Personal Key Indicators of Heart Disease

Charles Hwang

Table of Contents

Introduction

- I wanted to predict whether someone has heart disease using different variables
- Dataset is reliable (federal data and available on Kaggle)
  - Kaggle is a website that houses free-to-use datasets and hosts competitions for predictive analytics
  - 237,624 views and 35,303 downloads
- Significance: Heart disease is the leading cause of death in the United States[1]

Dataset

- Available dataset includes 319,795 rows and 18 columns[2]
- Centers for Disease Control and Prevention (CDC) conducts an annual telephone survey of US residents to gather various health data
- Survey in 2020 included “all 50 states as well as the District of Columbia and three U.S. territories”
    - CDC claims it is “the largest continuously conducted health survey system in the world”
- Full survey is 401,958 rows and 279 columns

Variables[2]

- HeartDisease (response; binary)
  - Coronary heart disease (CHD) or myocardial infarction (MI)
- BMI* (kg/m^2)
- Smoking (5+ packs lifetime; binary)
- AlcoholDrinking (“heavy”; binary)
- Stroke (binary)
- PhysicalHealth* (poor days in last 30)
- MentalHealth* (poor days in last 30)
- DiffWalking (binary)
- Sex (binary)
- AgeCategory (18-24, 25-29, … , 80+)
- Race (White, Hispanic, Black, Asian, American Indian/Alaskan Native, other)
- Diabetic (No, Yes, Borderline, Preg)
- PhysicalActivity (past 30 days; binary)
- GenHealth (Likert: EX, VG, G, F, P)
- SleepTime* (average hours per day)
- Asthma (binary)
- KidneyDisease (binary)
- SkinCancer (binary)

