Charles Hwang
Dr. Xi
STAT 408-001
18 December 2022

Personal Key Indicators of Heart Disease Final Project Report

Introduction

The dataset I chose for this project is the "Personal Key Indicators of Heart Disease" dataset (Pytlak, 2022) on Kaggle, a website with free-to-use datasets for statistical analysis. The goal of this project is to create a regression model using the variables in the dataset to predict whether or not a respondent has heart disease. The nature of regression analysis makes it difficult to determine the "best" possible model, as this depends on the context of the problem and goals of the model (e.g., whether to minimize the false positive rate or the false negative rate). However, criteria like the Akaike information criterion (AIC) can be used to help guide this process.

The motivation behind choosing the dataset was mainly due to its stability and security. The data were collected by a federal government agency, which may be subject to more data cleaning and regulations than the average dataset. The data are also relatively recent, having been collected approximately two-and-a-half years ago. The dataset is publicly available for download on Kaggle, a reliable website which I have used for class projects in prior semesters, and (as of the time this report was completed) has 240,616 views and 35,688 downloads (Pytlak, 2022). The significance of the topic of this dataset is that heart disease is the leading cause of death in the United States and has been for at least two decades ("Heart", 2022), which has been a longstanding cause of concern for cardiologists and doctors worldwide.

Dataset

The original dataset was compiled from a 2020 telephone survey of 401,958 United States residents with 279 variables taken by the Centers for Disease Control and Prevention (CDC). This survey to gather health data is taken annually across "all 50 states as well as the District of Columbia and three U.S. territories" and the CDC claims it is "the largest continuously conducted health survey system in the world" (Pytlak, 2022). However, the dataset on Kaggle only contains 319,795 observations and 18 variables as the user who originally published the dataset, Kamil Pytlak, removed a proportion of rows and columns. In explaining the reasoning and justification behind this decision, Pytlak writes: "I noticed many different factors (questions) that directly or indirectly influence heart disease, so I decided to select the most relevant variables from it and do some cleaning so that it would be usable for machine learning projects." The 18 variables in the dataset are the following:

1. Whether a respondent has reported ever having heart disease (binary),
    a. A respondent is reported as having had heart disease if they ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
2. Body mass index (kg/m^2),
3. Whether a respondent smoked 100+ cigarettes in their life (binary),
4. Heavy ("Alcohol", 2022) alcohol consumption (binary),
5. Whether a respondent ever had a stroke (binary),
6. Number of days of poor physical health in the last 30 days (discrete),
7. Number of days of poor mental health in the last 30 days (discrete),
8. Whether a respondent has difficulty walking (binary),
9. Sex (female, male),
10. Age category (18-24, 25-29, 30-34, …, 75-79, 80 or older),
11. Race (White, Hispanic, Black, Asian, American Indian/Alaska native, other),
12. Whether a respondent has diabetes (none, borderline, during pregnancy, yes),
13. Whether a respondent has reported doing physical activity or exercise in the last 30 days other than as part of their regular job (binary),
14. General health (excellent, very good, good, fair, poor),
15. Average hours of sleep (discrete),
16. Whether a respondent has ever had asthma (binary),
17. Whether a respondent has ever had kidney disease (binary), and
18. Whether a respondent has ever had skin cancer (binary).

We can see that there are four quantitative variables: (2) BMI, (6) physical health, (7) mental health, and (15) sleep time. According to the webpage on Kaggle, there are no mismatched or missing data out of all 319,795 * 18 = 5,756,310 values.

In visualizing the data, Kaggle is helpful in this regard as the webpage provides pie charts for Boolean (yes/no) variables (variables 1, 3-5, 8, 13, and 16-18), proportions of the two most common levels for non-Boolean factor variables (variables 9-12, and 14), and histograms for quantitative variables (variables 2, 6-7, and 15). Kaggle also provides the number of levels for non-Boolean factor variables and the mean, standard deviation, and interquartile range (IQR) for quantitative variables. The pie chart for the (1) heart disease variable and a table I created show approximately 8.6 percent of respondents (27,373) in the data were reported as being diagnosed with heart disease. With there being 14 categorical variables, I did not find it practical or meaningful to compare all ${}_{14}C_2 = 91$ pairs of variables, so I did not print any frequency tables.

I created a boxplot for (2) BMI and saw it had several unreasonable outliers with a maximum of 94.85 (note that a BMI of 40+ is defined by the CDC as "class III obesity"). After inspecting the equation for BMI (kg/m^2) and entering weight and height values via trial-and-error, we can see a value this high is virtually impossible. I inferred observations like this could be due to respondents and/or data collectors inadvertently entering weight, either in pounds or

kilograms, instead of BMI. I also created boxplots for the other three quantitative variables and saw there were several values considered statistical outliers. For example, it is possible to have 6+ days of poor physical health and/or 8+ days of poor mental health in the last 30 days, but it seemed unreasonable for a legal adult to sleep for 11+ hours per day *on average*. I inferred observations like this could be from misunderstanding or misinterpreting the question (e.g., thinking it was asking about the hours of sleep last night instead of on average, inverse question (hours awake), etc.) or unserious responses. The histograms on Kaggle also showed a local mode at 30 for (6-7) days of poor mental and physical health in the last 30 days (19,505 responses), which could similarly be from misunderstanding the question as the inverse or a reactionary/spontaneous pessimistic response.

