Heart Disease: A Logistic Regression Analysis of Risk Factors

Colin Busby
BSTA 661
Fall 2022

I. Introduction

The primary research objective of this data analysis was to explore a heart disease data set, and utilize logistic regression to analyze the explanatory variables present in the data set to determine which are the best at predicting the presence of heart disease. This particular topic used for this analysis was chosen due to the huge impact of heart disease on the American public, and by extension, the rest of the world.

According to the Centers for Disease Control and Prevention (CDC), heart disease, also called cardiovascular disease (CVD), is the leading cause of death for the majority of ethnic and racial groups in America: in 2020 alone approximately 697,000 people died from heart disease in the United States. CVD isn't just one disease, it covers several related types of heart conditions. The most common is Coronary Artery Disease which affects 20.1 million Americans, and causes decreased blood flow to the heart, eventually leading to a heart attack. Even if you ignore the human toll of cardiovascular disease, CVD has a financial cost: from 2017 to 2018, heart disease cost the United States over $229 billion. This financial cost is due to a variety of reasons, such as the cost of healthcare services such as hospital stays and ER visits, medicinal costs, and the cost of lost productivity from absence and death. From these numbers alone one can see the impact cardiovascular disease has on the world, thus why it's vital that the risk of heart disease be accurately assessed.

II. Data Description

The data set used in this exploration and analysis is Personal Key Indicators of Heart Disease, from the Kaggle user Kamil Pytlak. The data itself consists of information gathered from a telephone survey as part of the Behavioral Risk Factor Surveillance System (BRFSS), a CDC system of "health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services" (CDC - BRFSS). Pytlak cleaned and compiled the raw data, reducing the initial 300 variables down to 18 related to heart disease.

The dimensions of the data set are 319795 rows (observations) by 18 columns (variables). The variables consist of the nominal response variable *HeartDisease*, and 17 explanatory variables: *BMI* (continuous), *Smoking* (nominal, Yes/No), *AlcoholDrinking* (nominal, Yes/No), *Stroke* (nominal, Yes/No), *PhysicalHealth* (interval), *MentalHealth* (interval), *DiffWalking* (nominal, Yes/No), *Sex* (nominal, Male/Female), *AgeCategory* (ordinal, "18-24" years, etc.), *Race* (nominal, American Indian/Asian/Black/Hispanic/White), *Diabetic* (nominal, 4

levels), *PhysicalActivity* (nominal, Yes/No), *GenHealth* (ordinal, 5 levels), *SleepTime* (continuous), *Asthma* (nominal, Yes/No), *KidneyDisease* (nominal, Yes/No), and *SkinCancer* (nominal, Yes/No). It should be noted that for *Smoking,* 'Yes' is defined as having smoked over 100 cigarettes during a person's life while 'No' is defined as smoking none or less than 100 cigarettes. Finally, *AlcoholDrinking* is defined as heavy drinking: 14 or more drinks a week for adult men and 7 or more drinks a week for adult women.

There were several observations of note during the data exploration. The gender ratio was 52.47% women and 47.53% male, while the largest racial/ethnic category was White at 76.68%, with Hispanic and Black being the next two largest groups at 8.58% and 7.17%, respectively. In terms of being diabetic, 87% of respondents did not have or were borderline diabetic, with no significance between the two. Approximately 67.82% of respondents reported having a BMI of at least 25, indicating they were overweight or obese. For the variable *Stroke*, 96.23% of respondents had never had a stroke, while the remaining 3.77% had previously had a stroke (Figure 1). Finally, 58.75% of respondents reported that they had smoked over 100 cigarettes during their lifetime.

Stroke	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	307726	96.23	307726	96.23
Yes	12069	3.77	319795	100.00

Figure 1. Proportion of respondents who have/haven't had a stroke

III. Methods and Results

The initial step of the data analysis was to recode several of the variables, either due to format or non-significance. The variable *AgeCategory* was re-coded into the continuous variable *age* by taking the median of each category and dividing by 5. The variable "Diabetic" had the level "Yes (during pregnancy)" re-coded into "No" due to it being non-significant. Finally, the variable *Race* had the levels "Ameri" [American Indian/Alaskan Native] and "Other" re-coded as "White" due to non-significance.

