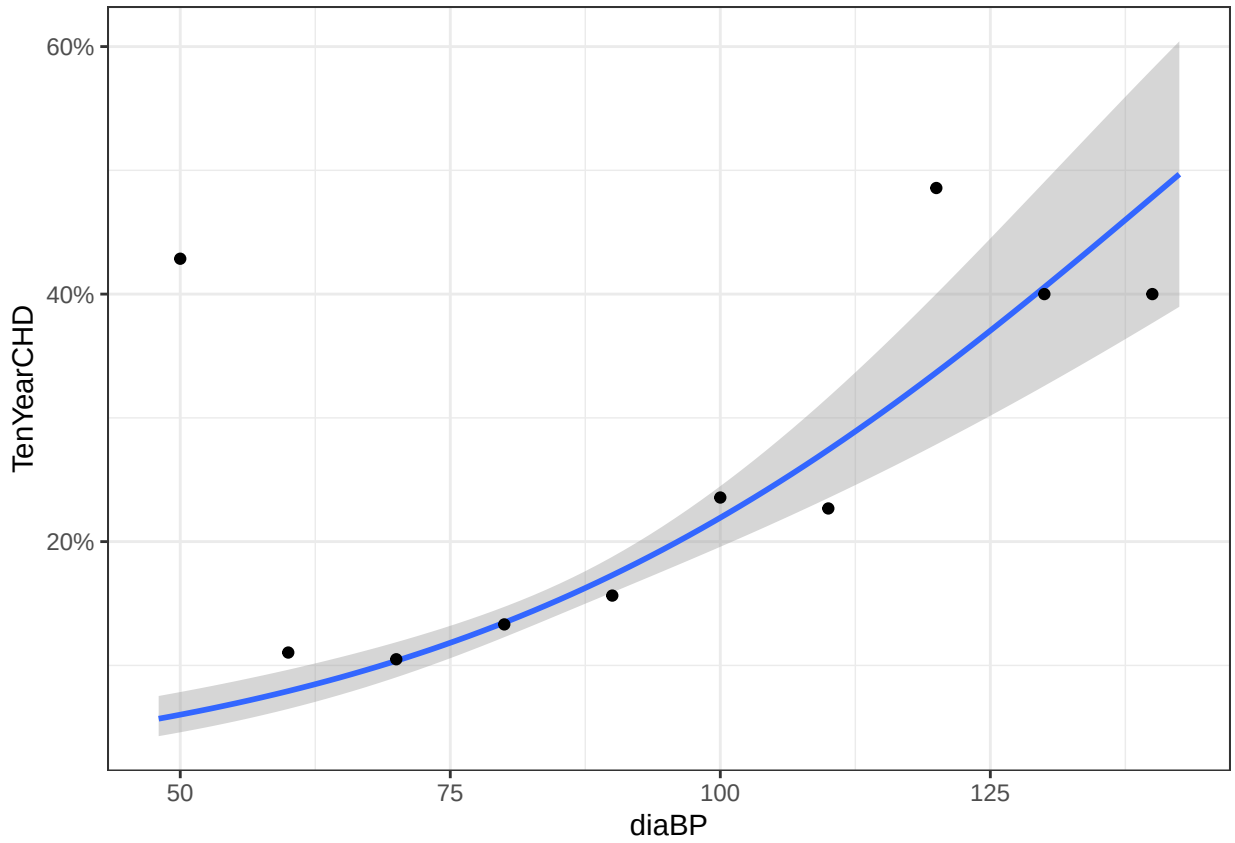
```
train_data <- read.csv("~/Documents/MS-BDSA 2023/ITM 883/Group Project/train.csv")
test_data <- read.csv("~/Documents/MS-BDSA 2023/ITM 883/Group Project/test.csv")

# Remove ID Variable
train <- train_data[, -1]
test <- test_data[, -1]

# Creating binary variables for sex

train$sex <- ifelse(train$sex == "M", 1, 0)
test$sex <- ifelse(test$sex == "M", 1, 0)

# Checking and removing NA records
train <- train[!is.na(train$totChol),]

# Visualizing probability distribution
ggplot(data=train, aes(x=sysBP, y=TenYearCHD)) +
  geom_smooth(method="glm", method.args = list(family = "binomial")) +
  stat_summary(data=train, aes(x=round(sysBP,-1), y=TenYearCHD), fun=mean, geom="point") +
  scale_y_continuous(labels = scales::percent) +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(data=train, aes(x=diaBP, y=TenYearCHD)) +
  geom_smooth(method="glm", method.args = list(family = "binomial")) +
  stat_summary(data=train, aes(x=round(diaBP,-1), y=TenYearCHD), fun=mean, geom="point") +
  scale_y_continuous(labels = scales::percent) +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Logistic Regression model predicting TenYearCHD using sex, systolic BP and
# Total Cholesterol as independent variables.
lr_model1 <- glm(TenYearCHD ~ sex + sysBP + totChol, data = train, family = binomial(link = "logit"))
summary(lr_model1)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ sex + sysBP + totChol, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6009  -0.5995  -0.4813  -0.3634   2.6896
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.255536   0.383437 -16.314  < 2e-16 ***
## sex          0.623075   0.101640   6.130 8.77e-10 ***
## sysBP        0.023580   0.002081  11.331  < 2e-16 ***
## totChol      0.004147   0.001084   3.826  0.00013 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2838.0  on 3351  degrees of freedom
```

```
## Residual deviance: 2651.9  on 3348  degrees of freedom
## AIC: 2659.9
##
## Number of Fisher Scoring iterations: 5
```

