STAT 495 Project

Personal Key Indicators of Heart Disease

(Binary Logistic Regression)

Report Prepared By
Ethan Huang

May 3, 2022

Table of Contents

I. Introduction

For this project, I wanted to find a health-related dataset for which I could run a regression analysis. Regression analysis is a key component of statistics as it explains why something occurs or fails to occur. Within the first 10 minutes of looking through Kaggle, I stumbled upon a usable dataset suitable for building regression and machine learning models. There is a binary response variable (HeartDisease) along with 19 predictor variables related to the patient's health profile. Thus, I could apply a Binary Logistic Model, a topic covered in class. I was very excited to understand the factors influencing heart disease and to predict one's odds of heart disease based on his or her health profile.

II. Background

Each year, the CDC conducts telephone surveys asking adults about their health status through the Behavioral Risk Factor Surveillance System. It contains questions about demographics, underlying conditions, and habits. Around 300,000-400,000 adults are interviewed each year [1].

According to the 2020 survey, 8% of all survey participants have heart disease. Many of them exhibit risk factors including drug use, obesity, and low physical activity. Heart disease is a leading cause of death for many Americans, and 47% of all Americans exhibit at least one of three key risk factors: high blood pressure, high cholesterol, and tobacco use [2]. Other risk factors include lack of sleep, excessive alcohol use, diabetes, and many other factors.

III. The Dataset

The data, found through Kaggle, is a compilation of the 319,796 surveys administered in 2020. It features the 20 top variables influencing a person's presence or absence of heart disease. The explanatory variables include BMI, Smoking, Alcohol, Stroke, and many more variables [1].

However, I made some modifications to the data, and I condensed the number of predictor variables from 19 to 11. First, I changed the categorical predictor AgeCategory to the numerical predictor Age as there were too many age categories. To do so, I replaced each category with the median of the original

category. For example, if the original category was 30-34, then the new value is 32. Second, I narrowed the values for Diabetes to 'Yes' and 'No'. Prior to the change, there were values for gestational and borderline diabetes, which I changed to 'Yes' and 'No' respectively.

IV. Fitted Model

Both the SAS and R outputs show that all the predictors are significant at the 1% level. The fitted model is as follows:

$$P(\text{HeartDisease}) = \frac{\exp(-8.0927301+0.0207536*\text{BMI}+0.4919958*\text{Smoker}+0.2840615*(\text{AlcoholUse=No})+1.3447983*\text{Stroke}+0.6639444*\text{Male}+0.0608215*\text{Age}+0.3163944*\text{Black}+0.4298358*\text{Hispanic}+0.6885143*\text{NativeAmerican}+0.5754208*\text{Other}+0.4874692*\text{White}+0.7406995*\text{Diabetic}+0.3319632*\text{NoPhysicalActivity}-0.0531171*\text{SleepTime}+0.4879486*\text{Asthma})}{[1+\exp(-8.0927301+0.0207536*\text{BMI}+0.4919958*\text{Smoker}+0.2840615*(\text{AlcoholUse=No})+1.3447983*\text{Stroke}+0.6639444*\text{Male}+0.0608215*\text{Age}+0.3163944*\text{Black}+0.4298358*\text{Hispanic}+0.6885143*\text{NativeAmerican}+0.5754208*\text{Other}+0.4874692*\text{White}+0.7406995*\text{Diabetic}+0.3319632*\text{NoPhysicalActivity}-0.0531171*\text{SleepTime}+0.4879486*\text{Asthma})]}$$

V. Interpretation of Selected Coefficients

Smoking: A smoker is 163.558% more likely to develop heart disease than a non-smoker.

$\exp(0.4919958)*100=163.558$

PhysicalActivity=No: Someone who lacks physical activity is 139.37% more likely to develop heart disease than someone who has physical activity.