Before the analysis was performed, a sample size of 5000 was taken from the data set using sampling without replacement, as any sample larger than that would lock up and crash the SAS program on the laptop used due to the amount of memory utilized for the analysis.

The initial procedure used to analyze the data was the PROC GENMOD procedure, in order to fit a generalized linear model to all of the variables, using *HeartDisease* as the response variable and the other 17 variables as the explanatory variables. The binomial distribution was used along with a Logit link function, while the reference levels of the class variables were set to "No", for ease of interpreting the results. The resulting Analysis of Maximum Likelihood Parameter Estimates found only one variable, *MentalHealth*, to have a p-value greater than 0.05 at 0.6182, leaving the other 16 explanatory variables in place. The AIC provided by the analysis was extremely high, with a value of 147820.3044 (Figure 2).

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Log Likelihood		-73887.1522	
Full Log Likelihood		-73887.1522	
AIC (smaller is better)		147820.3044	
AICC (smaller is better)		147820.3079	
BIC (smaller is better)		148065.8394	

Figure 2. Goodness-of-Fit for generalized linear model

To corroborate this model, a second procedure, PROC LOGISTIC, was employed to generate another model using logistic regression alongside the Backward Elimination Procedure. This analysis resulted in the elimination of nine variables: *PhysicalHealth*, *BMI*, *Race*, *PhysicalActivity*, *SkinCancer*, *SleepTime*, *Asthma*, *MentalHealth*, and *AlcoholDrinking*. The resulting model was thus made up of *Smoking, Stroke, DiffWalking, Sex, age, Diabetic, GenHealth,* and *KidneyDisease*. While the removal of *BMI* and *AlcoholDrinking* from the model was a surprise, given that excessive drinking and obesity are good indicators for the risk of heart disease, this could be explained as *BMI* and *AlcoholDrinking* showing correlation with variables in the final model, such as *Smoking* or *Diabetic*. The resulting AIC score was 3073.408, a significant improvement over the PROC GENMOD AIC score. Additionally, a ROC Curve for the generated model was created (Figure 3), showing an AUC of 0.8469, well within the >0.80 criterion for a good score, suggesting that the generated model has an 84.69% of correctly guessing the presence of heart disease based upon the provided parameters. Further confirmation of the model's fit was provided by the Hosmer and Lemeshow Goodness-of-Fit Test, with a p-value of 0.2659, suggesting that

the model is a good fit for the sample used. The final model is

HeartDisease =
-4.4022+0.4717Smoking+1.5323Stroke+0.2926DiffWalking-0.8296Sex+0.237age+0.55Diabetic+GenHealth(-1.9847=Excellent,-1.478=VeryGood,-0.8787=Good)+0.4324KidneyDisease