Logistic Regression Model1 Interpretation:

1. Predicted Probability of Chronic Heart Disease (TenYearCHD) =

$$\frac{e^{(-6.256+0.623(sex=male)+0.024(sysBP)+0.0041(totChol))}}{1+e^{(-6.256+0.623(sex=male)+0.024(sysBP)+0.0041(totChol))}}$$

2. *Sex:* The very small p-value indicates that the predicted odds of CHD for males is significantly different than that of females, when controlling for systolic BP and total cholesterol The predicted odds for Males is 1.86 times as that of Females, for the same sysBP and total cholesterol This indicates that Males are at a higher risk of Congenital Heart Disease as compared to Females with the same level of systolic Blood Pressure and Total Cholestrol.

3. *sysBP:* The very low p-value indicates that there is a statistically significant impact on the odds for a change in sysBP Considering a person of same sex and total cholesterol level, the predicted odds of CHD increases by around 27% for every 10 units(mmHg) increase in systolic BP.

4. *totChol:* The very low p-value indicates that there is a statistically significant impact on the odds for a change in totChol Considering a person of same sex and sysBP levels, the predicted odds of CHD increases by around 4% for every 10 units(mg/dL) increase in total Cholestrol.

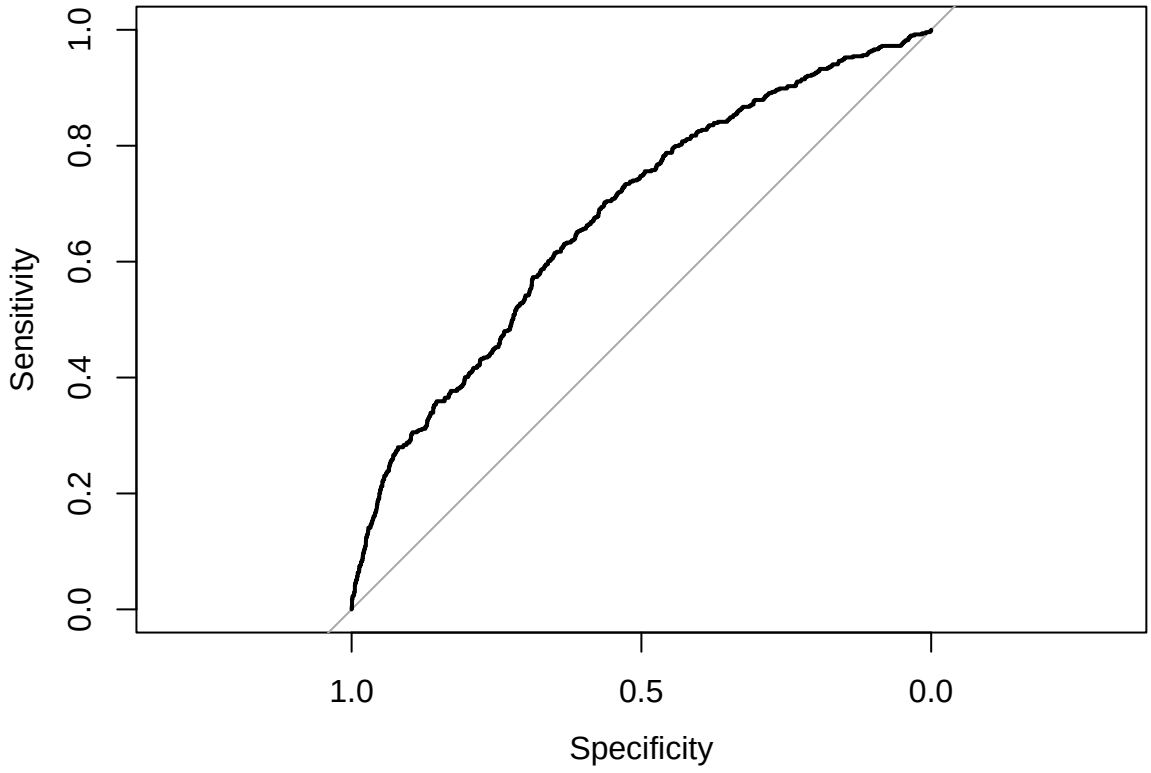
Logistic Regression Model1 Evaluation:

```
# Prediction
PredictedProbability <- predict(lr_model1, train, type = "response")

# ROC curves
MyROC <- roc(train$TenYearCHD, PredictedProbability)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(MyROC)
```




```
treshold <- coords(MyROC, "best")[1]
coords(MyROC, "best")
```

```
##   threshold specificity sensitivity
## 1  0.133796   0.5635534    0.702381
```

```
MyROC
```

```
