Machine Learning: Using the Logistic Regression Model to Predict Coronary Heart Disease

Victor Wright

September 2019

Abstract

Logistic regression is a supervised machine learning algorithm that can be applied to some classification problems. In "Machine Learning: Using the Logistic Regression Model to Predict Coronary Heart Disease", we first provide a summary of the logistic regression model and identify similarities between it and linear regression. We also discuss the structure of the target variable, explain the modeling objective of logistic regression, introduce the concept of utilizing conditional probability in classification problems, and present the two forms of the logistic regression model to the reader. Next, we provide a definition of coronary heart disease (CHD) and identify and discuss the major risk factors of CHD recognized by government health organizations as well as health care professionals and researchers. For continuous risk factors, we indicate how these quantities have and can represent different categories, or levels, of certain medical conditions. For example, the condition of obesity is identified if a patient's body mass index (BMI) measurement falls into a certain range. Moreover, obesity is a condition that can also be further subdivided into different stages of obesity based upon certain measurements of BMI according to government and health care professionals. Following our discussion of the major CHD risk factors, we explain our target variable and give meaning to the values it contains. Next, we supply to the reader the data source, provide a brief history of the Framingham Heart study, show our data cleansing procedure, and describe a routine that can be used to produce empirical logit plots for categorical and continuous independent variables. Finally, in the later sections of this paper, we provide a description of variables, fit logistic regression models, analyze receiver operating characteristics curves, explore confusion matrices, and discuss variable selection methods used. In particular, we make use of the Boruta (search) algorithm as well as variable importance rankings to help us construct a robust model we use for classification.

A Brief Discussion of the Logistic Regression Model

The *logistic regression model* is somewhat similar to the linear regression model. It is similar because like linear regression it is also a *supervised learning method* which can be estimated using information in a *data matrix*, J, that contains k *variables* and n *observations* (*i.e.*, $\dim(J) = n \times k$) [**?**]. One being the *dependent/response* variable and *k-1 independent/predictor* variables. However, in comparison to linear regression, the modeling objective is somewhat different [1, 2**?**].

Like linear regression, we still want to regress Y on the independent variables $X = [\vec{X_1}, \vec{X_2}, ..., \vec{X_k}]$ [1, 2]. *I.e.*, explain the *linear dependency* between Y and X (where $X, Y \in J$). And like linear regression, logistic regression is considered to be a flexible modeling approach. Flexibility is not declared about f but declared since the model can be applied to model a variety of processes in various fields. However, the information contained in Y is not considered to be continuous. In fact, Y is a *categorical variable* [1, 2**?**]. Furthermore, if Y only contains two categories, these categories are often represented with a *binary coding* (Y can also be referred to as an *indicator variable* or *binomial response*) [1, 2**?**]. *I.e.*, either $Y = 1$ or $Y = 0$ where $Y = 1$ represents *the event of success* and $Y = 0$ represents *the event of failure* [2]. Moreover, in this type of regression problem, we want to model the event of $Y = 1$ for any observation X [1]. In other words, we wish to *classify* which category any observation X_n belongs to based on its *feature values* $[X_1 = x_{n,1}, X_2 = x_{n,2}, ..., X_k = x_{n,k}]$ regardless if X_n is contained in a training or test set. [**?**]. Therefore, this type of problem is a regression problem where the objective is *classification* [**?**]. Finally, we use the concept of *probability* to assign, or classify, which category ($Y = 1$ or $Y = 0$) each observation belongs to [**?**]. A mathematical model for events that are said to be uncertain is probability (probability always lies in the interval $[0, 1]$) which we use to measure the likelihood of uncertain events [3, 4]. In other words, we use the concept of probability to help us model the event $Y = 1$ for any observation X_n [1**?**].

We previously mentioned that logistic regression is similar to the linear regression model. Thus, some of the modeling assumptions for linear regression carry over to the logistic regression model. In particular, the three assumptions that carry over are *linearity*, *independence*, and *randomness* [1]. Independence and randomness can be determined based on how the data were collected [1]. The factor that creates a difference between the two is the modeling objective of logistic regression; modeling the probability of $Y = 1$ [1, 2**?**].

The probability of success, the probability of $Y = 1$ given X_n, can (and has been) represented by various notations. Some examples are $\Pr(Y = 1), \Pr(Y = 1 \mid X = x), p, \pi$, and $p(x)$ for any of the $i = 1, 2, ..., n$ observations in X [1, 2**?**]. In this paper, we will use p which is a *conditional probability*. This just means that the value of p is dependent on some other event that has already been observed so p is not constant; the value of p is dependent on X_n [1, 3, 4]. There are two forms of the logistic regression model that can be used to describe the

likelihood of the event $Y = 1$ given X_n. They are the *logit form (log odds)* and *probability form* of the model [1]. The logit form is

$$\ln(\frac{p}{1-p}) = \beta_0 + \sum_{j=1}^{k} \beta_j X_j \tag{1}$$

and

$$p = \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j X_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j X_j)} \tag{2}$$

is the probability form of the logistic regression model [1]. Additionally, (1) and (2) are the logit and probability forms of a *multiple logistic regression* model for $j = 1, 2, ..., k$. [1]. Note that in between (1) and (2) there is an intermediate transformational step which is called *odds* and is

$$\frac{p}{1-p} = \exp(\beta_0 + \sum_{i=1}^{k} \beta_i X_i) \tag{3}$$

[1]. Linearity can be verified is by assessing the relationship between the *empirical logit* and the predictor variables [1]. We assess linearity later. Finally, *functional transforms* can be applied to the data in the predictors when they are continuous and then included in the logistic regression model as well as *dummy variables* [1**?**]. In fact, linearity is automatic when dummy variables are used in the logistic regression model [1].

Definition of Coronary Heart Disease (CHD) and Major Risk Factors and Conditions

The National Cancer Institue defines *coronary heart disease (CHD)* as a cardiovascular disease (CVD) that is caused by plaques and fatty material build-up that occurs in the coronary arteries of the heart (atherosclerosis) such that the build-up of the material is strong enough to narrow or entirely block the heart's coronary arteries which transport oxygen and blood to the heart [5]. Healthcare professionals claim that any amount of cigarette consumption, diabetes mellitus, obesity, family history of premature CHD, elevated blood pressure, elevated serum total cholesterol (TC), elevated low-density lipoprotein cholesterol (LDL-C), and low serum high-density lipoprotein cholesterol (HDL-C), left ventricular hypertrophy, and advancing age have been considered to be independent major risk factors of CHD [6, 7]. Much scientific and medical research has been conducted on well-known risk factors, and through research findings, decisions and definitions have been made about what measurements categorize such risk factors. *E.g.*, an individuals body mass index (BMI) of a patient has been defined as the calculation of the patient's weight (measured in kilograms) divided by

their height (measured in square meters) and has been accepted by the American Heart Association as an appropriate characteristic that can be used to measure *adiposity* (being categorized as severely overweight or obese) in male and female patients [8]. Moreover, BMI has also been used as a metric in other research papers that discuss obesity and acknowledge it as a risk factor of CHD [6, 7, 9].

In *Assessment of Cardiovascular Risk by Use of Multiple-Risk-Factor Equations* by *Grundy, Pasternak, et al.* the *normal weight* of a patient is defined as $18.5 \leq BMI \leq 24.9$ in $\frac{kg}{m^2}$. They also define the condition of being *overweight* as $25 \leq BMI \leq 29$ in $\frac{kg}{m^2}$, and *obesity* as $BMI > 30$ $\frac{kg}{m^2}$ [7]. Additionally, they state that *obesity class I* is present when $30.0 \leq BMI \leq 34.9$ in $\frac{kg}{m^2}$ is observed, *obesity class II* is present when $35.9 \leq BMI \leq 39.9$ in $\frac{kg}{m^2}$ is observed, and *obesity class III* is present $BMI > 50.0$ in $\frac{kg}{m^2}$ is observed [7]. However, adipose tissue distribution and patient sex can pose difficulties when BMI is used to identify levels of obesity in a patient because this metric does not adjust for where and how fat is distributed in the body (pelvis, abdomen, etc.) which makes it difficult to define universal thresholds, or rankings, of BMI that indicates different levels of unhealthy BMI [8]. In *Current Estimates of the Economic Cost of Obesity in the United States* by *Wolf and Colditz* overweight is considered to be $25.0 \leq BMI \leq 29.9$ in $\frac{kg}{m^2}$ and obesity is considered to be $BMI > 30$ in $\frac{kg}{m^2}$ which corresponds to the AHA's current definition of overweight and obesity based on the BMI metric [9, 10]. Neither parties acknowledge stage I, II, or III. Finally, the definition of normal body weight in the work by *Grundy, Pasternak, et al.* corresponds to the AHA's current definition of normal body weight [6, 10].