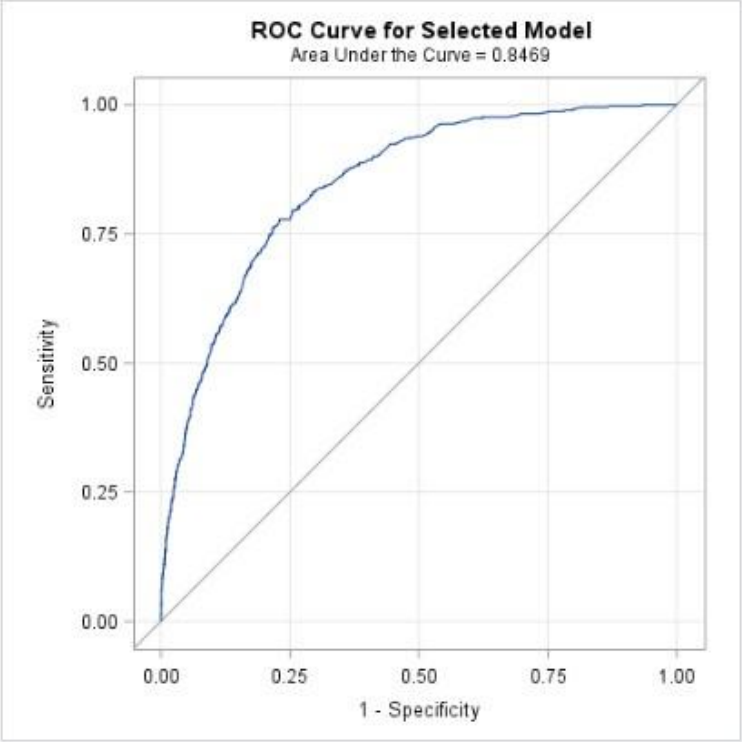

Figure 3. ROC Curve and AUC score for logistic regression model

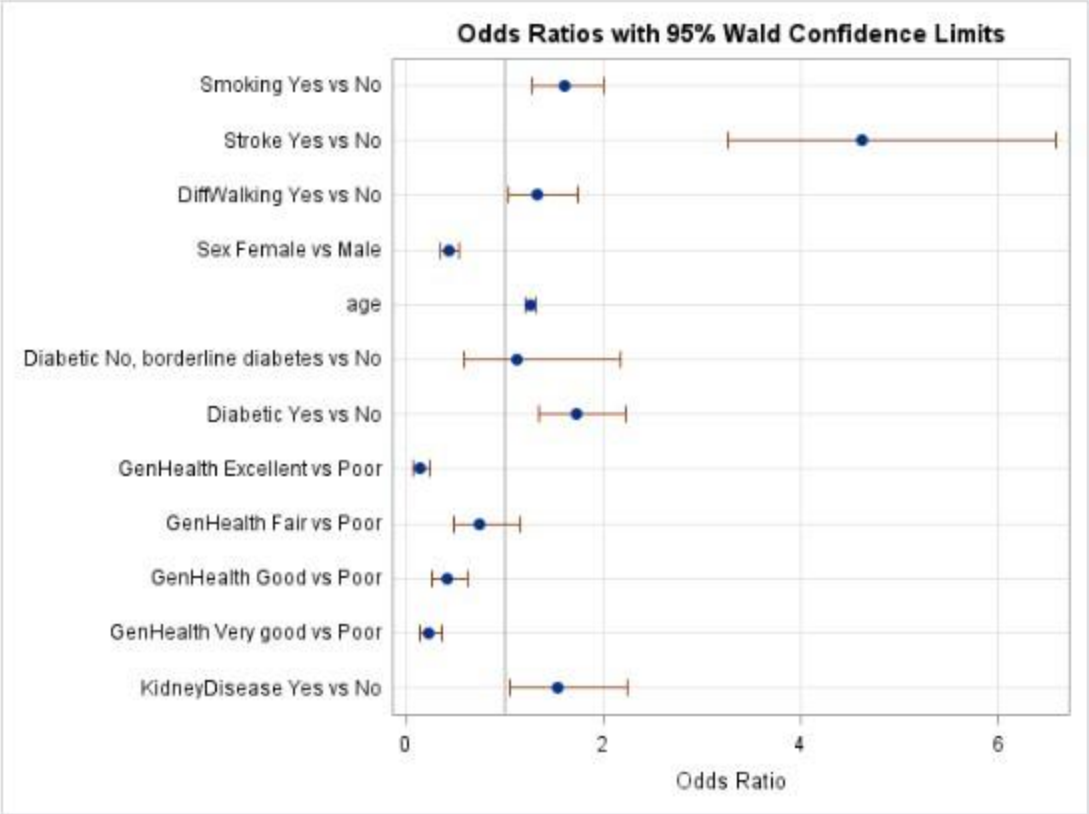

Figure 4. Odds ratios of explanatory variables in comparison to baseline references

For analysis of the results, an Odd Ratios plot was generated as part of the logistic regression. The resulting graph (Figure 4) shows that the greatest factor in increasing the odds of having heart disease is whether or not someone has ever had a stroke, with the odds of having heart disease being 4.6 times greater than if you never had a stroke. Additionally, having had a stroke does drastically increase your odds of heart disease regardless of your degree of general health (Figure 5). Smoking and diabetes are the next two largest, with a person who smokes having a 60.3% increase in the odds of having cardiovascular disease, while the odds of having heart disease are 1.733 times greater than if you don't have diabetes. Finally, the odds of having heart disease if you are a woman are 0.436 times that of a man.