##
## Call:
## roc.default(response = train$TenYearCHD, predictor = PredictedProbability)
##
## Data: PredictedProbability in 2848 controls (train$TenYearCHD 0) < 504 cases (train$TenYearCHD 1).
## Area under the curve: 0.6774
```

```
treshold <<- treshold$threshold


# Testing validation accuracy
library(boot)
# Defining cost functions
cost.error <- function(r, pi=0){
  # Using mean as values are 1 and 0
  # If |r-pi| > 0.5, then the prediction is wrong as 0.5 is used as the treshold
```

```
  mean(abs(r - pi)>treshold)
}

cost.accuracy <- function(r, pi=0){
  # |r-pi| < 0.5 gives correct predictions, and mean gives accuracy
  mean(abs(r-pi)<treshold)
}

cost.specificity <- function(r, pi=0){
  # Sum of cases where positive is predicted accurately
  TN = sum((pi<treshold)&(r==0))
  # Sum of Cases where predicted as 1 but actual is 0
  FP = sum((pi>treshold)&(r==0))

  return(TN/(TN+FP))
}
  cost.sensitivity <- function(r, pi=0){
    # Sum of cases where positive is predicted accurately
    TP = sum((pi>treshold)&(r==1))
    # Sum of Cases where predicted as 1 but actual is 0
    FN = sum((pi<treshold)&(r==1))

    return(TP/(TP+FN))
  }


# Cross-validation
set.seed(36)

cv.error <- cv.glm(data = train, glmfit = lr_model1, cost = cost.error, K = 5)
cv.accuracy <- cv.glm(data = train, glmfit = lr_model1, cost = cost.accuracy, K = 5)
cv.sensitivity <- cv.glm(data = train, glmfit = lr_model1, cost = cost.sensitivity, K = 5)
cv.specificity <- cv.glm(data = train, glmfit = lr_model1, cost = cost.specificity, K = 5)


print(list(error = cv.error$delta[1], accuracy = cv.accuracy$delta[1],
           sensitivity = cv.sensitivity$delta[1], specificity = cv.specificity$delta[1]))
```

```
## $error
## [1] 0.525358
##
## $accuracy
## [1] 0.4788186
##
## $sensitivity
## [1] 0.6849978
##
## $specificity
## [1] 0.5644697
```

Interpretation of low threshold and accuracy:

```
table(train$TenYearCHD)
```

```
##
##    0    1
## 2848  504
```