$\exp(0.3319632)*100=139.370$

SleepTime: For each additional hour of sleep, the estimated odds of heart disease decrease by 5.172%.
(exp(-0.0531171)-1)*100=-5.172

BMI: For each increase in BMI, the estimated odds of heart disease increase by 2.098%.
(exp(0.0207536)-1)*100=2.098

VI. Predicted Probability

Someone with a high chance of heart disease is an older Native American male with a higher BMI. His underlying conditions include diabetes, asthma, and stroke. He smokes, drinks, sleeps less, and lacks physical activity. We set age as 80, SleepTime at 3 hours, and BMI as 35. Based on the SAS and R codes, he has approximately an 89% chance of developing heart disease.

VII. Conclusion

In short, all of the predictors are significant. The model highlights the habits and underlying conditions contributing to heart disease. People with underlying conditions are more likely to get heart disease than people who do not have underlying conditions. To reduce one's risk of heart disease, one can refrain from drug use, get enough exercise, and get a good night's sleep. Overall, this is an informative dataset as I got to apply concepts from STAT 495.

Appendix

SAS Code

```
/*To Import CSV File*/
proc import out=heartdisease datafile="C:\Users\colle\Downloads\archive
(6)\heart_2020_modified.csv"
dbms=csv replace;

proc genmod data=heartdisease;
/*To set categorical and reference variables*/
class Smoking (ref="No") AlcoholDrinking (ref="Ye") Sex (ref= "Female") Race
(ref= "Asian") Diabetic (ref= "No")
Stroke (ref="No") PhysicalActivity  (ref= "Yes") Asthma (ref = "No");
/*To fit the model*/
model HeartDisease(event="Yes")= Smoking AlcoholDrinking Sex AgeCategory
Race Diabetic Stroke PhysicalActivity Asthma BMI SleepTime/
        dist=binomial link=logit;
run;

/*To find the predicted value of one with the highest odds of Heart Disease*/
data prediction;
input Smoking $ AlcoholDrinking $ Sex $ AgeCategory Race $ Diabetic $ Stroke
$ PhysicalActivity $ Asthma $ BMI SleepTime;
cards;
Yes Yes Male 80 NativeAmerican Yes Yes No Yes 35 3
;

data heartdisease;
set heartdisease prediction;
run;

proc genmod data=heartdisease;
class Smoking AlcoholDrinking Sex Race Diabetic Stroke PhysicalActivity
Asthma;
```

```
model HeartDisease(event="Yes")= Smoking AlcoholDrinking Sex AgeCategory
Race Diabetic Stroke PhysicalActivity Asthma BMI SleepTime/
        dist=binomial link=logit;
output out=outdata p=pred_probdisease;
run;

/*To print our predicted value*/
proc print data=outdata(firstobs=319796)noobs;
var pred_probdisease;
run;
```

SAS Output
Binary Logistic Model

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-8.0927	0.0918	-8.2727	-7.9128	7771.71	<.0001
Smoking	Yes	1	0.4920	0.0139	0.4648	0.5192	1254.08	<.0001
Smoking	No	0	0.0000	0.0000	0.0000	0.0000	.	.
AlcoholDrinking	No	1	0.2841	0.0329	0.2195	0.3486	74.46	<.0001
AlcoholDrinking	Ye	0	0.0000	0.0000	0.0000	0.0000	.	.
Sex	Male	1	0.6639	0.0140	0.6365	0.6914	2242.59	<.0001
Sex	Female	0	0.0000	0.0000	0.0000	0.0000	.	.
Age		1	0.0608	0.0006	0.0597	0.0620	10569.5	<.0001
Race	Black	1	0.3164	0.0714	0.1765	0.4563	19.65	<.0001
Race	Hispa	1	0.4298	0.0720	0.2886	0.5710	35.59	<.0001
Race	Nativ	1	0.6885	0.0827	0.5263	0.8507	69.24	<.0001
Race	Other	1	0.5754	0.0762	0.4260	0.7248	57.00	<.0001
Race	White	1	0.4875	0.0665	0.3572	0.6177	53.80	<.0001
Race	Asian	0	0.0000	0.0000	0.0000	0.0000	.	.
Diabetic	Yes	1	0.7407	0.0158	0.7098	0.7716	2205.51	<.0001
Diabetic	No	0	0.0000	0.0000	0.0000	0.0000	.	.
Stroke	Yes	1	1.3448	0.0219	1.3019	1.3877	3781.18	<.0001
Stroke	No	0	0.0000	0.0000	0.0000	0.0000	.	.
PhysicalActivity	No	1	0.3320	0.0149	0.3028	0.3612	496.87	<.0001
PhysicalActivity	Yes	0	0.0000	0.0000	0.0000	0.0000	.	.
Asthma	Yes	1	0.4879	0.0186	0.4515	0.5244	688.82	<.0001
Asthma	No	0	0.0000	0.0000	0.0000	0.0000	.	.
BMI		1	0.0208	0.0011	0.0186	0.0229	350.38	<.0001
SleepTime		1	-0.0531	0.0045	-0.0619	-0.0444	141.18	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