Similar to BMI, blood pressure is also a continuous quantity, which is measured in *mm Hg*, and certain *interval readings* of a patient's blood pressure can be used to classify what type of blood pressure a patient has [6, 7, 11]. The JNC-VII ranks blood pressure readings of *systolic BP* < 120 or *diastolic BP* < 80 as *optimal* blood pressure which corresponds to the JNC-VI optimal blood pressure category [11]. When $120 \leq$ *systolic BP* < 130 or $80 \leq$ *diastolic BP* < 85 the JNC-VI classified this as *normal BP* and when $130 \leq$ *systolic BP* < 140 or $85 \leq$ *diastolic BP* < 89 as *borderline high BP* [11]. However, these readings are acknowledged as *prehypertension* according to the JNC-VII [11]. The JNC-VI and VII label *systolic BP* ≥ 140 or *diastolic BP* ≥ 140 as *hypertension* [11]. Moreover, hypertension can be further categorized into levels/stages of hypertension based on different interval readings [6, 11]. For instance, the JNC-VI claims that hypertension can be broken down into three stages, the JNC-VII claims two stages, and *Wilson, Agostino, et al.* state that when *systolic BP* ≥ 160 or *diastolic BP* ≥ 100 this defines *stages II-IV hypertension* [6, 11]. According to the JNC-VI *stage I hypertension* is present when $140 \leq$ *systolic BP* < 160 or $90 \leq$ *diastolic BP* < 100, *stage II hypertension* is present when $160 \leq$ *systolic BP* < 180 or $100 \leq$ *diastolic BP* < 110, and *stage III hypertension* when *systolic BP* ≥ 180 or *diastolic BP* ≥ 110 [11]. And, according to the JNC-VII, the definition of stage I hypertension remains the same as the JNC-VI definition [11]. However, the JNC-VII combines the JNC-VI

definitions of stage II and stage III hypertension into just stage II hypertension [11].

Like using BMI to classify patients who are overweight, obese, or rank in a different obesity stage, it is also difficult to label what "kind" of smoker someone is. *Schane, Ling, and Glance* illustrate this idea by discussing that the action of smoking less than one pack of cigarettes per day, smoking 1 - 39 cigarettes per week, less than ten cigarettes per day, less than 15 cigarettes per day have all been considered as levels of cigarette consumption that classify cigarette smokers as "light smokers" [12]. However, they do claim that cardiovascular disease risk levels remain about the same when comparing the habit of daily smoking with the actions of light and intermittent smoking [12]. Finally, they also consider heavy smoking to be consuming twenty-three or more cigarettes and light smoking to be consuming anywhere in between four and seven cigarettes per day implying that moderate smoking can be defined as $7 < \text{Cigarettes smoked day} < 23$ according to their work [12]. Aside from just advancing age being an absolute short-term risk factor for CHD the National Cholesterol Education Program (NCEP) Expert Panel on Evaluation, Detection, and Treatment of High Blood Cholesterol in Adults (ATP III) have concluded that age can further be decomposed into sex-specific risk factor categories [13]. NCEP ATP III claims that being male *and* forty-five years of age or older is a major risk factor for CHD development [13]. Additionally, being female *and* fifty-five years of age and older [13]. Both being risks that exclude LDL-C [13].

It is easy to see that there are many risk factors of CHD and certain *levels*, or *categories*, of these risk factors as well that have yet to be well-defined most likely due to new findings in observational studies and research. Finally, *Grundy, Pasternak, et al.* state that the probability of oneself developing CHD over a given period of time is the definition of *absolute risk* [7]. Moreover, they state that if the given time period is 10 years or less, then the probability of oneself developing CHD within this timeframe is called the *absolute short-term risk* of CHD development [7]. In this paper, we wish to predict absolute short-term risk of CHD development.

Explanation of the Data Source, Variables, and Data Clean Up

We obtained a data set from Kaggle.com. The URL is https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression [14]. And in fact, this data set has 4240 records and fifteen variables each containing medical visit information for people living in the city of Framingham, Massachusettes and is also part of a longstanding cardiovascular study which began in 1948 and formally called the "Framingham Heart Study" [14, 15, 16]. The National Heart, Lung, and Blood Institute (as of 1976) was the organization that administered and organized the study [15, 16]. Moreover, the observational subjects who were recruited for this

study were supposedly absent of heart disease symptoms at the time of recruitment [15]. The Framingham Heart study was designed to track each patient, of whom had not endured either stroke or heart attack, during long time periods in hopes to uncover what may be common conditions or features that promote cardiovascular disease [15]. In this project, we use this data set and create a new data set to analyze that is dim $= 4142 \times 35$. The 19 new variables given a binary coding to represent various conditions that the patient could have. The new variables have been coded with the help of Excel. We load the new data set into R and clean it up a little using the program. The code is provided below.

```
> CHD.Data <- read.csv("CHD.Data.csv", header = T, sep = ",")
> # The dimension of the data matrix is
>
> dim(CHD.Data)

[1] 4240   38

> # and it contains
> missing.observations <- sum(is.na(CHD.Data))
> missing.observations

[1] 645

> # incomplete observations on thirty-eight variables.
>
> # The names of the variables are
>
> names(CHD.Data)

 [1] "male"                 "age"                  "Age.Risk.Women"
 [4] "Age.Risk.Men"         "education"            "Heavy.Smoker"
 [7] "Moderate.Smoker"      "Light.Smoker"         "currentSmoker"
[10] "cigsPerDay"           "BPMeds"               "prevalentStroke"
[13] "prevalentHyp"         "diabetes"             "Borderline.High.TC"
[16] "Very.High.TC."        "totChol"              "High.sys.BP"
[19] "High.dia.BP"          "High.BP"              "Hyper.sys.BP"
[22] "Hyper.dia.BP"         "Hypertension"         "Advanced.Hyper.sys"
[25] "Advanced.Hyper.dia"   "Advanced.Hypertension" "sysBP"
[28] "diaBP"                "Normal.Body.Weight"   "Overweight"
[31] "Obesity"              "ClassOne.Obesity"     "ClassTwo.Obesity"
[34] "ClassThree.Obesity"   "BMI"                  "heartRate"
[37] "glucose"              "TenYearCHD"

>
> # where male (column 1), age (column 2), education (column 5),
> # currentSmoker (column 9), cigsPerDay (column 10), BPMeds (column 11),
> # prevalentStroke (column 12), prevalentHyp (column 13), diabetes
```

```
> # column(14), totChol (column 17), sysBP (column 27) , diaBP (column 28),
> # BMI (column 35), heartRate (column 36), glucose (column 37), and
> # TenYearCHD (column 38) are all part of the original data set. The rest
> # are categorical variables that have been given a binary coding that
> # represents certain medical conditions the patient does or does not
> # have using the information contained in the variables from
> # the original data set.

> # Our goal is to preseve as much data from the original data set as we
> # can. So we next identify missing values in the original sample
> # (we do not look at calculated variables here) by doing the following
>
> v1 <- sum(is.na(CHD.Data[,1])) # Total missing values in male.
> v2 <- sum(is.na(CHD.Data[,2])) # Total missing values in age.
> v5 <- sum(is.na(CHD.Data[,5])) # Total missing values in education.
> v9 <- sum(is.na(CHD.Data[,9])) # Total missing values in currentSmoker.
> v10 <- sum(is.na(CHD.Data[,10])) # Total missing values in cigsPerDay.
> v11 <- sum(is.na(CHD.Data[,11])) # Total missing values in BPMeds.
> v12 <- sum(is.na(CHD.Data[,12])) # Total missing values in prevalentStroke.
> v13 <- sum(is.na(CHD.Data[,13])) # Total missing values in prevalentHyp.
> v14 <- sum(is.na(CHD.Data[,14])) # Total missing values in diabetes.
> v17 <- sum(is.na(CHD.Data[,17])) # Total missing values in totChol.
> v27 <- sum(is.na(CHD.Data[,27])) # Total missing values in sysBP.
> v28 <- sum(is.na(CHD.Data[,28])) # Total missing values in diaBP.
> v35 <- sum(is.na(CHD.Data[,35])) # Total missing values in BMI.
> v36 <- sum(is.na(CHD.Data[,36])) # Total missing values in heartRate.
> v38 <- sum(is.na(CHD.Data[,37])) # Total missing values in glucose.
> # The number of missing observations in each variable are
>
> variable.name <- c("male", "age", "education", "currentSmoker",
+                    "cigsPerDay", "BPMeds", "prevalentStroke",
+                    "prevalentHyp", "diabetes", "totChol", "sysBP",
+                    "diaBP", "BMI","hearRate", "glucose")
> missing.observations.in.variable <- c(v1,v2,v5,v9,v10,v11,v12,v13,v14,
+                                       v17,v27,v28,v35,v36,v38)
> missing.observations.by.variable <- as.matrix(
+ data.frame(variable.name, missing.observations.in.variable), byrow = TRUE )
> missing.observations.by.variable

      variable.name     missing.observations.in.variable
 [1,] "male"            "  0"
 [2,] "age"             "  0"
 [3,] "education"       "105"
 [4,] "currentSmoker"   "  0"
 [5,] "cigsPerDay"      " 29"
 [6,] "BPMeds"          " 53"
```

```
 [7,] "prevalentStroke" "  0"
 [8,] "prevalentHyp"    "  0"
 [9,] "diabetes"        "  0"
[10,] "totChol"         " 50"
[11,] "sysBP"           "  0"
[12,] "diaBP"           "  0"
[13,] "BMI"             " 19"
[14,] "hearRate"        "  1"
[15,] "glucose"         "388"

>
```