Requires data imbalance handling, since training data contains majority FALSE classified observations, which leads to the model having a bias. This is evident with the low best threshold value

Logistic Regression Model2

```
train <- read.csv("~/Documents/MS-BDSA 2023/ITM 883/Group Project/train.csv")
test <- read.csv("~/Documents/MS-BDSA 2023/ITM 883/Group Project/test.csv")

colSums(is.na(train))
```

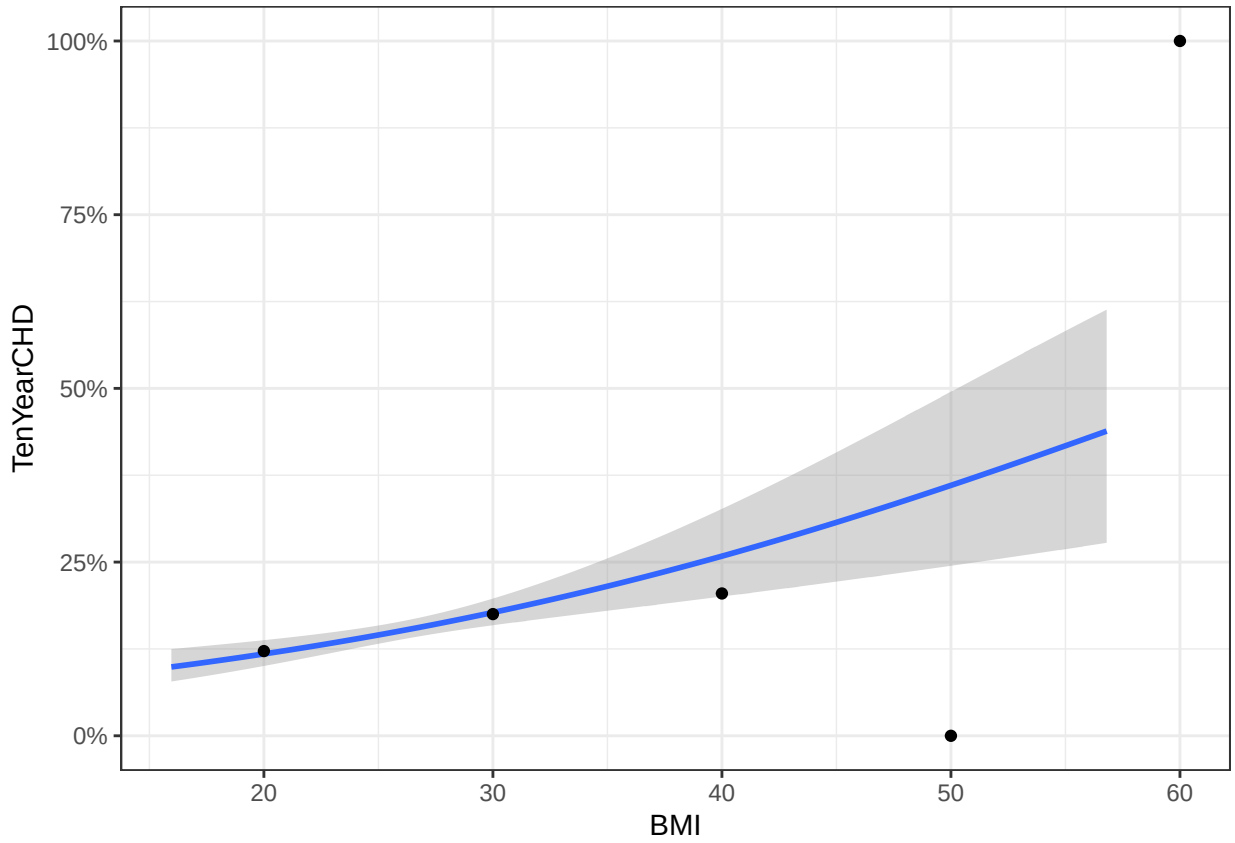
```
##              id             age       education             sex      is_smoking
##               0               0              87               0               0
##      cigsPerDay          BPMeds prevalentStroke    prevalentHyp        diabetes
##              22              44               0               0               0
##         totChol           sysBP           diaBP             BMI       heartRate
##              38               0               0              14               1
##         glucose      TenYearCHD
##             304               0
```

```
train = na.omit(train)

colSums(is.na(test))
```

```
##              id             age       education             sex      is_smoking
##               0               0              18               0               0
##      cigsPerDay          BPMeds prevalentStroke    prevalentHyp        diabetes
##               7               9               0               0               0
##         totChol           sysBP           diaBP             BMI       heartRate
##              12               0               0               5               0
##         glucose
##              84
```