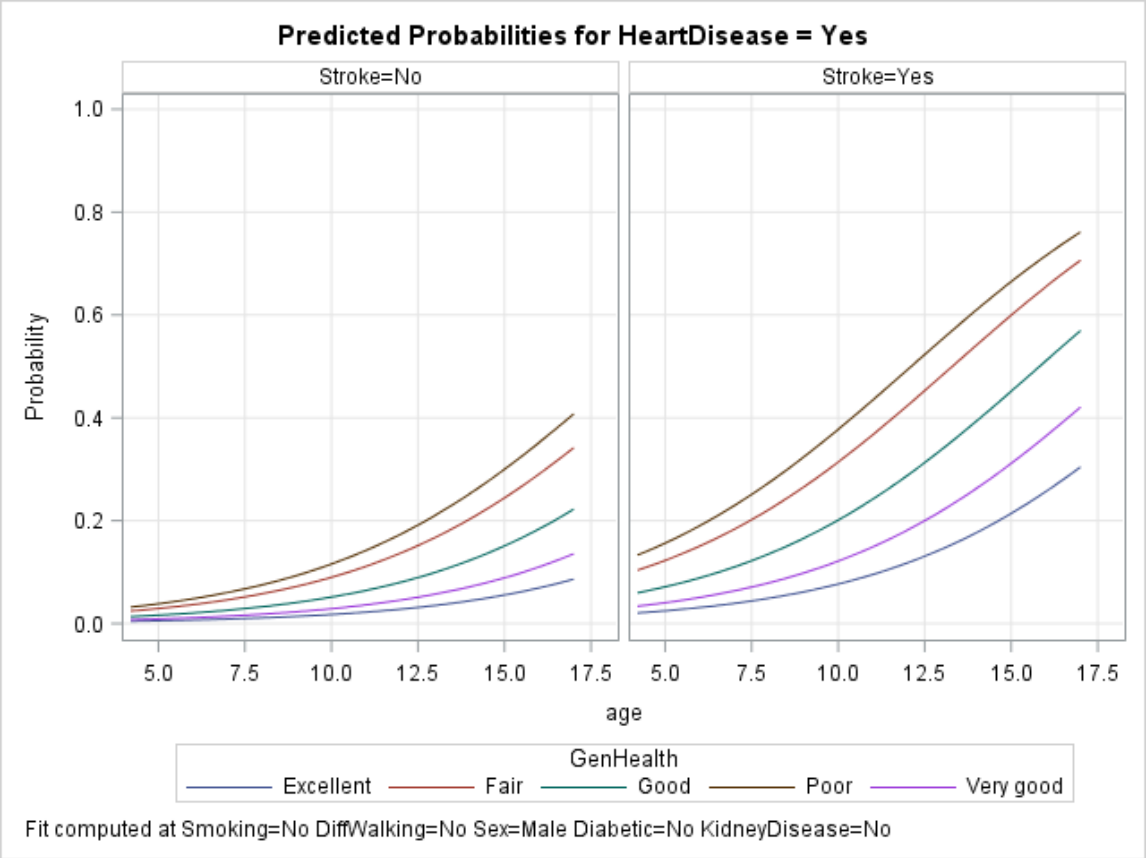

Figure 5. Predicted probabilities for heart disease based on general health and stroke history.

IV. Conclusion

Overall, the logistic regression model was found to be a much greater fit than the generalized linear model for interpreting the relationship between the response variable *HeartDisease* and the explanatory variables. One caveat to the model is that it was only tested with a sample size of 5000. Larger samples may prove the model to be a poor fit.

In terms of heart disease, based upon the logistic regression model, the average person at the greatest risk of heart disease is an older man in poor health, potentially overweight, suffering from diabetes, has a history of strokes, smokes, and has difficulty walking. To put it another way, living a healthy lifestyle significantly cuts down a person's risk of heart disease. While we cannot control our age or biological gender, we can control factors such as exercise, weight, diet, and smoking that contribute to obesity and strokes.