These R commands and functions have allowed us to identify all of the missing cases in each variable. For the purposes of exploring the effect of major risk factors on absolute short-term coronary heart disease, we remove the variables *education, BPMeds*, and *heartRate*. The variable *cigsPerDay* contains twenty-nine missing values. We do not exactly know why this variable contains some null values. However, we do need to leave it in our data set because the values of *cigsPerDay* (which is the average number of cigarettes per day the patient smokes) have been used in the construction of other categorical variables that have been calculated by us. We do not remove *totChol* for the same reason. On another note, we notice that the variable *currentSmoker* does not have any null values. Since this is the case, we hypothesize about why there are missing values in *cigsPerDay*. Some of them could of these hypotheses are:

- The patient is an intermittent smoker (they don't smoke every day)
- They are trying to quit and during cessation they became labeled as an intermittent smoker
- They have recently quit. However, they have not been smoke-free long enough to be labeled as a non-smoker

. Similarly, we don't know exactly why the variables *BPMeds, BMI, heartRate*, and *glucose* have null values. We are not interested in quantifying the effect of blood pressure medication on absolute short-term CHD risk. We avoid these missing cases by simply removing the *BPMeds* from our data set. Finally, we do not know exactly why there are some missing measurements of *BMI, heartRatem*, and *glucose*. However, we do not use any estimation method that can be used to fill in null values. Finally, we remove *glucose* and rely on the dummy variable *diabetes* to represent the diabetes mellitus risk factor rather than glucose measurements. We also eliminate *heartRate*.

```
> Cleaned.CHD.Data <- na.omit(data.frame(CHD.Data[,-c(5,11,37)]))
> # Dimensions of the new data matrix
>
> dim(Cleaned.CHD.Data)
```

```
[1] 4142   35

> # Number of complete cases
>
> sum(complete.cases(Cleaned.CHD.Data))

[1] 4142

> # Number of missing cases
>
> sum(is.na(Cleaned.CHD.Data))

[1] 0
```

Data Visualization: Exploring the Relationship between the Logit Form of the Model and Categorical and Continuous Predictors

The logistic regression model is a *parametric* learning method that assumes there is a *linear* relationship between the logit form of the logistic regression model and predictor(s) [1, 2**?**]. There are two elementary approaches that we can use to assess whether or not the linearity assumption is met [1]. Moreover, these two approaches are dependent upon the type of data that comprises the predictor(s) [1]. As we stated before, if the predictor variable(s) are categorical linearity will be automatic [1]. When the predictor(s) are continuous, assessing the linearity assumption becomes a little more involved. In either case, we need an estimator for $\ln(\frac{p}{1-p}$. We refer to this estimator as the *empirical logit* which is part of the *empirical logit plot* that is used to verify if whether or not the linearity assumption is met [1].

Case I: Epirical Logit Plots for Categorical Predictors

When the predictor is categorical, or coded in binary, the goal is to find an estimator for p and $1-p$ in *each* level/category of the predictor [1]. For example, suppose we consider the variable currentSmoker. The coding scheme for current smoker is

$$currentSmoker \quad = \quad \begin{cases} 1, \text{ the patient is a current smoker} \\ 0, \text{ the patient is not a current smoker} \end{cases}$$

We have already discussed odds in this paper. But these odds are in terms of the logistic regression model (3). However, (3) is currently unknown to us because we do not know what the βs are. Let p_s be the probability of success

(being labeled with absolute short-term CHD risk) for smoking patients. Let $p_{n'}$ be the probability of success (being labeled with absolute short-term CHD risk) and a non-smoker. Then using the information in our data matrix we estimate these probabilities and compute empirical logit values of absolute short-term CHD risk for smokers and non-smokers which are $\ln(\frac{\hat{p_s}}{1-\hat{p_s}})$ and $\ln(\frac{\hat{p_{n'}}}{1-\hat{p_{n'}}})$ respectively. However, we do not use probabilities to obtain empirical logit values for each level of *currentSmoker*. The approach is to count how many smokers are in *currentSmoker* and how many non-smokers are in *currentSmoker* that are labeled with and without absolute short-term risk of CHD development. Let t_s and $t_{n'}$ be the number of smokers and non-smokers without absolute short-term CHD risk respectively. Then, let r_s and $r_{n'}$ be the number of smokers and non-smokers who have an absolute short-term risk of CHD respectively. The empirical logit values for smokers and non-smokers are then $\ln(\frac{r_s}{t_s+r_s})$ and $\ln(\frac{r_{n'}}{t_{n'}+r_{n'}})$. So the estimators of p_s and $p_{n'}$ are *sample proportions*. For more information on computing empirical logit values for categorical predictors see *STAT2: building models for a world of data* by *Cannon, Cobb, et al.* We next produce an empirical logit plot for the variable *currentSmoker*.

```
> library(dplyr)
> # The number of smokers with absolute short-term CHD Risk is
> Cleaned.CHD.Data %>%
+   filter(currentSmoker == 1 & TenYearCHD == 1) %>%
+   select(currentSmoker, TenYearCHD) %>%
+ summarise(sumsmoke = sum(currentSmoker))

  sumsmoke
1      322

> # The number of smokers without absolute short-term CHD Risk is
>
> Cleaned.CHD.Data %>%
+ filter(currentSmoker == 1 & TenYearCHD == 0) %>%
+   select(currentSmoker, TenYearCHD) %>%
+ summarise(nonsmoker = sum(currentSmoker))

  nonsmoker
1      1705

> # The number of non-smokers with absolute short-term CHD Risk
> Cleaned.CHD.Data %>%
+   filter(currentSmoker == 0 & TenYearCHD == 1) %>%
+   select(currentSmoker, TenYearCHD) %>%
+   summarise(sum_non_CHD = sum(TenYearCHD))

  sum_non_CHD
1         300
```

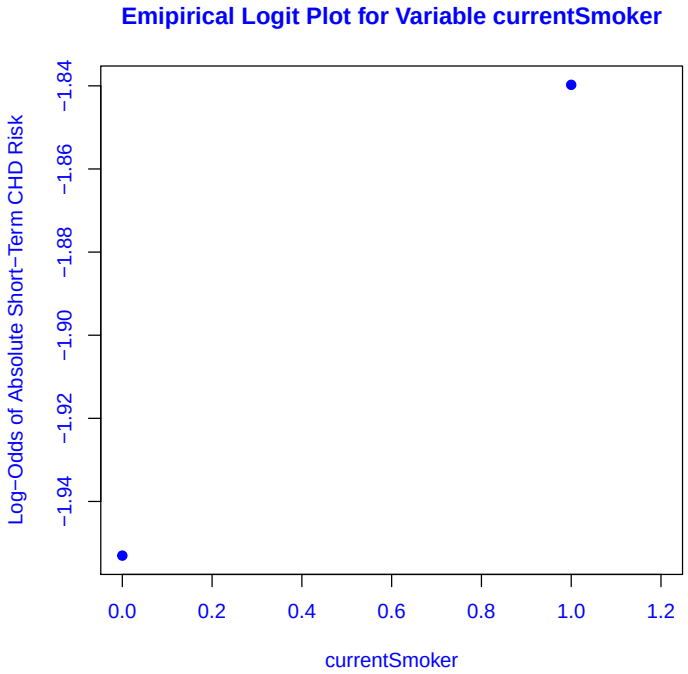
```
> # The number of non-smokers without CHD Risk is
> Cleaned.CHD.Data %>%
+   filter(currentSmoker == 0, TenYearCHD == 0) %>%
+ select(currentSmoker) %>%
+   summarise(count.non.smoker = n())

  count.non.smoker
1             1815

> # As a check
> 322 + 1705 + 300 + 1815

[1] 4142

> # The empirical logit value is computed with the following
>
> r_s <- 322 # Number of smokers with absolute short-term CHD risk
> t_s <- 1705 # Number of smokers without absolute short-term CHD risk
> r_non <- 300 # Number of non-smokers with absolute short-term CHD risk
> t_non <- 1815 # Number of non-smokers without absolute short-term CHD risk
> # Then the empirical logit values are
>
> empirical.logit.smokers <- log(r_s/(r_s + t_s), base = exp(1))
> empircal.logit.non.smokers <- log(r_non/(t_non + r_non), base = exp(1))
> currentSmoker.levels <- c(0, 1)
> empirical.logit.values <- c(empircal.logit.non.smokers, empirical.logit.smokers)
>
>
```



As we can see linearity is automatic for this type of variable.