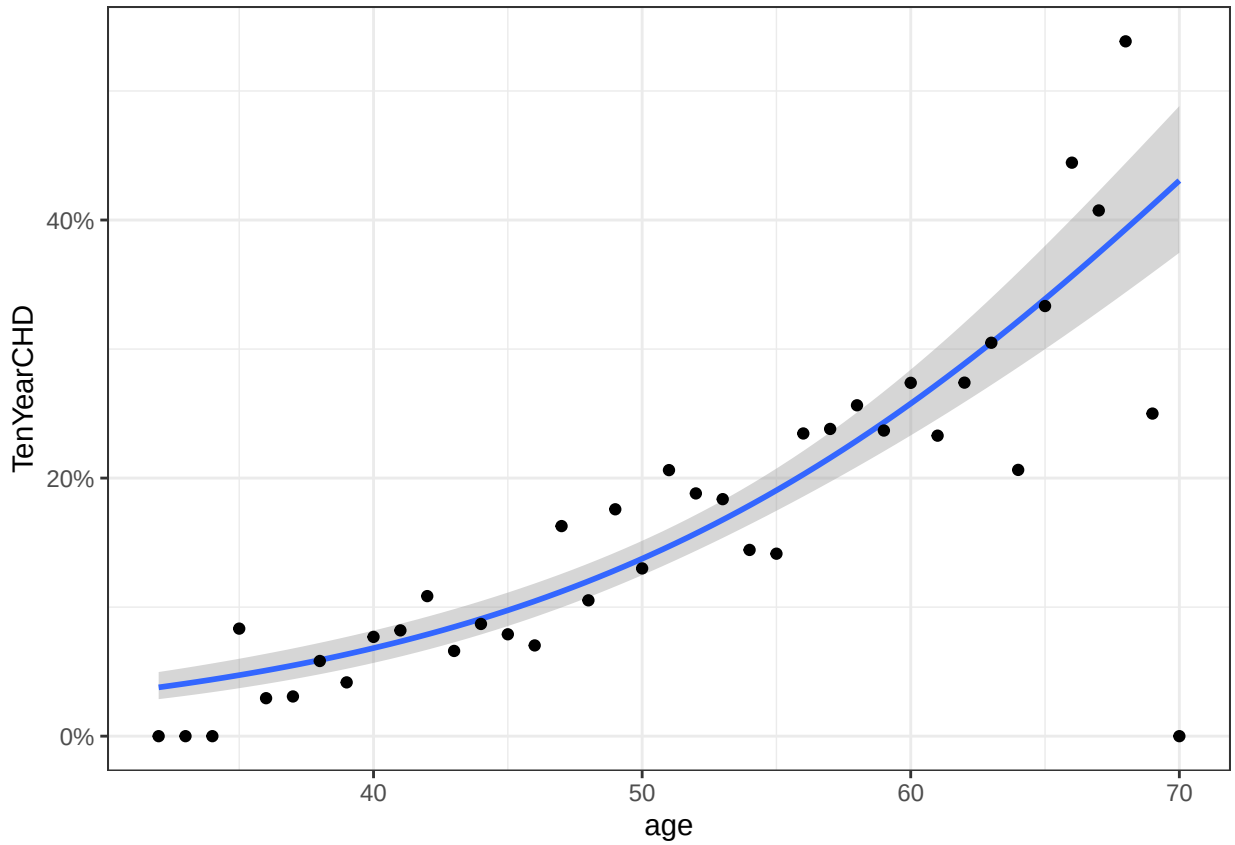
```
test = na.omit(test)

# Visualizing the data based on independent variables

library(ggplot2)
ggplot(data=train) + geom_smooth(aes(x=BMI, y=TenYearCHD), method="glm",
                                 method.args = list(family = "binomial")) +
  stat_summary(data=train, aes(x=round(BMI,-1), y=TenYearCHD), fun=mean, geom="point") +
  scale_y_continuous(labels = scales::percent_format()) +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(data=train) + geom_smooth(aes(x=age, y=TenYearCHD), method="glm",
                                 method.args = list(family = "binomial")) +
  stat_summary(data=train, aes(x=age, y=TenYearCHD), fun=mean, geom="point") +
  scale_y_continuous(labels = scales::percent_format()) +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Logistic Regression Model to predict dependent variable TenYearCHD using Age,Prevalent Stroke,
# Prevalent Hyp, diaBP, BMI as independent variables
lr_model2 = glm(TenYearCHD ~ age + prevalentStroke + prevalentHyp + diaBP + BMI, data = train,
             family=binomial(link="logit"))
summary(lr_model2)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ age + prevalentStroke + prevalentHyp +
##     diaBP + BMI, family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3388  -0.6109  -0.4464  -0.3385   2.6076
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -6.412951   0.586124 -10.941  < 2e-16 ***
## age              0.067803   0.006692  10.132  < 2e-16 ***
## prevalentStroke  0.770105   0.511480   1.506  0.13216
## prevalentHyp     0.403090   0.138996   2.900  0.00373 **
## diaBP            0.010266   0.005399   1.901  0.05726 .
## BMI              0.006554   0.013270   0.494  0.62137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2491.6  on 2926  degrees of freedom
## Residual deviance: 2295.7  on 2921  degrees of freedom
## AIC: 2307.7
##
## Number of Fisher Scoring iterations: 5
```

