A Statistical Analysis of Early Diabetes Indicators

Ashler Herrick and Branden Ciranni

December 01, 2020

1 Introduction

While Glycosated Hemoglobin Levels are usually taken as an indicator of Type 2 Diabetes, which other early indicators are present that make one more likely to develop the disorder? It is known that obesity is a key factor, but specifically which features are statistically significant to justify this claim? The aim of this analysis is to identify such features using measurements from a 1997 study conducted on 403 black Americans in Virginia who were interviewed with the goal of determining the prevalence of obesity, diabetes and other cardiovascular risk factors in the community.[1] While it has been suggested that the Waist/Hip ratio may be an accurate predictor, we test it to be significant, but find it weak compared to other, more powerful indicators.

Our analysis begins with an exploratory analysis of the data in order to gain a better understanding of all of our possible predictors. Section 2 provides a description of the data fields, and describes additional transformations that were performed. Following this in Section 3, we verify the claim that the Waist-Hip ratio is significantly different between the diabetic and non-diabetic groups using an Independent Samples T-test. Then in Section 4, we evaluate feature correlations to identify feature variables for input into our logistic regression model, and minimize the effects of Multicollinearity.

The final section includes the fitting of a logistic regression model to the dataset. We began with a model using six predictors: total cholesterol, stabilized glucose, high density lipoprotein age, weight to height ratio, and waist to hip ratio. We find that a model including stabilized glucose, total cholesterol, weight to height ratio and age provide the best Akaike Information-Criterion for the model. We then test the model on an out of sample testing data set to determine the power of the model as a classifier for patients at risk for diabetes. Finally, we examine the effect of both missing completely at random value and non-ignorable missing values.

2 Data Overview

The data is provided by Vanderbilt Biostatistics, and is from a study by Dr. John Schorling on the prevalence of obesity, diabetes, and other cardiovascular risk factors in the African

American community. It consists of 19 variables and 403 participants, all of whom are African Americans in Central Virginia. A summary of all fields in the raw data is provided below in the *Data Dictionary*.

Data Dictionary		
Feature	Type	Description
id	int	Unique Identification for study participant.
chol	double	Total Cholesterol Level.
stab.glu	double	Stabilized Glucose Level - For those with diabetes, the body does not produce enough insulin, which regulates blood sugar levels, leading to higher-than-normal levels of Glucose in the blood.
hdl	double	High Density Lipoprotein. Sometimes referred to as "Good Cholesterol", HDL particles are responsible for transporting cholesterol to the liver for removal from the body.
ratio	double	Cholesterol / HDL Ratio.
glyhb	double	Glycosated Hemoglobin level - the average glood glucose measurement over 3-4 months. Glycosated Hemoglobin greater than 7.0 is taken as a diagnosis for diabetes.
location	string	Geographical location in Virginia. Takes values in {"Buckingham", "Louisa"}
age	int	Age of study participant.
gender	categorical	Gender of study participant. Takes values in {"male", "female"}.
height	int	Height (in inches).
weight	int	Weight (in pounds).
frame	categorical (ordinal)	Description of build of study participant - Takes values in {"small", "medium", "large"}.
bp.1s	int	First Systolic Blood Pressure Reading.
bp.1d	int	First Diastolic Blood Pressure Reading.
bp.2s	int	Second Systolic Blood Pressure Reading. (Not Required).
bp.2d	int	Second Diastolic Blood Pressure Reading (Not Required).
waist	int	Waist Measurement (Inches)
hip	int	Hip Measurement (Inches)
time.ppn	int	Postprandial Time (Minutes) - Time since the last meal, before blood was drawn.

2.1 Exploratory Analysis and Transformations

In order to identify risk factors for diabetes, we first create an indicator variable for whether a study participant has diabetes. We use the definition mentioned in the data dictionary, and create the boolean column `has_diabetes = data$glyhb > 7.0`. Given this, we note that the data is heavily skewed towards a negative diabetes diagnosis, with $n_{neg} = 330$, and $n_{pos} = 60$. This is acknowledged later in our Logistic Regression Model. Following this, we inspect the data for null values, finding 65% null values in the `bp.2s` and `bp.2d` columns. These are dropped, and the remaining columns are kept with negligible null values which are ignored during aggregate calculations. Finally, we map any ordinal categorical columns to ordered integers so they may later be included in correlation calculations. The only such column is `frame`, and the values [small, medium, large] are mapped to [1, 2, 3]. Following this, the distributions of each feature are plotted and explored, with more detail on this in the accompanying `analysis.Rmd` notebook.[3]