SAS code

Github Repository: BusbyCI11/BSTA-661-Project

Note: Change datafile to location of heart_2020_cleaned.csv

```
PROC IMPORT OUT= WORK.HEART_DISEASE
            DATAFILE= "G:\SAS Projects\BSTA 661 Project\heart_2
020_cleaned.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

proc contents data=WORK.HEART_DISEASE order=varnum;
run;

data hd_num; set WORK.HEART_DISEASE;
      if AgeCategory='18-24' then agecatnum=21;
      if AgeCategory='25-29' then agecatnum=27;
      if AgeCategory='30-34' then agecatnum=32;
      if AgeCategory='35-39' then agecatnum=37;
      if AgeCategory='40-44' then agecatnum=42;
      if AgeCategory='45-49' then agecatnum=47;
      if AgeCategory='50-54' then agecatnum=52;
      if AgeCategory='55-59' then agecatnum=57;
      if AgeCategory='60-64' then agecatnum=62;
      if AgeCategory='65-69' then agecatnum=67;
      if AgeCategory='70-74' then agecatnum=72;
      if AgeCategory='75-79' then agecatnum=77;
      if AgeCategory='80 or older' then agecatnum=85;
      if Diabetic='Yes (during pregnancy)' then Diabetic='No';
      if Race='Ameri' then Race='White';
      if Race='Other' then Race='White';

run;

data hd_num; set hd_num;
age = agecatnum/5;
age2 = age*age;
run;
```

```
proc freq data=heart_disease;
table sex*AgeCategory/nocol norow;
run;

proc freq data=heart_disease;
     table Race*AgeCategory/nocol norow;
run;

proc freq data=hd_num;
     table Diabetic;
run;
proc freq data=hd_num;
     table Smoking;
run;

proc freq data=hd_num;
     table Stroke;
run;

proc univariate data=hd_num;
     var BMI;
     histogram;
run;

proc freq data=hd_num nlevels;
        table BMI Smoking AlcoholDrinking Stroke PhysicalHealth
MentalHealth PhysicalActivity DiffWalking Sex agecatnum Race Diabetic
GenHealth SleepTime Asthma KidneyDisease SkinCancer Sex*Stroke /
noprint;
run;

proc surveyselect data = hd_num
     out = hd_num_sample
     method = SRS rep = 1
     sampsize = 5000
     seed = 12345;
run;

proc genmod descending data=hd_num;
```

```
class Smoking(ref="No") AlcoholDrinking(ref="No") Stroke(ref="No")
DiffWalking(ref="No") Sex agecatnum(ref='21') Race Diabetic(ref="No")
GenHealth(ref="Poor") Asthma(ref="No") KidneyDisease(ref="No")
SkinCancer(ref="No") PhysicalActivity(ref="No");
model HeartDisease = BMI Smoking AlcoholDrinking Stroke
PhysicalHealth MentalHealth DiffWalking age PhysicalActivity Race
Diabetic
GenHealth SleepTime Asthma KidneyDisease SkinCancer / dist=bin
link=logit ;
output out=temp p=pred upper=ucl lower=lcl;
run;

ods graphics on;

proc logistic descending data=hd_num_sample plots=oddsratio;
class Smoking(ref="No") AlcoholDrinking(ref="No") Stroke(ref="No")
DiffWalking(ref="No") Sex agecatnum(ref='21') Race Diabetic(ref="No")
GenHealth(ref="Poor") Asthma(ref="No") KidneyDisease(ref="No")
SkinCancer(ref="No") PhysicalActivity(ref="No") / param=ref;
 model HeartDisease = BMI Smoking AlcoholDrinking Stroke
PhysicalHealth MentalHealth DiffWalking Sex age PhysicalActivity Race
Diabetic
GenHealth SleepTime Asthma KidneyDisease SkinCancer /
selection=backward lackfit aggregate=(BMI Smoking AlcoholDrinking
Stroke PhysicalHealth MentalHealth DiffWalking Sex age
PhysicalActivity Race Diabetic
GenHealth SleepTime Asthma KidneyDisease SkinCancer) outroc=classif1;
output out = prob PREDPROBS=I;
store logiModel;
run;

title "Predicted Probabilities of Heart Disease";
proc plm source=logiModel;
     effectplot slicefit(x=age sliceby=GenHealth plotby=Smoking);
     effectplot slicefit(x=age sliceby=Sex plotby=Smoking);
     effectplot slicefit(x=age sliceby=Sex plotby=Stroke);
     effectplot slicefit(x=age sliceby=Stroke plotby=Sex);
     effectplot slicefit(x=age sliceby=Stroke plotby=Smoking);
     effectplot slicefit(x=age sliceby=GenHealth plotby=Stroke);
```

```
run;

ods graphics off;
```