Case II: Empirical Logit Plots for Continuous Variables

We have produced an empirical logit plot to examine the relationship between the empirical logit and *currentSmoker* variable in our data set. Here we produce an empirical logit plot for the continuous *age* variable. To produce an empirical logit plot for any continuous variable we follow the steps below:

I. Identify the sample size, sort the values of the predictor in ascending order, and then break the sorted values of the predictor into $i^* = 1, 2, ..., a$ intervals. Each interval containing about an equal number of observations u_{i^*}

II. Then compute $\frac{1}{u_{i^*}} \sum_{v_{i^*}=1}^{u_{i^*}} X_{k,v_{i^*}}$ for predictor k in interval i^*. *I.e.*, the sample mean of the organized continuous information in predictor k that lies in interval i^* which contains $v_{i^*} = 1, 2, ..., u_{i^*}$ observations

III. Compute the *observed proportion of successes* in each interval i^*. Let the proportion of successes in interval i^* be $\hat{p_{i^*}}$. $\hat{p_{i^*}} = \frac{1}{u_{i^*}} \sum_{v_{i^*}=1}^{u_{i^*}} Y_{v_{i^*}}$ where the values of $Y_{v_{i^*}}$ are either 0 or 1 in each interval

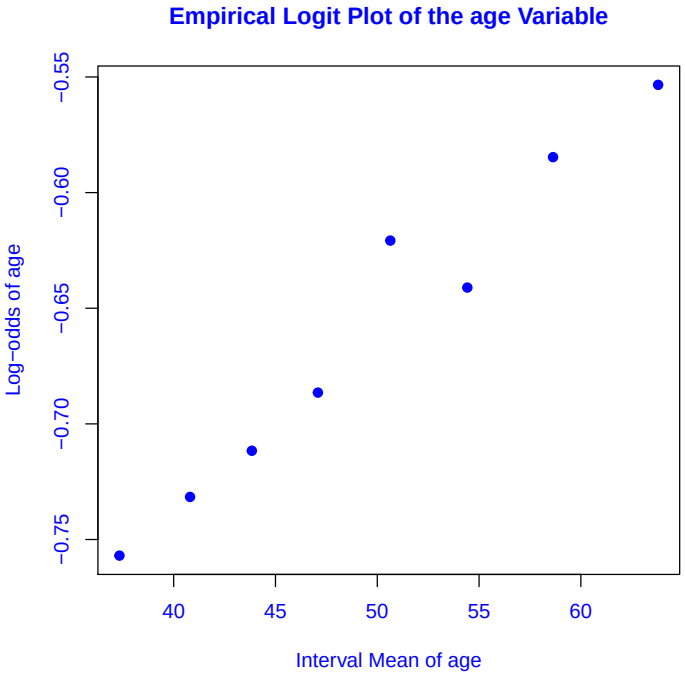
IV. Compute $\ln(\frac{\hat{p_{i^*}}}{1-\hat{p_{i^*}}})$ for each i^*

V. Construct the empirical logit plot for the continuous predictor. This is a *scatter plot* of $\ln(\frac{\hat{p_{i^*}}}{1-\hat{p_{i^*}}})$ *vs.* $\frac{1}{u_{i^*}}\sum_{v_{i^*}=1}^{u_{i^*}} X_{k,v_{i^*}}$ for each i^*. *I.e., ordered pairs* $(\frac{1}{u_{i^*}}\sum_{v_{i^*}=1}^{u_{i^*}} X_{k,v_{i^*}}\ , \ln(\frac{\hat{p_{i^*}}}{1-\hat{p_{i^*}}}))$

[1]. These instructions are a very "mathematized" version of the approach given by *Cannon Cobb, et al.* in *STAT2: Building Models for a World of Data.* One final note, when n is large it is beneficial to have $a \geq 3$ in order to obtain insight on any departure of linearity, direction, and magnitude of the relationship between the empirical logit and the continuous predictor [1]. We produce an empirical logit plot for the age variable using R below. Code is provided to show how we create the plot.

```
> # Sort the information in age in ascending
> sorted.age <- order(Cleaned.CHD.Data$age, decreasing = FALSE)
> sorted.CHD.Data <- Cleaned.CHD.Data[sorted.age,]
> m1 <- mean(sorted.CHD.Data[1:518,2])
> m2 <- mean(sorted.CHD.Data[519:1037,2])
> m3 <- mean(sorted.CHD.Data[1038:1556,2])
> m4 <- mean(sorted.CHD.Data[1557:2074,2])
> m5 <- mean(sorted.CHD.Data[2075:2593,2])
> m6 <- mean(sorted.CHD.Data[2594:3112,2])
> m7 <- mean(sorted.CHD.Data[3113:3631,2])
> m8 <- mean(sorted.CHD.Data[3632:4142,2])
> p1 <- (sum(sorted.CHD.Data[1:518, 35]) )/(length(sorted.CHD.Data[1:518,35]))
> p2 <- (sum(sorted.CHD.Data[519:1037,35]) )/(length(sorted.CHD.Data[519:1037,35]))
> p3<-(sum(sorted.CHD.Data[1038:1556,35]) )/(length(sorted.CHD.Data[1038:1556,35]))
> p4<-(sum(sorted.CHD.Data[1557:2074,35]) )/(length(sorted.CHD.Data[1557:2074,35]))
> p5<-(sum(sorted.CHD.Data[2075:2593,35]) )/(length(sorted.CHD.Data[2075:2593,35]))
> p6<-(sum(sorted.CHD.Data[2594:3112,35]) )/(length(sorted.CHD.Data[2594:3112,35]))
> p7<-(sum(sorted.CHD.Data[3113:3631,35]) )/(length(sorted.CHD.Data[3113:3631,35]))
> p8<-(sum(sorted.CHD.Data[3632:4142,35]) )/(length(sorted.CHD.Data[3632:4142,35]))
> l1 <- log(p1, base = exp(1))/(1 - log(p1, base = exp(1)))
> l2 <- log(p2, base = exp(1))/(1 - log(p2, base = exp(1)))
> l3 <- log(p3, base = exp(1))/(1 - log(p3, base = exp(1)))
> l4 <- log(p4, base = exp(1))/(1 - log(p4, base = exp(1)))
> l5 <- log(p5, base = exp(1))/(1 - log(p5, base = exp(1)))
> l6 <- log(p6, base = exp(1))/(1 - log(p6, base = exp(1)))
> l_7 <- log(p7, base = exp(1))/(1 - log(p7, base = exp(1)))
> l8 <- log(p8, base = exp(1))/(1 - log(p8, base = exp(1)))
> Interval.means <- c(m1,m2,m3,m4,m5,m6,m7, m8)
> Interval.logit <- c(l1,l2,l3,l4,l5,l6,l_7, l8)

> plot(x = Interval.means, y = Interval.logit,
+      xlab = "Interval Mean of age", ylab = "Log-odds of age",
+      main = "Empirical Logit Plot of the age Variable",
+      col = "blue", pch = 21, bg = "blue", col.lab = "blue",
+      col.axis = "blue", col.main = "blue")
```



The relationship between the empirical logit and the age variable shows a strong linear relationship. Additionally, the above code can be manipulated slightly to produce empirical logit plots for other continuous predictors in our data matrix. If one produces empirical logit plots for the variables *cigsPerDay, sysBP,* and *diaBP* in this data set, you will be able to verify that the relationship between these variables and their empirical logits is linear. The relationship between the *BMI* and the empirical logit is linear. But the relationship is weak. Each empirical logit plot was produced using *eight* intervals.