Logistic Regression Model 2 Interpretation:

1. Predicted probability for TenYearCHD =

$$\frac{e^{(-6.41+0.068(age)+0.77(prestroke=yes)+0.40(preHyp=yes)+0.1(diaBP)+0.006(BMI))}}{1+e^{(-6.41+0.068(age)+0.77(prestroke=yes)+0.40(preHyp=yes)+0.1(diaBP)+0.006(BMI))}}$$

2. *Intercept:* The p-value for the intercept is very small, which means we are confident that the predicted probability is not equal to 0 when the value independent variables are 0.

3. *Age:* The very small p-value indicates that there is statistically significant difference in the predicted probability of TenYearCHD for people with different ages. When controlling for variables prevalentStroke, prevalentHyp, diaBP and BMI, the predicted odds for TenYearCHD increases by 7% with every one-year increase in age. This indicates that, there is an increased chance in coronary heart disease with increase in age of people.

4. *PrevalentStroke:* When controlling for variables age, prevalentHyp, diaBP and BMI, the high p-value indicates that there is not a statistically significant impact on TenYearCHD whether the person had prevalent stroke or not.

5. *PrevalentHyp:* The very small p-value indicates that there is statistically significant difference in the predicted probability of TenYearCHD for people who have prevalent hypertension. When controlling for variables age, prevalentstroke, diaBP and BMI, the predicted odds for TenYearCHD is 1.496 times more than that of a person who does not have prevalent hypertension.

6. *diaBP:* When controlling for variables age, prevalentstroke, prevalentHyp, and BMI, the high p-value indicates that there is not a statistically significant impact on TenYearCHD with change in diaBP.

7. *BMI:* When controlling for variables age, prevalentstroke, prevalentHyp and diaBP, the high p-value indicates that there is not a statistically significant impact on TenYearCHD for people with different BMI values.