3 Testing Waist/Hip Ratio as a Significant Predictor

It has been suggested by Vanderbilt Statistics that the Waist/Hip Ratio may be a good indicator of diabetes. We test this claim using a two-population independent samples T-test. Because gylcosated hemoglobin level, `glyhb > 7.0` is a common diagnosis for diabetes, we separate the population by this condition, and test the mean Waist/Hip ratio of the two populations. Define:

$$\mu_{pos} = \text{waist/hip ratio for population testing positive for diabetes}$$
$$\mu_{neg} = \text{waist/hip ratio for population testing negative for diabetes}$$

We then define the hypotheses as follows:

$$H_0 : \mu_{pos} = \mu_{neg} \quad vs. \quad H_a : \mu_{pos} \neq \mu_{neg}$$

3.1 Normality Tests

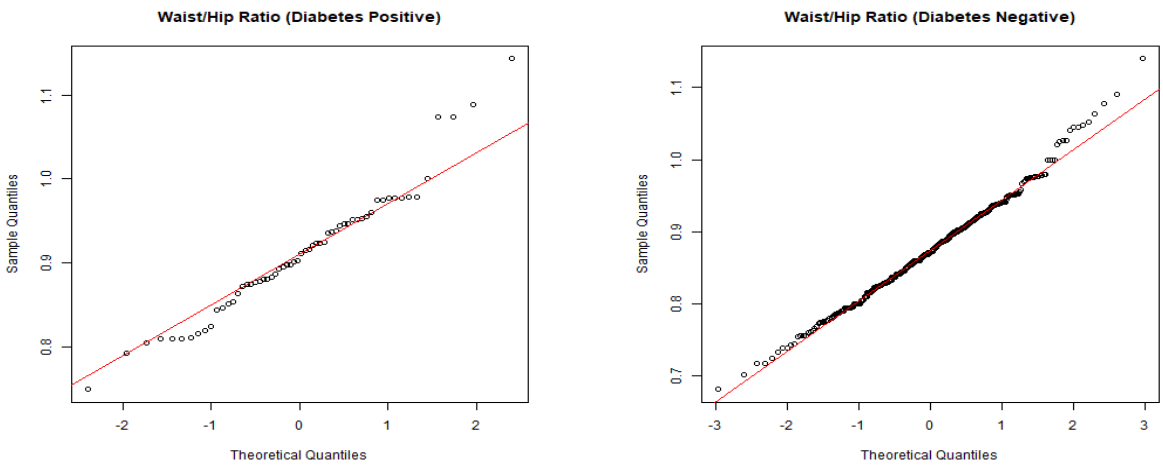
Since the T-test assumes normality in the data, we first test this in both samples, using both a normal-QQ plot for each population, and the Shapiro-Wilk Test. The normal plots are shown below in figures 1a and 1b.

The data is approximately normal by inspection, and further verification by the Shapiro-Wilk Test verifies this normality at $\alpha = 0.05$.

3.2 Testing Homogeneity of Variance

We use a two-sided F-test for the Homogeneity of variance between the samples, with the test statistic $F = s^2_{pos}/s^2_{neg}$ at $\alpha = 0.05$. Given $n_{pos} = 60$ positive samples, and $n_{neg} = 328$ negative samples (NA values dropped), the rejection regions are:

$$F > f_{59,327,0.025} \text{ or } F < f_{59,327,0.975}$$



(a) With Diabetes (b) Without Diabetes

Figure 1: Normal QQ Plots

The F-test in R has the output below.

```
F = 1.1148, num df = 59, denom df = 327, p-value = 0.5516
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.7717174 1.7007293
sample estimates:
ratio of variances
          1.114773
```

Because $p > 0.05$, and $1 \in (0.77, 1.70)$, the 95% confidence interval, we fail to reject the homogeneity of variances.[3]