Predicting Absolute Short-Term CHD Risk Using Logistic Regression

In this section we model the probability of being labeled with absolute short-term risk of CHD by fitting and interpreting two kinds of multiple logistic regression models. Models with only categorical independent variables and models containing both continuous and categorical independent variables. Moreover, we select variables to include in each model based upon our research findings in the previous section titled "Definition of Coronary Heart Disease (CHD) and Major Risk Factors and Conditions" without utilizing variable selection methods (variable selection methods are discussed later). Additionally, we pick combinations of variables we think should include in our models such that each of

our choosings will not cause *multicollinearity* issues. For example, one may hypothesize that there is strong relationship between diastolic and systolic blood pressure. *I.e.*, the variables are dependent upon each other. Since this is the case, it is likely that including both of these variables in any logistic regression model, or categorical variables that represent both types of blood pressure, will likely cause multicollinearity issues. For a thorough discussion of collinearity and multicollinearity see "Applied Linear Regression" second edition by Sanford Weisberg.

Model I: Modeling with Continuous and Categorical Predictors

According to our research, we have seen that some of the major risk factors of CHD are advancing age, *sex-specific advancing age*, various levels of BMI, being diabetic, and smoking. Let's model absolute short-term risk of CHD risk for patients using the risk factors mentioned in the previous sentence. To present model nicely, we let $age = X.1$, $male = X.2$ ($X.2 = 1$ if patient is male. $X.2 = 0$ is female), $BMI = X.3$, $diabetes = X.4$ ($X.4 = 1$ if the patient is diabetic. $X.4 = 0$ if the patient is not), $Light.Smoker = X.5$ ($X.5 = 1$ if the patient is a light smoker. $X.5 = 0$ if the patient is not), $Moderate.Smoker = X.6$ ($X.6 = 1$ if the patient is a moderate smoker. $X.6 = 0$ if the patient is not), and $Heavy.Smoker = X.7$ ($X.7 = 1$ if the patient is a heavy smoker. $X.7 = 0$ if the patient is not), and have p_1 be the probability of labeling a patient with absolute short-term risk of CHD with this model. So our model can be written as

$$\ln(\frac{p_1}{1-p_1}) = \beta_0+\beta_1 X.1+\beta_2 X.2+\beta_3 X.3+\beta_3 X.3+\beta_4 X.4+\beta_5 X.5+\beta_6 X.6+\beta_7 X.7$$

. We use R to fit the model. Code is provided below.

```
> attach(Cleaned.CHD.Data)
> CHD.df.1 <- data.frame(age, male, BMI, diabetes,
+                        Light.Smoker, Moderate.Smoker, Heavy.Smoker, TenYearCHD)
> colnames(x = CHD.df.1) <- c("X.1", "X.2", "X.3", "X.4",
+                             "X.5", "X.6", "X.7", "Risk")
> Model.I <- glm(Risk ~. , data = CHD.df.1, family = binomial)
> # The family argument tells the glm function the response we are modeling
> # is a binomial response.
>
> summary(Model.I)

Call:
glm(formula = Risk ~ ., family = binomial, data = CHD.df.1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-1.4180  -0.6050  -0.4430  -0.3138   2.6683

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.364189   0.435012 -16.929  < 2e-16 ***
X.1          0.083053   0.005748  14.448  < 2e-16 ***
X.2          0.410093   0.096110   4.267 1.98e-05 ***
X.3          0.034548   0.011003   3.140 0.001691 **
X.4          0.784731   0.219410   3.577 0.000348 ***
X.5         -0.023000   0.190619  -0.121 0.903961
X.6          0.433897   0.111625   3.887 0.000101 ***
X.7          0.780114   0.148397   5.257 1.46e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3504.1  on 4141  degrees of freedom
Residual deviance: 3192.6  on 4134  degrees of freedom
AIC: 3208.6

Number of Fisher Scoring iterations: 5
```

. The only insignificant variable is $X.5$. All others are significant. The estimated model in probability form is

$$\hat{p_1} = \frac{e^{(-7.364+0.083X.1+0.410X.2+0.035X.3+0.785X.4-0.023X.5+0.434X.6+0.780X.7)}}{1+e^{(-7.364+0.083X.1+0.410X.2+0.035X.3+0.785X.4-0.023X.5+0.434X.6+0.780X.7)}}$$

. As we previously discussed in the section of this paper called "A Brief Discussion of the Logistic Regression Model" we use the concept of probability to label which category an observation X_n, or an unseen observation (not in the training set), belongs to [**?**]. When we use the model to label any observation based on its feature values we label the observation as 1 if $p_1 > c$ and 0 otherwise where c is a *posterior probability threshold* [**?**]. There is no specific value of c that must be used in a classification problem [**?**]. In fact, depending on the application, c can be varied to *minimize* model *error rates* or *maximize* rates at which it correctly assigns observations to a certain class that are contained in Y [**?**]. We assess this model by producing and analyzing a *receiver operating characteristics curve* (*ROC curve*) and compute the area under the curve for this model [**?**]. Again R can help us do this and also return a probability threshold that yields a *maximized sensitivity* and *specificity*. We produce a ROC curve and provide code below.

```
> library(pROC)
> attach(CHD.df.1)
```

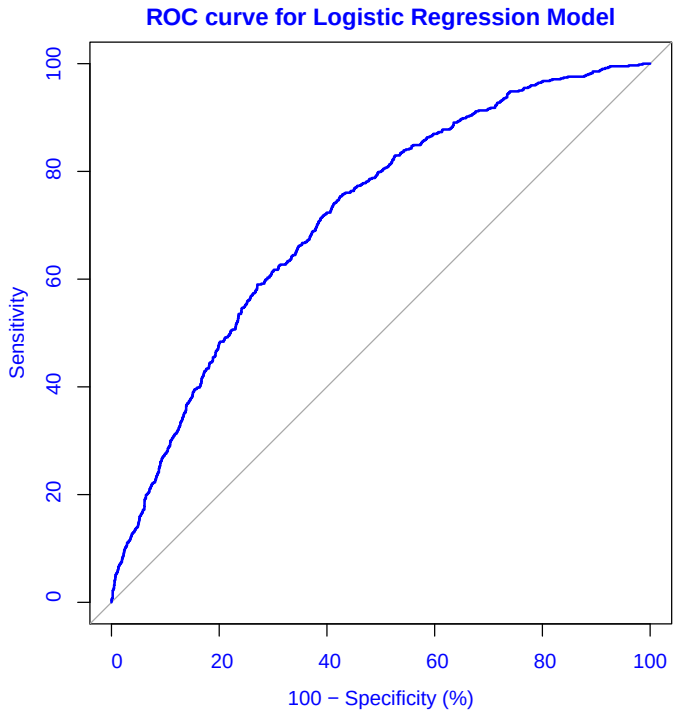
```
> Model.I.probs <- predict(object = Model.I,
+                          newdata = CHD.df.1,
+                          type = "response")
> # Create a ROC curve using the entire sample
>
> par(pty = "s")
> Model.I.ROC <- roc(response = Risk,
+                    predictor = Model.I.probs, percent = TRUE,
+                    plot = TRUE, legacy.axes = TRUE,
+                    ylab = "Sensitivity", col = "blue",
+                    col.axis = "blue", col.lab = "blue",
+                    col.main = "blue",
+                    main = "ROC curve for Logistic Regression Model")
> Model.I.ROC$auc

Area under the curve: 71.53%

> coords(roc = Model.I.ROC, x = "best", best.method = "youden")

  threshold specificity sensitivity
  0.1289227  57.5284091  75.4019293

>
> # When x is set to "best" in the coords function, and best.method is set  # to youden,
> # the coords function returns a posterior probability
> # threshold that maximizes the classifier's sensitivity and specificity.
>
```



We see that the optimized sensitivity of this model is 75.40%, the optimized specificity is 57.53%, and the area under the ROC curve is 71.53%. But this is the *training sensitivity* and *training specificity* which is usually an overestimate of these metrics [**?**]. Also, we see that the probability threshold that maximizes them is 0.1289227. So we use this threshold and a $50 - 50$ split of the data to estimate the test sensitivity, specificity, and accuracy.

```
> set.seed(16)
> CHD.df.1.train <- sample(x = 4142, size = 2071, replace = FALSE)
> CHD.df.1.test <- CHD.df.1[-CHD.df.1.train,]
> # Estimate the model on the training set
>
> train.set.Model.I <- glm(Risk ~. , data = CHD.df.1,
+                                    family = binomial,
+                          subset = CHD.df.1.train)
> # Predict absolute short term risk of CHD on unseen cases
>
> Model.I.validation.probs <- predict(object = train.set.Model.I,
+                                     newdata = CHD.df.1.test,
+                                     type = "response")
> # Create a vector of labels and use probability
> # threshold to name successes
>
```

```
>
> Model.I.validation.predictions <- rep(x = 0, times = 2071)
> Model.I.validation.predictions[Model.I.validation.probs >= 0.1289227] <- 1
> # Produce a confusion matrix
>
> table(Model.I.validation.predictions, CHD.df.1.test[,8])

Model.I.validation.predictions   0   1
                             0 982  86
                             1 786 217

>
>
```

The estimated test sensitivity is $100 * (\frac{217}{217+86})\% = 71.62\%$, the estimated test specificity is $100 * (\frac{982}{982+786})\% = 55.54\%$, and the estimated test accuracy is $100 * (\frac{217+982}{2071})\% = 57.89\%$.