Predicted Value

pred_probdisease
0.88992

R Code

```
#We Use read.csv to import the dataset
heartdisease <- read.csv("C:\\Users\\colle\\Downloads\\archive
(6)\\heart_2020_modified.csv")

#We set our reference variables
HeartDisease.rel <- relevel(as.factor(heartdisease$HeartDisease),ref="No")
AlcoholDrinking.rel <- relevel(as.factor(heartdisease$AlcoholDrinking),ref="Yes")
Race.rel <- relevel(as.factor(heartdisease$Race),ref="Asian")
PhysicalActivity.rel <- relevel(as.factor(heartdisease$PhysicalActivity),ref="Yes")

#We fit our model with family=binomial, link=logit
summary(fitted.model<- glm(HeartDisease.rel ~ BMI + Smoking +
AlcoholDrinking.rel + Stroke +
 Sex + Age + Race.rel + Diabetic + PhysicalActivity.rel +
 SleepTime + Asthma ,data=heartdisease, family=binomial(link=logit)))

#We print the predicted value
print(predict(fitted.model,type="response",data.frame(Race.rel="NativeAmerican",
BMI=35,Smoking="Yes",
AlcoholDrinking.rel="No",Stroke="Yes",Sex="Male",Age=80,Diabetic="Yes",
PhysicalActivity.rel="No",SleepTime=3,Asthma="Yes")))
```

R Output
Binary Logistic Model

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.0927301  0.0917920 -88.164  < 2e-16 ***
BMI                      0.0207536  0.0011087  18.719  < 2e-16 ***
SmokingYes               0.4919958  0.0138928  35.414  < 2e-16 ***
AlcoholDrinking.relNo    0.2840615  0.0329195   8.629  < 2e-16 ***
StrokeYes                1.3447983  0.0218695  61.492  < 2e-16 ***
SexMale                  0.6639444  0.0140200  47.357  < 2e-16 ***
Age                      0.0608215  0.0005916 102.814  < 2e-16 ***
Race.relBlack            0.3163944  0.0713646   4.433 9.27e-06 ***
Race.relHispanic         0.4298358  0.0720419   5.966 2.42e-09 ***
Race.relNativeAmerican   0.6885143  0.0827403   8.321  < 2e-16 ***
Race.relOther            0.5754208  0.0762118   7.550 4.34e-14 ***
Race.relWhite            0.4874692  0.0664553   7.335 2.21e-13 ***
DiabeticYes              0.7406995  0.0157719  46.963  < 2e-16 ***
PhysicalActivity.relNo   0.3319632  0.0148922  22.291  < 2e-16 ***
SleepTime               -0.0531171  0.0044703 -11.882  < 2e-16 ***
AsthmaYes                0.4879486  0.0185914  26.246  < 2e-16 ***
---
```

Predicted Value

```
        1
0.8899156
```

References

[1] “Personal Key Indicators of Heart Disease Dataset”, Kaggle.com, https://www.kaggle.com/datasets/kamil pytlak/personal-key-indicators-of-heart-disease, 2022

[2] “Know Your Risk For Heart Disease”, CDC.gov, https://www.cdc.gov/heartdisease/risk_factors.htm, 2019