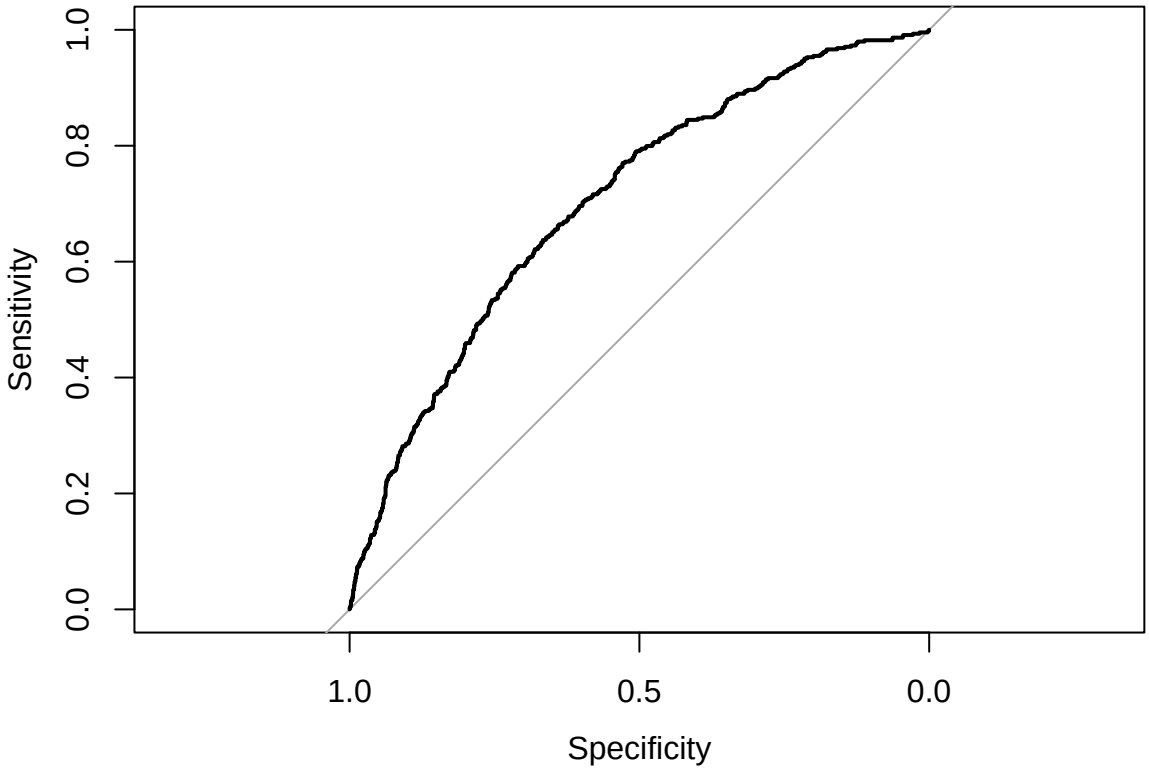
Logistic Regression Model 2 Evaluation:

```
Predicted_prob = predict(lr_model2, data=train, type="response")
ROC = roc(train$TenYearCHD,Predicted_prob)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(ROC)
```




```
# Finding best cut-off value
coords(ROC,"best")
```

```
##   threshold specificity sensitivity
## 1 0.1582665   0.6665324   0.6373874
```

```
table(train$TenYearCHD)
```

```
##
##    0    1
## 2483  444
```

```
ROC
```

```
##
## Call:
## roc.default(response = train$TenYearCHD, predictor = Predicted_prob)
##
## Data: Predicted_prob in 2483 controls (train$TenYearCHD 0) < 444 cases (train$TenYearCHD 1).
## Area under the curve: 0.7027
```

The model has an AUC of 0.71 indicating that it has moderate discriminatory power. The best cut off value i.e threshold is given as 0.16 which means if the predicted probability of the variable TenYearCHD is higher than 0.16, then the person is classified as having a high risk of developing coronary heart disease. The

threshold value obtained from the model is very low due to the imbalanced distribution of the target variable i.e TenYearCHD in the dataset. Therefore, it can be said that the model is biased towards predicting a negative outcome for the target variable.

```
# Confusion matrix
threshold = 0.16
predicted_label <- ifelse(Predicted_prob > threshold , 1, 0)
actual_label <- train$TenYearCHD
confusion_matrix <- table(predicted_label, actual_label)
print(confusion_matrix)
```

```
##                actual_label
## predicted_label    0    1
##               0 1674  167
##               1  809  277
```

```
# Calculate accuracy, precision and recall
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
precision <- confusion_matrix[2,2] / sum(confusion_matrix[,2])
recall <- confusion_matrix[2,2] / sum(confusion_matrix[2,])
```

```
## Accuracy: 0.6665528
```

```
## Precision: 0.6238739
```

```
## Recall: 0.2550645
```

We observe an accuracy of 66.65% for our model with low precision, this indicates that the model is not quite highly accurate in predicting the risk of having coronary heart disease in people. Low accuracy can be explained by the imbalance in the data which caused the bias in the model to predict many FALSE positives as the best cutoff value is quite low at 0.16.