3.3 Comparing Mean Waist/Hip Ratio between Diabetes Diagnoses

We have shown the data to be approximately normal across both samples, and having equal variance. Thus, we can use a pooled variance T-Test on the two independent groups to test the difference in means. Our test statistic becomes

$$t = \frac{\bar{x}_{pos} - \bar{x}_{neg} - (\mu_{pos} - \mu_{neg})}{s\sqrt{\frac{1}{n_{pos}} + \frac{1}{n_{neg}}}}$$

The T-test in R has the output below.[3]

```
t = 3.5464, df = 386, p-value = 0.0004385
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 0.01601855 0.05587835
sample estimates:
mean of x mean of y
0.9113142 0.8753658
```

From this test, since $p < 0.05$, we conclude at $\alpha = 0.05$ that there is a significant difference between the means. By the confidence interval, we verify that the waist/hip ratio is higher in patients with diabetes than those without.

This descriptive test does tell us something about the average at an aggregate population level, that the waist hip ratio is higher in diabetic patients, but does not necessarily give us much insight into the quality of the Waist/Hip Ratio as a predictor for a single individual. Although there is a difference between the means in the two populations, that does not guarantee that it is a good predictor on its own. After all, there are non-diabetic patients with higher waist/hip ratios than their diabetic counterparts. Often, multiple variables work together to produce a compound effect, and only with a set of several predictors can we make a good inference. Therefore, we conclude that there is a difference in the mean waist/hip ratio at the aggregate level, but we opt to look into feature correlations to determine additional candidates to consider together as predictors of diabetes.

4 Analysis of Feature Correlations

4.1 Multicollinearity

An important part of linear modeling is feature selection. Specifically, we want to minimize Multicollinearity, the event where our predictor variables are linearly dependent. For a linear model, consider X to be our matrix of predictor variables, and y our vector of target values. Then the least squares estimate $\hat{\beta}$ is the solution to the equation

$$(X'X)\beta = X'y$$

Multicollinearity presents three main challenges:

1. **Instability:** If we have perfect linear dependence between two predictor variables, two columns of X will be dependent, meaning X is no longer full rank nor invertible. The more likely case is if predictor variables are approximately dependent, in which $X'X$ will be nearly singular, making $\hat{\beta}$ numerically unstable to compute.

2. **Sensitivity to Small Fluctuations:** Such a model may vary largely with only small changes to the input features, due to the redundancy of attention on similar predictors.

3. **Large Variance in** β: Most coefficients of a Multicollinear model tend to have very large standard errors. [4]

4.2 Correlation Coefficients

To address this, we calculate pairwise correlation coefficients between our predictor variables. Since some our variables cannot be assumed to be normal or of the same distribution, and are subject to outliers as shown in the Exploratory Analysis, we use Kendall's Rank Correlation over the Pearson Correlation Coefficient. The assumptions for Kendall's Coefficient are:

- Variables in question are Ordinal or Continuous
- Data appears to follow a Monotonic Relationship

While it is not a linear correlation, it does still give a measurement of the dependence of our predictor variables. We calculate the pairwise correlation matrix (2)[3] for all continuous or ordinal variables using Kendall's Rank coefficient, and note the obvious relationships between `weight`, `waist`, `hip`, and `wh_ratio`. Thus we, will not use all of these in the Logistic Regression model. Additionally, we see `stab.glu` has a moderate correlation with the column corresponding to our target variable, `glyhb`. Therefore, it may be a good predictor, and should be considered in the model. We follow similar logic to choose additional features to include or exclude in the linear model, and further evaluate the presence of multicollinearity by use of the Variance Inflation Factor during Logistic Regression.