Model II: Modeling with Categorical Predictors

In this section, we build a model that uses all of the major risk factors previously discussed. Furthermore, the variables we created have been given a binary coding which represents various levels of each *continuous risk factor*. *E.g.*, $X = 1$ if $25 \leq BMI < 29.9999$ representing the condition of obesity and $X = 0$ otherwise when the patient's BMI doesn't fall within this interval. For this model, the variables are as follows:

- $X.1 = 1$ if the patient is male. $X.1 = 0$ if female.
- $X.2 = 1$ if the patient is female and 55 or older. $X.2 = 0$ if female and younger than 55.
- $X.3 = 1$ if the patient is male and 45 or older. $X.3 = 0$ if male and younger than 45.
- $X.4$ is the dummy variable that represents whether or not the patient is a current smoker. $X.4 = 1$ if yes $X.4 = 0$ if they are not
- $X.5$ is the dummy variable that represents the presence of diabetes. It is defined the same as in model I
- $X.6 = 1$ if the patient is overweight. $X.6 = 0$ if the patient has a BMI that doesn't fall into the definition of being overweight
- $X.7 = 1$ if the patient is obese. $X.7 = 0$ if the patient has a BMI that doesn't fall into the definition of being obese
- $X.8 = 1$ if the patient has borderline high total cholesterol. $X.8 = 0$ if they do not

- $X.9 = 1$ if the patient has very high total cholesterol. $X.9 = 0$ if they do not
- $X.10 = 1$ if the patient has high systolic blood pressure. $X.10 = 0$ if they do not
- $X.11 = 1$ if the patient has hypertensive systolic blood pressure $X.11 = 0$ if they do not.
- $X.12 = 1$ if the patient has advanced hypertensive systolic blood pressure. $X.12 = 0$ if they do not

. For a further explanation of our dummy variable coding schemes please see the project's data dictionary available to view or download from the Wright Analytics website. Using R code that is similar to the code used to produce model I we train model II on the entire data set. The summary of the logistic regression output is provided below along with a confusion matrix. The distribution of the data in the test and training set used to produced the confusion matrix is again a 50-50 split.

```
Call:
glm(formula = Risk ~ ., family = binomial, data = CHD.df.2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3602  -0.6166  -0.4126  -0.3251   2.5423

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.19146    0.16099 -19.823  < 2e-16 ***
X.1          0.20430    0.17440   1.171 0.241410
X.2          1.01229    0.14131   7.164 7.86e-13 ***
X.3          1.05411    0.15897   6.631 3.34e-11 ***
X.4          0.32526    0.09689   3.357 0.000788 ***
X.5          0.77409    0.21963   3.525 0.000424 ***
X.6          0.04403    0.10345   0.426 0.670382
X.7          0.17214    0.14204   1.212 0.225558
X.8          0.17191    0.14313   1.201 0.229706
X.9          0.27766    0.13874   2.001 0.045362 *
X.10         0.24208    0.13329   1.816 0.069327 .
X.11         0.59873    0.12176   4.917 8.78e-07 ***
X.12         1.03788    0.13689   7.582 3.40e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

```
    Null deviance: 3504.1  on 4141  degrees of freedom
Residual deviance: 3188.6  on 4129  degrees of freedom
AIC: 3214.6

Number of Fisher Scoring iterations: 5

Area under the curve: 71.59%

  threshold specificity sensitivity
   0.178634   71.590909   60.932476

Model.II.validation.predictions    0    1
                              0 1254  143
                              1  495  179
```

The *estimated sensitivity* of the model using the validation set approach with a 50-50 split is $100 * (\frac{179}{179+143})\% = 55.56\%$ which is significantly smaller than the training sensitivity obtained which is 60.93%. Notice that the conditional probability threshold for this model is much bigger than the conditional probability threshold for model I. Again, there is no one conditional probability threshold that must be used. If we were to use the conditional probability threshold from model I with model II the estimated sensitivity is

```
Model.II.validation.predictions    0    1
                              0 1017   92
                              1  732  230
```

$100 * (\frac{230}{230+92})\% = 71.43\%$. On another topic, model II has many more predictor coefficients that have *not* been found to be significantly different from zero. Does this mean that these risk factor categories have no impact on absolute short-term CHD risk for a patient? No. Probably not. However, we do not have enough numerical evidence that they *are* significantly different from zero. Thus, it makes no sense to interpret them.

Model III: A Well-Balanced Model

In this section, we model the event of patients being labeled with absolute short-term CHD risk with a logistic regression model that contains continuous and categorical predictors. The independent variables we include are $male = X.1$, $age = X.2$, $Age.Risk.Men = X.3$, $Age.Risk.Women = X.4$, $BMI = X.5$, and $diabetes = X.6$, $Light.Smoker = X.7$, $Moderate.Smoker = X.8$, $Heavy.Smoker = X.9$, $cigsPerDay = X.10$, and $sysBP = X.11$. Using R, the summary of the model is

```
Call:
glm(formula = Risk ~ ., family = binomial, data = CHD.df.3)

Deviance Residuals:
```

```
    Min       1Q   Median       3Q      Max
-1.5345  -0.5994  -0.4203  -0.2932   2.8073

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.679131   0.550285 -13.955  < 2e-16 ***
X.1          0.352412   0.184522   1.910  0.05615 .
X.2          0.054380   0.009272   5.865 4.49e-09 ***
X.3          0.358607   0.205415   1.746  0.08085 .
X.4          0.304328   0.183371   1.660  0.09699 .
X.5          0.006854   0.011636   0.589  0.55583
X.6          0.702127   0.222306   3.158  0.00159 **
X.7         -0.071963   0.197078  -0.365  0.71500
X.8          0.223370   0.232861   0.959  0.33744
X.9          0.329608   0.440297   0.749  0.45410
X.10         0.012323   0.012086   1.020  0.30794
X.11         0.017228   0.002123   8.116 4.83e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3504.1  on 4141  degrees of freedom
Residual deviance: 3120.5  on 4130  degrees of freedom
AIC: 3144.5

Number of Fisher Scoring iterations: 5
```

. The area under the ROC curve is

```
Area under the curve: 73.4%

  threshold specificity sensitivity
  0.1592199  69.6022727  67.3633441
```

73.40% which is an improvement compared to model I and II in terms of this metric. Additionally, we see that a posterior probability threshold of 0.1592199, that optimizes sensitivity, returns a training sensitivity of 67.36% but *also* a training *specificity* of 69.60%. Specificity is the rate at which the model labels true events of failure as events of failure. A confusion matrix for this model, using a 50-50 split, is

```
Model.III.validation.predictions    0    1
                               0 1242  124
                               1  507  198
```

. The estimated test sensitivity is $100 * (\frac{198}{198+124})\% = 61.49\%$. The estimated test specificity is $100 * (\frac{1242}{1242+507})\% = 71.01\%$. The estimated error rate using

this threshold is $100 * (1 - \frac{1242+198}{2071})\% = 30.47\%$. Since the training sensitivity and specificity are well balanced, the AUC of the ROC curve is higher compared to models I and II, and the estimated test sensitivity is lower than the training sensitivity we conclude that model III is a well-balanced model despite the fact that there may be slight collinearity issues present. This is seen by the negative slope for $Light.Smoker$.

In this section, we trained three logistic regression models that contained various combinations of variables that represented major risk factors of CHD. The model that had the highest estimated test sensitivity was model I. The area under model III's ROC curve was the largest. Moreover, we decided that model three was well balanced because there was no significant difference between its training specificity and sensitivity. Also, its training accuracy is 69.53%.

Feature Selection for the Logistic Regression Model

In the section "Predicting Absolute Short-Term CHD Risk Using Logistic Regression" we built three models using predictor variables that we hand-picked based on our research findings. We engineered some of these variables to represent certain levels of major CHD risk factors (*e.g.*, we coded categorical variables to represent certain patient conditions such as being overweight or obese). The data set that we have been analyzing has thirty-four variables that we can consider adding to logistic regression models to model the probability of $Y = 1$. How do we know if we included the most *relevant* variables in the models we fit in the previous section? We don't. We made an educated guess what variables *should* be included in the model based on our research findings. Since we have many variable options, we discuss *feature selection* (variable selection) methods in this section and apply it to this application to see if we can uncover a more effective model compared to the ones that we have already obtained. One algorithm we can use is the *Boruta Algorithm*. We do not explain how this algorithm performs feature selection in-depth because the explanation is beyond the scope of this paper. We provide R code for the Boruta algorithm below.