*numeric variable

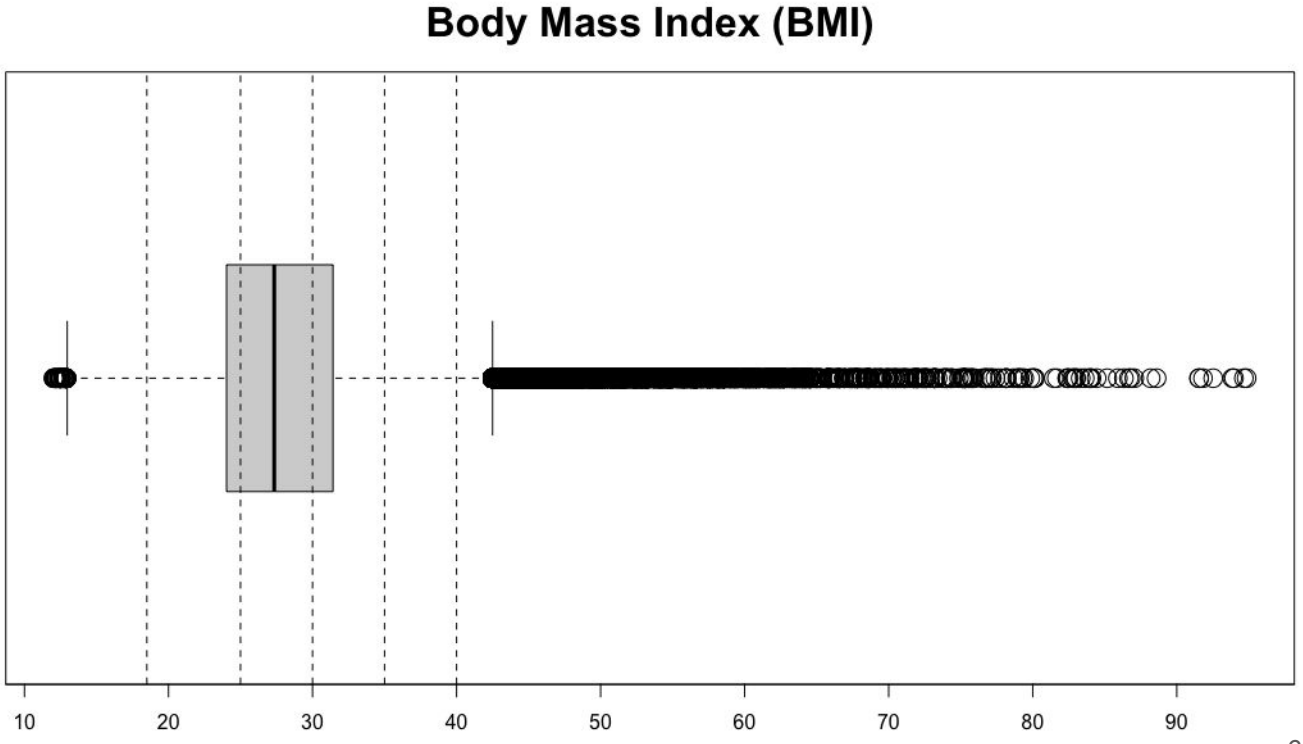
Data Visualization

Classification Table

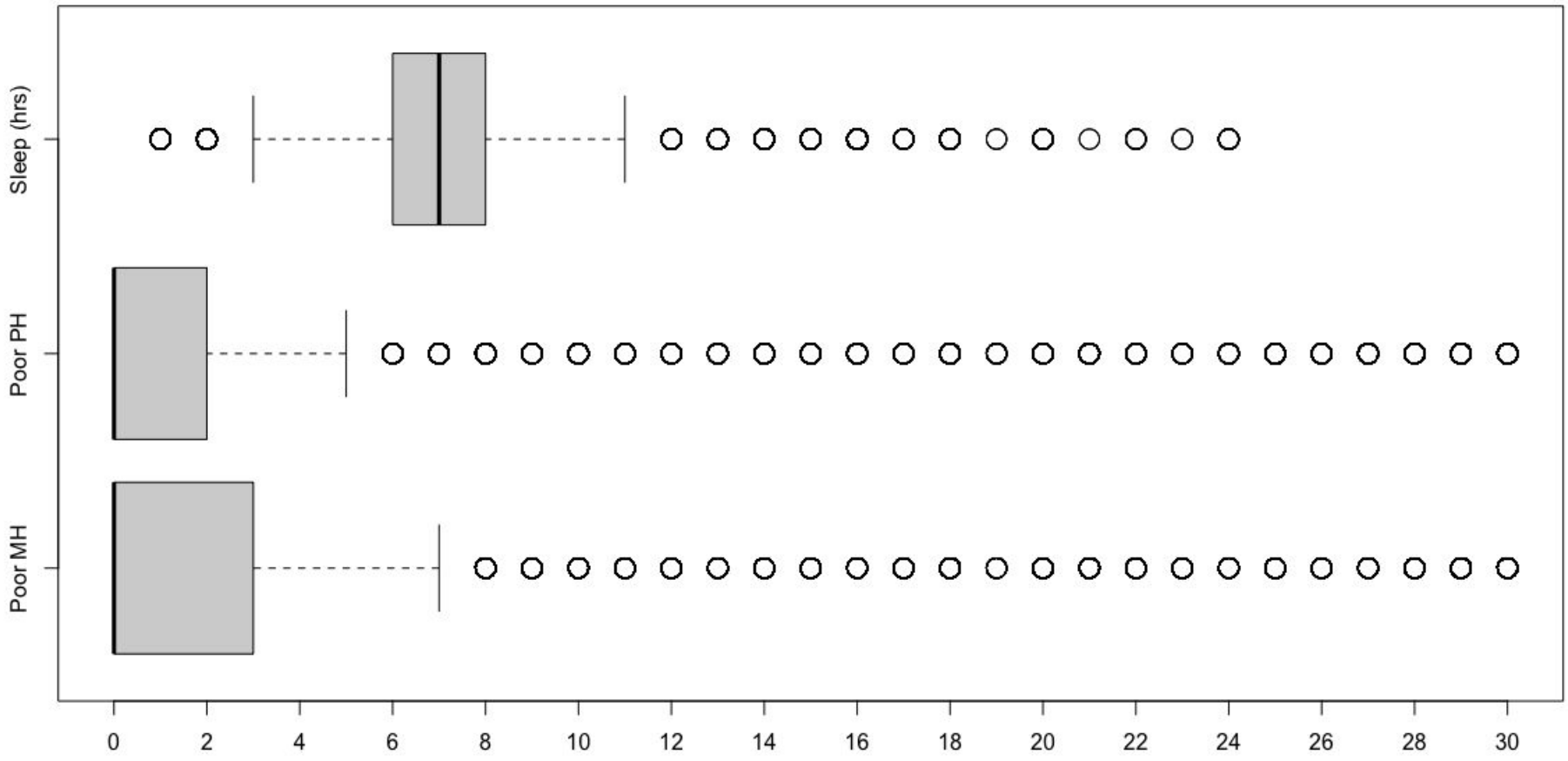
No	Yes
292422	27373

$p = 0.0855955$

$1 - p = 0.9144045$



of Poor Mental/Physical Health Days (Last 30) and Hours of Sleep



Correlation Table of Numeric Variables

	BMI	PhysicalHealth	MentalHealth	SleepTime
BMI	1.00000000	0.10978754	0.06413057	-0.05182225
PhysicalHealth	0.10978754	1.00000000	0.28798667	-0.06138663
MentalHealth	0.06413057	0.28798667	1.00000000	-0.11971679
SleepTime	-0.05182225	-0.06138663	-0.11971679	1.00000000

Methods

- Since response variable is binary, logistic regression is the clear choice
- Data was split into training (80%) and test (20%) sets
- Generalized linear model (GLM) with all variables and `family=binomial`
- Backward selection with the `step` function produced the same full model

Assumptions

- Binary response (slide 5) ✓
- Independence (slide 4) ✓
- Linearity between logit of response and each predictor variable ✓*
  - * We have no reason to believe otherwise
- Absence of multicollinearity (slide 8) ✓
- Absence of outliers and influential points (slides 6-7) ✗
- Large sample size (slide 4) ✓

Results

- Accuracy rate: 0.9140856
- Sensitivity (TP): 0.1011115
- Specificity (TN): 0.991761018
- Positive predictive value: 0.9246553
    - False discovery rate: 0.07534472