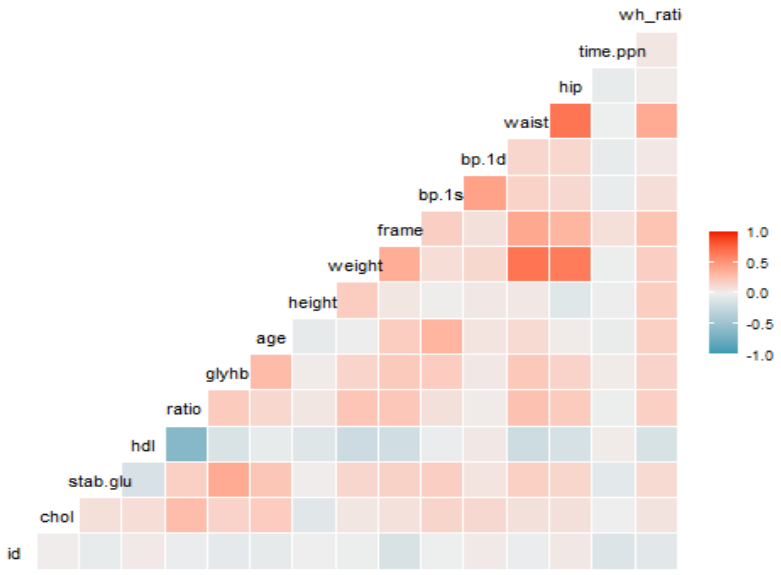


Figure 2: Kendall Rank Correlation Matrix

5 Determining Risk Factors via Logistic Regression

5.1 Model Definition

If we are trying to determine whether or not a patient has diabetes given some set of measurements, we want to determine what measurements are best to use, and how much they affect the odds that the patient has diabetes. Logistic regression is a generalized linear model with a random component that assumes a binomial distribution for the response variable, a systematic component that assumes there is a set of explanatory variables that describe the distribution of the response, and a logit link function that specifies the relationship between the response variable and the explanatory variables.[2] The assumptions for logistic regression can be summarized as follows:

- The data are independently distributed.
- The explanatory variables are independent of one another.
- The dependent variable is binomially distributed.
- There is linear relationship between the logit function of p and the explanatory variables.

Notice that errors need not be normally distributed, and there is no homogeneity of variance assumption. One drawback of logistic regression is that it uses maximum likelihood estimation, which requires a larger sample than ordinary least squares to ensure the model is not overfit. In the case of modeling whether or not a patient has diabetes with logistic regression we assume that $Y \sim \text{Binomial}(n, p)$ where p is the probability of diabetes, and Y is whether or not a patient has diabetes. There is then a set of explanatory variables $X_1, X_2, ..., X_6$ which are total cholesterol, stabilized glucose, high density lipoprotein age, weight to height ratio, and waist to hip ratio. The model then assumes that

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_6 x_6.$$

The quantity $\frac{p}{1-p}$ is referred to as the odds ratio, so logistic regression can be understood as modeling the log odds ratio as a linear function of explanatory variables.

5.2 Model Fitting and Diagnostics

Using R, we can perform a stepwise logistic regression to determine which predictors are the best. To understand how the stepwise regression works, the concept of Akaike Information Criterion must first be defined. AIC is defined as

$$\text{AIC} = 2k - 2\ln(\hat{L}),$$

where $\hat{L}$ is that maximum value of the likelihood function for the model, and k is the number of parameters in the model. The idea of AIC is that it penalizes a model that is overfit by

assigning a penalty for the number of terms, and rewards a model with good fit, as indicated by the log maximum likelihood. It is important to note that a lower AIC is an indication of a better model.

We split the data into training data and testing data, the former for fitting the coefficients of the model and the latter for testing the efficacy on out of sample data. The stepwise logistic regression calculates the AIC for the model with each predictor removed, and then removes the predictor that gives the greatest decrease in AIC. The stepwise removal terminates when the removal of any predictor gives a higher AIC than the model with the predictor included. We find that the AIC is minimized for the following model

$$\log\left(\frac{p}{1-p}\right) = -14.514 + .0485x_1 + .0098x_2 + .7249x_3 + .0538x_4$$

where x_1 is stable glucose, x_2 is cholesterol, x_3 is weight to height ratio, and x_4 is age.

We can use variance inflation factor to determine the presence of multicollinearity of predictors. Variance inflation factor is calculated by taking a predictor, and performing a regression against every other predictor in the model. The formula is

$$\text{VIF}_i = \frac{1}{1-R_i^2},$$

where R_i^2 is the r-squared of the model with predictor i as the dependent variable. VIF is bounded below by one, and unbounded above. Typical cutoff values to conclude multicollinearity are 5 and 10. We want to test for multicollinearity because if the predictors are correlated then it violates our model assumptions. It poses numerical challenges as well, inflating the standard error of the coefficients, which can lead to a false rejection of the null hypothesis that the coefficient is equal to zero. The variance inflation factors for the fitted logistic model with predictors stable glucose, cholesterol, weight to height ratio, and age are, respectively, 1.0699, 1.0859, 1.1543, and 1.1718. These are all close to one, which indicates that multicollinearity is not a significant problem for the model.