```
> # Implementation of the Boruta Algorithm for feature selection
> library(Boruta)
> Boruta.Search <- Boruta(TenYearCHD ~., data = Cleaned.CHD.Data,
+                         doTrace = 0, pValue = 0.01)
>

> Boruta.Search.Significants <- getSelectedAttributes(x = Boruta.Search,
> # List the significant variables using a significance level of 0.01
>
> print(Boruta.Search.Significants)

 [1] "age"                   "Age.Risk.Women"        "Age.Risk.Men"
 [4] "Heavy.Smoker"          "cigsPerDay"            "prevalentHyp"
```

```
 [7] "diabetes"              "Very.High.TC."         "totChol"
[10] "High.BP"               "Hyper.sys.BP"          "Hyper.dia.BP"
[13] "Hypertension"          "Advanced.Hyper.sys"    "Advanced.Hyper.dia"
[16] "Advanced.Hypertension" "sysBP"                 "diaBP"
[19] "Overweight"            "Obesity"               "BMI"
```

```
> # Use R to uncover and return variable importance from the search
>
> Boruta.Var.Importance.1 <- attStats(x = Boruta.Search)
> Boruta.Var.Importance.2 <- Boruta.Var.Importance.1[Boruta.Var.Importance.1$decision != c(
> # Return importances in data frame
>
> Boruta.Var.Importance.2[order(-Boruta.Var.Importance.2$meanImp,
+                               decreasing = FALSE),
+                         ]
```

```
                        meanImp  decision
sysBP                 22.679823 Confirmed
diaBP                 20.755834 Confirmed
age                   18.391570 Confirmed
prevalentHyp          13.992514 Confirmed
totChol                9.619838 Confirmed
BMI                    9.321212 Confirmed
Advanced.Hypertension  8.996512 Confirmed
Hypertension           8.787567 Confirmed
Advanced.Hyper.sys     8.169908 Confirmed
diabetes               6.965158 Confirmed
Age.Risk.Men           6.892927 Confirmed
Advanced.Hyper.dia     6.330359 Confirmed
Hyper.sys.BP           6.292310 Confirmed
cigsPerDay             6.269478 Confirmed
Very.High.TC.          5.762492 Confirmed
Age.Risk.Women         5.752920 Confirmed
Obesity                4.319457 Confirmed
Heavy.Smoker           4.095624 Confirmed
Hyper.dia.BP           3.961423 Confirmed
Overweight             3.278796 Confirmed
High.BP                3.187701 Confirmed
Normal.Body.Weight     3.168324 Tentative
male                   2.671110 Tentative
heartRate              2.653958 Tentative
ClassOne.Obesity       2.437533 Tentative
High.sys.BP            2.221297 Tentative
Moderate.Smoker        1.856407 Tentative
```

```
> # Perform variable importance search on a full model and return variable importance to
> # get an idea of what should be kept in the model
```

```
> library(caret)
> library(arm)
> library(VGAM)
> set.seed(100)
> # Method is bayesglm for regression and classification
> model.full <- train(TenYearCHD ~., data = Cleaned.CHD.Data,
+                     method = "bayesglm")
> model.importance <- varImp(model.full)
> print(model.importance)

loess r-squared variable importance

  only 20 most important variables shown (out of 34)

                     Overall
age                  100.000
sysBP                 88.396
prevalentHyp          58.443
Age.Risk.Men          49.723
Advanced.Hyper.sys    45.188
Advanced.Hypertension 42.903
diaBP                 40.320
Advanced.Hyper.dia    32.631
Age.Risk.Women        19.795
male                  16.300
diabetes              15.538
Hypertension          14.947
totChol               14.320
BMI                   13.054
Hyper.sys.BP          11.195
Normal.Body.Weight     9.230
Very.High.TC.          8.578
Hyper.dia.BP           7.640
cigsPerDay             7.251
Heavy.Smoker           6.752

>
```

The Boruta search Algorithm returns 22 variables and labels them as confirmed meaning that we should consider adding some or all of them in a logistic regression model. The variable importance method returns fifteen variables that have importance rankings above the value of 10. Moreover, twenty of the most important variables out of thirty-four that we should consider adding in a logistic regression model. Next, we fit two models using *variable selection recommendations* obtained by using these selection methods. One model aims to be an improvement of model III. It contains the original variables contained in model III and other significant/relevant variables the selection methods recommended.

The second model we train in this section contains mostly categorical variables which are recommended as most relevant/significant. Note, both methods are supervised search methods because we have specified that *TenYearCHD* depends on the remaining variables in R. Note, we have renamed *TenYearCHD* as *Risk* in the previous models and continue to use this name for the dependent variable. The models are found below along with their AUCs and confusion matrices computed using a 50-50 split.

For the first model, the one that aims to improve model III, the subset of variables is as follows: $male = X.1$, $age = X.2$, $Age.Risk.Men = X.3$, $Age.Risk.Women = X.4$, $BMI = X.5$, $diabetes = X.6$, $cigsPerDay = X.7$, $currentSmoker = X.8$, $sysBP = X.9$, $prevalentHyp = X.10$, $totChol = X.11$, and $prevalentStroke = X.12$. The estimated model using all of the data is

```
Call:
glm(formula = Risk ~ ., family = binomial, data = CHD.df.4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4021  -0.5978  -0.4150  -0.2879   2.7961

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.651663   0.632726 -12.093  < 2e-16 ***
X.1          0.354895   0.185473   1.913 0.055689 .
X.2          0.053010   0.009300   5.700 1.20e-08 ***
X.3          0.361367   0.205614   1.758 0.078832 .
X.4          0.265776   0.183617   1.447 0.147771
X.5          0.003468   0.011700   0.296 0.766943
X.6          0.713541   0.221758   3.218 0.001292 **
X.7          0.020536   0.005786   3.549 0.000387 ***
X.8          0.046869   0.147553   0.318 0.750756
X.9          0.014065   0.002737   5.139 2.76e-07 ***
X.10         0.197996   0.127667   1.551 0.120929
X.11         0.002011   0.001040   1.933 0.053203 .
X.12         0.901467   0.464089   1.942 0.052083 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3504.1  on 4141  degrees of freedom
Residual deviance: 3111.8  on 4129  degrees of freedom
AIC: 3137.8

Number of Fisher Scoring iterations: 5
```

. Next, we compute the optimal threshold, area under the ROC curve, and a confusion matrix for this model using a 50-50 split (50-50 split for the matrix only).

```
  threshold specificity sensitivity
   0.141433   64.034091   73.311897

Area under the curve: 73.67%

Model.IV.validation.predictions    0    1
                              0 1193   88
                              1  591  199
```

For the second model, the subset of variables is as follows: $male = X.1$, $age = X.2$, $Age.Risk.Men = X.3$, $Age.Risk.Women = X.4$, $prevalentHyp = X.5$, $prevalentStroke = X.6$, $Advanced.Hyper.sys = X.7$, $Hyper.sys.BP = X.8$, $High.sys.BP = X.9$, $Normal.Body.Weight = X.10$, $Overweight = X.11$, $Obesity = X.12$, $Borderline.High.TC = X.13$, $Very.High.TC = X.14$, $currentSmoker = X.15$, and $cigsPerDay = X.16$. The estimated model using all of the data is

```
Call:
glm(formula = Risk ~ ., family = binomial, data = CHD.df.5)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3933  -0.6074  -0.4167  -0.2943   2.6844

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.552745   0.616934  -9.001  < 2e-16 ***
X.1          0.345256   0.185970   1.857 0.063381 .
X.2          0.057194   0.009228   6.198 5.73e-10 ***
X.3          0.344265   0.205264   1.677 0.093508 .
X.4          0.271847   0.183012   1.485 0.137437
X.5          0.344108   0.140827   2.443 0.014547 *
X.6          0.905795   0.465361   1.946 0.051602 .
X.7          0.613548   0.185126   3.314 0.000919 ***
X.8          0.259594   0.158692   1.636 0.101874
X.9          0.148136   0.136917   1.082 0.279278
X.10        -0.214673   0.428971  -0.500 0.616767
X.11        -0.138515   0.430235  -0.322 0.747490
X.12        -0.036071   0.441580  -0.082 0.934897
X.13         0.124484   0.144687   0.860 0.389589
X.14         0.202399   0.140125   1.444 0.148621
X.15         0.036308   0.146957   0.247 0.804857
```

```
X.16         0.020568   0.005770   3.565 0.000364 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3504.1  on 4141  degrees of freedom
Residual deviance: 3140.2  on 4125  degrees of freedom
AIC: 3174.2

Number of Fisher Scoring iterations: 5
```

. Next, we compute the optimal threshold, area under the ROC curve, and a confusion matrix for this model using a 50-50 split (50-50 split for the matrix only).