Confusion Matrix

	0	1
No	57900	481
Yes	5014	564

Matrix of Results

	Positive	Negative
True	0.1011115	0.991761018
False	0.008238982	0.8988885

```
Call:
glm(formula = HeartDisease ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1133  -0.4104  -0.2428  -0.1270   3.6221

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -6.3097610  0.1299880 -48.541  < 2e-16 ***
BMI                 0.0090913  0.0012816   7.094 1.30e-12 ***
SmokingYes          0.3602878  0.0161066  22.369  < 2e-16 ***
AlcoholDrinkingYes -0.2507708  0.0378075  -6.633 3.29e-11 ***
StrokeYes           1.0453483  0.0252846  41.343  < 2e-16 ***
PhysicalHealth      0.0034601  0.0009667   3.579 0.000345 ***
MentalHealth        0.0046946  0.0009894   4.745 2.08e-06 ***
DiffWalkingYes      0.1958066  0.0203425   9.626  < 2e-16 ***
SexMale             0.7183172  0.0163257  43.999  < 2e-16 ***
AgeCategory25-29    0.1345623  0.1419976   0.948 0.343314
AgeCategory30-34    0.4766847  0.1272601   3.746 0.000180 ***
AgeCategory35-39    0.6005253  0.1217014   4.934 8.04e-07 ***
AgeCategory40-44    1.0230442  0.1142758   8.952  < 2e-16 ***
AgeCategory45-49    1.3435366  0.1103239  12.178  < 2e-16 ***
AgeCategory50-54    1.7801817  0.1064447  16.724  < 2e-16 ***
AgeCategory55-59    2.0062542  0.1048671  19.131  < 2e-16 ***
AgeCategory60-64    2.2845783  0.1039234  21.983  < 2e-16 ***
AgeCategory65-69    2.5199432  0.1036303  24.317  < 2e-16 ***
AgeCategory70-74    2.8095883  0.1035570  27.131  < 2e-16 ***
AgeCategory75-79    3.0046664  0.1041518  28.849  < 2e-16 ***
AgeCategory80 or older  3.2566497  0.1038827  31.349  < 2e-16 ***
RaceAsian          -0.5125952  0.0925055  -5.541 3.00e-08 ***
RaceBlack          -0.3968159  0.0643837  -6.163 7.12e-10 ***
RaceHispanic       -0.2985811  0.0655792  -4.553 5.29e-06 ***
RaceOther          -0.1165961  0.0713955  -1.633 0.102448
RaceWhite          -0.1187634  0.0573349  -2.071 0.038321 *
DiabeticBL          0.1431830  0.0468454   3.057 0.002239 **
DiabeticYes         0.4823947  0.0187059  25.788  < 2e-16 ***
DiabeticYesPreg     0.0544857  0.1198488   0.455 0.649382
PhysicalActivityYes 0.0298776  0.0180095   1.659 0.097117 .
GenHealth2          1.5092180  0.0366381  41.193  < 2e-16 ***
GenHealth3          1.0323838  0.0329548  31.327  < 2e-16 ***
GenHealth1          1.8878382  0.0457246  41.287  < 2e-16 ***
GenHealth4          0.4475175  0.0338513  13.220  < 2e-16 ***
SleepTime          -0.0263790  0.0048746  -5.412 6.25e-08 ***
AsthmaYes           0.2709828  0.0215686  12.564  < 2e-16 ***
KidneyDiseaseYes    0.5520620  0.0274444  20.116  < 2e-16 ***
SkinCancerYes       0.1037718  0.0218499   4.749 2.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 149034  on 255835  degrees of freedom
Residual deviance: 115592  on 255798  degrees of freedom
AIC: 115668

Number of Fisher Scoring iterations: 7
```

Conclusion

- Akaike information criterion (AIC): 115668
- Accuracy rate: 0.9140856
  - This is a relatively good result
  - Better than trivial predictions
- High false negative rate is difficult to avoid when p is close to 0
  - Similarly, a high false positive rate is difficult to avoid when p is close to 1

Classification Table (Test)

No	Yes
58381	5578

$p_{test} = 0.08721212$

$1 - p_{test} = 0.91278788$

Future Improvements

- Using other statistical methods outside of this class
- Using different combinations of variables and/or interaction terms
- Changing baseline levels of factor variables
- Analysis of outliers and influential points/reviewing assumptions
    - Reviewing “unusual” survey responses
- Transforming variables like BMI and Diabetes

References

1. “Heart Disease Facts.” *Centers for Disease Control and Prevention*, 14 Oct. 2022, https://www.cdc.gov/heartdisease/facts.htm.
    a. Centers for Disease Control and Prevention, National Center for Health Statistics. About Multiple Cause of Death, 1999–2020. CDC WONDER Online Database website. Atlanta, GA: Centers for Disease Control and Prevention; 2022. Accessed February 21, 2022.
2. Pytlak, Kamil. “Personal Key Indicators of Heart Disease.” Kaggle, 16 Feb. 2022, https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.

Questions?