In addition to the stepwise regression and variance inflation factor, we can also perform the hypothesis test $H_0 : \beta_i = 0$ vs. $H_a : \beta_i \neq 0$. Where the test statistic is $\frac{\beta_i}{\text{SE}(\beta_i)} \sim \text{N}(0,1)$. The tests for the model are summarized in the R output below.[3]

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -14.514263   2.744451  -5.289 1.23e-07 ***
stab.glu             0.048475   0.008459   5.730 1.00e-08 ***
chol                 0.009833   0.006800   1.446  0.14813
weight_height_ratio  0.724931   0.467542   1.551  0.12102
age                  0.053822   0.019421   2.771  0.00558 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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

    Null deviance: 231.694  on 264  degrees of freedom
Residual deviance:  94.435  on 260  degrees of freedom
AIC: 104.43
```

We now need to make a decision. Do we continue to include cholesterol and weight to height ratio as a risk factor because removing them would give worse AIC for the model, or do we remove them because we cannot significantly conclude that they are not equal to zero? I chose to keep them in the model despite the fact that we cannot conclude they are zero for the simple fact that they improve the information criterion for the model.

If we take the exponential of the coefficients, we can interpret them as odds ratios. This means that the odds multiply by e^{β_i} for every 1-unit increase in x_i. We can go one step further and calculate the models probability that a patient has diabetes given stable glucose, cholesterol, weight to height ratio and age of a patient via the formula

$$p(y|x) = \frac{e^{\hat{\beta}x}}{1 + e^{\hat{\beta}x}},$$

which can be derived from the definition of the model.

We can turn the model into a classifier by selecting a cutoff probability at which we determine diabetes to be likely. From a doctor's perspective, is it worse to not classify a patient as high risk for diabetes and find out the patient does have diabetes, or to classify a patient as high risk for diabetes and find out the they don't have diabetes, that is, is it more important for the test to have a low probability of false positive or low probability of false negative? The cutoff point is not arbitrary, and if we think about the classification from a testing perspective, it directly impacts the power of the test. If the goal is to classify risk factors for diabetes, we want a test with high power, that is, if someone has diabetes, we want to correctly identity from the risk factors that they are at risk, and recommend they be have their glycated hemoglobin tested. If we want to minimize the probability of a false negative with no constraints then we would simply classify everyone as being likely for diabetes. However, doing so will drastically reduce our overall classification rate. The question is then how do we maximize the power of the test subject to the constraint that the overall classification rate meets some threshold? There is no simple answer to this problem, therefore, I chose to test the model at a probability cutoff of 0.10, because it increases the power of the test while not drastically reducing the overall classification rate.

The overall classification rate of the model was found to be 0.80 on test data, and the observed power of the classifier i.e. the proportion of cases in which the model identified diabetes, given diabetes was present, was found to be 0.875 on the test data. The probability of a type two error, or false negative, is one minus the power of the test. Therefore, the model achieved a decently low false negative rate. However, there is still clearly some room for improvement with the model. This could come from adding predictors, or training the model on a larger dataset.

To analyze how the model responds to each predictor, we can graph the models estimated probability of diabetes, as a function of the predictor. The graphs are included in figure 3.

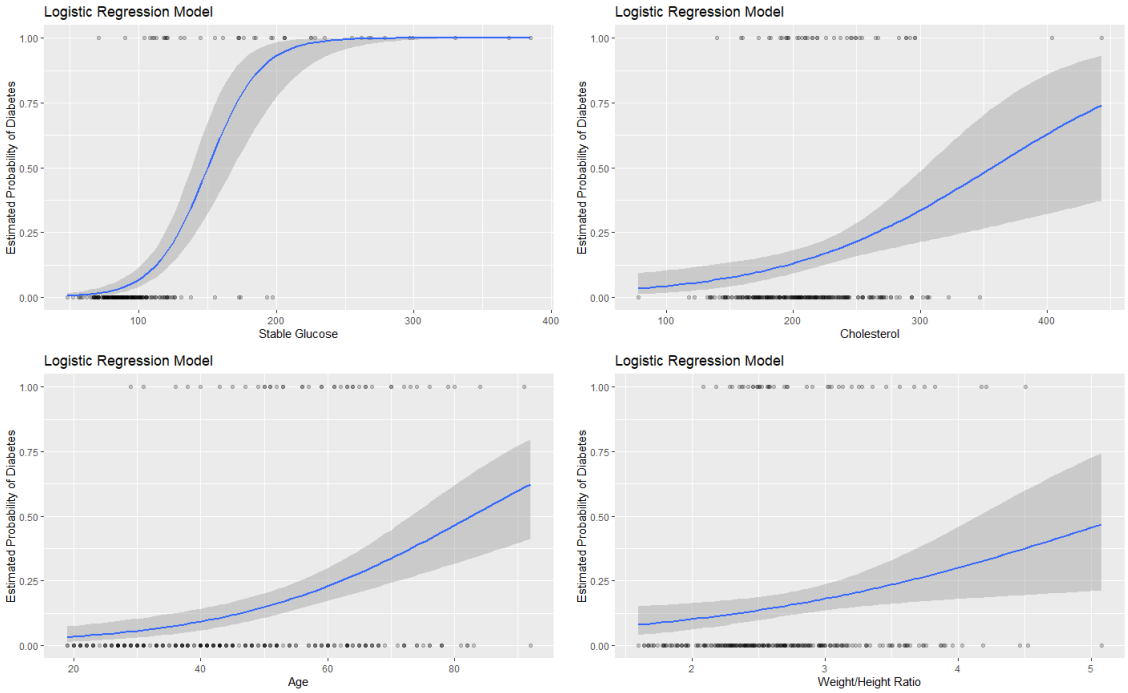


Figure 3: Estimated Probability of Diabetes as a function of the predictors.

5.3 Effects of Missing Values

5.3.1 Values Missing Completely at Random

To test the effect of missing values on the model, we randomly removed 20% of the training dataset, and then retrained the model on the remaining the data. The R output summarizes the results.[3]

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -13.967833   2.604209  -5.364 8.16e-08 ***
stab.glu             0.023688   0.005187   4.567 4.95e-06 ***
chol                 0.010757   0.006363   1.691 0.090921 .
age                  0.067351   0.020316   3.315 0.000916 ***
weight_height_ratio  1.143580   0.490627   2.331 0.019761 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