```
  threshold specificity sensitivity
  0.1695275  70.7954545  64.4694534

Area under the curve: 72.89%

Model.V.validation.predictions    0    1
                             0 1304  140
                             1  432  195
```

Discussion

In this paper we gave a brief discussion of the logistic regression model, listed major CHD risk factors, provided a history of the Framingham Heart study and data set components, conducted data visualization, provided routines for creating empirical logit plots for continuous and categorical predictors, estimated models, analyzed ROC curves and confusion matrices, and finally, performed variable selection and fit models on subsets of those recommendations. Model I was fit considering most major CHD risk factors. However, it does not account for the additional impact of being female and fifty years and older or male and forty-five years and old. It also ignores cholesterol level. Models II and III consider all major risk factors. The models we fit later consider all major risk factors represented by a combination of continuous and categorical variables (the models we fit after making use of the Boruta search algorithm and variable importance to aid us in variable selection). Moreover, these models incorporate patient medical history since *prevalentHyp* and *prevalentStroke* are independent variables in both. We then estimated test accuracy, sensitivity and specificity for models I, II, and III using the validation set approach with 50-50 splits. These rates we found using a posterior probability threshold that optimized model sensitivity and specificity. However, these thresholds did vary for each model. In the remainder of this section, we compute AUCs for each

model using validation sets to make a final choice on which model we wish to choose to model absolute short-term CHD risk for the population of Framingham, Massachusettes. The AUCs under the ROC curves for model I, model II, model III, model IV, and model V computed on validation sets containing 2071 test observations are shown belown in order.

```
Area under the curve: 69.43%

Area under the curve: 68.75%

Area under the curve: 71.11%

Area under the curve: 74.28%

Area under the curve: 51.43%
```

A large differences between the AUC for model IV and V is seen in the above output when evaluating their ROC curves on test sets. The difference is $74.28\% - 51.43\% = 22.85\%$. The AUC under the ROC curve for model IV is 74.28% which is a good indicator that this classifier is the most effective regardless of the threshold value used. Thus, we conclude that model IV was the best classifier we were able to find. on another topic, the threshold that optimizes model IV's sensitivity and specificity is 0.141433. Model IV's training sensitivity is 73.31% specificity is 64.03%. Model IV's estimated test sensitivity is $100 * \frac{199}{199+88}\% = 69.34\%$, estimated test specificity is $100 * \frac{1193}{1193+591} = 66.87\%$, and the estimated error rate is $100 * (1 - \frac{199+1193}{2071}) = 32.79\%$. To conclude this paper, we predict whether or not a male who is fifty years old, has a BMI of 30 $\frac{kg}{m^2}$, is diabetic, smokes twenty cigarettes per day (on average), has a systolic blood pressure reading of 150.0 in $mm\ Hg$ (is hypertensive), has a total cholesterol reading of 128.0 $\frac{mg}{dL}$, and has a history of stroke is likely to develop CHD within the next ten years. Then, estimate model metrics using a different threshold value. In general, Model IV can be written as

$$\begin{aligned} ln(\frac{\hat{p}}{1-\hat{p}}) = -7.652 + 0.355X.1 + 0.053X.2 + 0.361X.3 + 0.266X.4 \\ + 0.003X.5 + 0.713X.6 + 0.021X.7 + 0.047X.8 + 0.014X.9 \\ + 0.198X.10 + 0.002X.11 + 0.901X.12 \end{aligned}$$

in logit form. Plugging in the patient values yields $\ln(\frac{\hat{p}}{1-\hat{p}}) = 0.241$ and the probility of CHD development within ten years is $\frac{1}{1+e^{-(0.241)}} = 0.56$ or a 56.0% chance. Using R, the estimated probability is

```
> predict(object = Model.IV, newdata = data.frame(X.1 = 1, X.2 = 50, X.3 = 1,X.4 = 0,X.5=30,
+ X.6 = 1,X.7 = 20, X.8 = 1, X.9 = 150,
+ X.10 = 0, X.11 = 128, X.12 = 1),
+ type = "response")
```

```
        1
0.5643516
```

. The probability is much higher than the threshold that optimizes model IV's sensitivity and specificity. However, this threshold is only a recommendation. If we would like to have the model to have a higher sensitivity, we can change the threshold from 0.141433 to 0.10. The confusion matrix for model IV produced using the validation set approach and a threshold of 0.10 is found below.

```
> Model.IV.validation.predictions <- rep(x = 0, times = 2071)
> Model.IV.validation.predictions[Model.IV.validation.probs > 0.10] <-1
> table(Model.IV.validation.predictions, CHD.df.4.test$Risk)

Model.IV.validation.predictions   0   1
                              0 957  56
                              1 827 231

>
```

Under these circumstances, the estimated test sensitivity is $100 * \frac{231}{231+56}\% = 80.49\%$, the estimated specificity is $100 * \frac{957}{957+827}\% = 53.64\%$, and the estimated overall accuracy is $100 * \frac{957+231}{2071}\% = 57.36\%$. This has obviously caused a trade-off. However, under these circumstances, predictions using this threshold become less dangerous but also produces significantly more *false positive* predictions. Nevertheless, we still rely on model IV to classify whether or not patients in Framingham, Massachusettes are at risk of suffering from CHD within ten years.

References

[1] Ann R. Cannon, George W. Cobb, Bradley A. Hartlaub, Julie M. Legler, Robin H. Lock, Thomas L. Moore, Allan J. Rossman, and Jeffrey A. Witmer. *STAT2: building models for a world of data*. W.H. Freeman, 2013.

[2] Sanford Weisberg. *Applied linear regression*. John Wiley and Sons, 1985.

[3] Abraham Weishaus. *SOA Exam P study manual*. Actuarial Study Materials, 2017.

[4] Morris H. DeGroot and Mark J. Schervish. *Probability and statistics*. Addison-Wesley, 2012.

[5] Nci dictionary of cancer terms. URL `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/coronary-heart-disease`.

[6] Peter W. F. Wilson, Ralph B. D'Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz, and William B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, Dec 1998. doi: 10.1161/01.cir.97.18.1837.

[7] Scott M Grundy, Richard Pasternak, Philip Greenland, Sidney Smith, and Valentin Fuster. Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations. *Journal of the American College of Cardiology*, 34(4):1348–1359, 1999. doi: 10.1016/s0735-1097(99)00387-3.

[8] Robert H. Eckel and Ronald M. Krauss. American heart association call to action: Obesity as a major risk factor for coronary heart disease. *Circulation*, 97(21):2099–2100, Feb 1998. doi: 10.1161/01.cir.97.21.2099.

[9] Anne M. Wolf and Graham A. Colditz. Current estimates of the economic cost of obesity in the united states. *Obesity Research*, 6(2):97–106, 1998. doi: 10.1002/j.1550-8528.1998.tb00322.x.

[10] Body mass index (bmi) in adults. URL `https://www.heart.org/en/healthy-living/healthy-eating/losing-weight/bmi-in-adults`.

[11] Aram V. Chobanian, George L. Bakris, Henry R. Black, William C. Cushman, Lee A. Green, Joseph L. Izzo, Daniel W. Jones, Barry J. Materson, Suzanne Oparil, Jackson T. Wright, and et al. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. *Hypertension*, 42(6):1206–1252, 2003. doi: 10.1161/01.hyp.0000107251.49515.c2.

[12] Rebecca E. Schane, Pamela M. Ling, and Stanton A. Glantz. The health effects of light and intermittent smoking. *D23. Recent Advances On Smoking Health Effects*, 2010. doi: 10.1164/ajrccm-conference.2010.181.1$_m$*eetingabstracts.a*5433.

[13] Executive summary of the third report of the national cholesterol education program (ncep) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii). *JAMA: The Journal of the American Medical Association*, 285(19):2486–2497, 2001. doi: 10.1001/jama.285.19.2486.

[14] Dileep. Logistic regression to predict heart disease, Jun 2019. URL `https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression`.

[15] Framingham heart study. URL `http://www.framingham.com/heart/profile.htm`.

[16] Jorge Bacallao Gallestey. Framingham heart study. URL `https://www.britannica.com/event/Framingham-Heart-Studay`.