```
    Null deviance: 179.263  on 209  degrees of freedom
Residual deviance:  91.875  on 205  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 101.88
```

We can see the AIC for the model trained on the reduced data is better. This is not surprising because fewer data points will allow a closer fit, and this is also the reason we require a relatively large data set for the methods of logistic regression to be effective. We can also see that the coefficients of the model are different, and quite substantially in some cases. The p-values for the corresponding coefficient hypothesis tests have changed as well, being less than .10 in all cases. Again, we can attribute this to the fact that fewer data points allow a closer fit.

To analyze the efficacy of the model, we can test it on the out of sample data and observe the power. When testing the model trained on the reduced data set we saw an overall classification rate of 0.7727 and a power of 0.8125. Both the power and the overall classification rate saw a reduction, meaning the model is likely overfit to the reduced dataset, and is worse predicting out of sample.

5.3.2 Non-Ignorable Missing Values

If we are dealing with data that are not missing completely at random, then we need to consider why these data may be missing. There is the possibility that there are missing factors that explain the response variable, i.e. there are unmeasured or immeasurable characteristics that we have not incorporated into our model. This is likely given our model using only four predictors. In the case of non-ignorable missing values, those that are null as a result of some underlying latent feature in the data, a further investigation must be done to understand the reason for those missing values. In our data, for example, the 403 participants in the trial were only a subset of 1031 who were contacted. Out of those 1031, 403 received the screening, 197 refused, 316 could not be contacted, and 115 were contacted but could not schedule a screening time. [1] The non-respondents can not be ignored because there may be an ulterior reason for not participating. We will not make assumptions about these reasons, but do acknowledge that there are possible missing values that could impact our analysis.

References

[1] DE Hunt JP Willems, JT Saunders and JB Schorling. Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. 1997.

[2] Nelder J.A. McCullagh, P. *Generalized Linear Models*. CRC Press, 2 edition, 1989.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[4] Aijt Tamhane and Dorothy Dunlop. *Statistics and Data Analysis: From Elementary to Intermediate*. Pearson, Belmont, CA, USA, 1 edition, 2000.