I initially produced a plot matrix of the quantitative variables, but the large sample size made it impractical to interpret as almost all points overlapped with one another making it difficult to visualize any trends. In place of this, I produced a correlation matrix to see the relationships between the quantitative variables. The greatest linear correlation was between (6) days of poor physical and (7) mental health ($r = 0.2879867$), which indicates there is little to no linear correlation between any two of the quantitative variables. This made sense looking at the variables themselves, as it did not appear any of them are strongly correlated with one another, but it is possible the small values could be due to the large sample size of the dataset. Either way, the ${}_4C_2 = 6$ values of r in relation to each other also made sense intuitively.

Methods

I first read the dataset into *RStudio* and set the categorical variables as factor variables. For the (14) general health variable, *R* initially assigned the levels in alphabetical order, so I recoded them to match the implied levels with “excellent” being 5, “very good” being 4, “good” being 3, “fair” being 2, and “poor” being 1. After conducting the above univariate and bivariate analyses, I split the data into training and test datasets with an 80/20 split. Since there were no missing data, there was no need to impute the data with mean or median imputation or any other method.

Since the response variable, (1) heart disease, is a binary categorical variable, the natural and clear choice of methods is to perform logistic regression on the training data. Before running the code in *R*, I checked the assumptions of the regression. The six main assumptions of logistic regression are (Leung, 2022):

1. Binary, multinomial, or ordinal response variable
2. Linearity between each independent variable and logit of response variable
3. Absence of strongly influential outliers
4. Absence of collinearity between independent variables
5. Independence of observations
6. Large sample size

From the "Dataset" section, we can see assumptions 1 and 4-6 are clearly met: the heart disease variable is binary, there is little to no collinearity between the quantitative variables, each respondent was surveyed independently, and the sample size is 319,795. I did not specifically check the assumption for 2. Linearity between each independent variable and logit of response variable. Leung suggests a Box-Tidwell Test or a visual inspection of the plots of the logit of the response vs. each independent variable, but I am unfamiliar with both methods as I had never heard of the assumption or how to check it in prior classes and did not want to perform incorrect analysis or provide a misleading interpretation. Further, looking at the quantitative variables, it seemed reasonable this assumption would not be badly violated, and in the overall scope of the analysis being performed, it did not appear there would be significant consequences if the assumption was violated. Thus, I chose to proceed with the logistic regression, as there is no apparent reason to believe the assumption is violated.

We can see the final assumption of 3. Absence of strongly influential outliers is likely not met, as there were numerous outliers. However, the assumption is only violated with the presence of observations that are both outliers *and* influential points, and it is likely that not all of the previously observed outliers are influential points. Additionally, it is not apparent what transformation would fix the issue as the variables for (6-7) days of poor mental and physical health in the last 30 days are discrete, and with only four quantitative variables, it does not appear the assumption is "badly" violated. Thus, I chose to cautiously proceed with the logistic regression, noting that results should be interpreted carefully in the case of any statistically significant findings.

In performing the regression, I used a generalized linear model (GLM) and set the *family* parameter to *binomial* to indicate a logistic regression. I included all independent variables in the dataset to predict the response variable. I used the *step* function to perform backward selection, but it resulted in the same full model. Nearly all of the variables in the model were significant at the $\alpha = 0.05$ level, with the exception of the "25-29" level for (10) age, the "other" level for (11) race, the "during pregnancy" level for (12) diabetes, and, perhaps surprisingly, (13) physical activity. I infer that the reasoning for the first three levels of variables not being significant could be due to a low sample size for the particular level, and after creating tables for each, we can see there are 102, 83, and 700 respondents in each level in the training dataset with heart disease, respectively. However, in regards to physical activity, recall the variable simply measures whether a respondent has reported doing physical activity or exercise in the last 30 days other than as part of their regular job. It is possible, even probable, that the reported proportion is considerably higher than the true proportion due to response bias, as respondents may be inclined to answer affirmatively to portray a more positive self-image. This bias can be amplified in telephone surveys especially with a human administrator rather than a computerized one. Also, when closely looking at the variable definition, it would not seem that doing physical activity in the last 30 days is a significant predictor for heart disease.

We can see the Akaike information criterion (AIC) of the model is approximately 115,667.7. Looking at the coefficients, we can see age is the strongest predictor, followed by general health, with coefficient estimates strictly increasing by age category and strictly decreasing with better self-reported general health categories. Stroke history was also a strong predictor, followed by the male level of sex, and kidney disease history and diabetes were moderately strong. Smoking, asthma history, and difficulty walking (with asthma history being equidistant from the other two) rounded out the considerable variables, all of which had positive coefficients. Interestingly, heavy alcohol consumption was negatively associated with heart disease, which may suggest the existence of one or more confounding variables not in the available dataset.

In an attempt to produce a model with a lower AIC and/or false negative rate, I went back and manually performed backward selection on the full model, removing non-significant variables one at a time while decreasing the critical significance level to $\alpha = 0.001$ to ensure a substantially different model was produced. The resulting reduced model had a higher AIC (125,422.9) and the predicted values resulted in a higher false negative rate (92.26 percent), but the accuracy rate improved to 92.62 percent and the false positive rate was reduced to less than 0.64 percent. The precision also fell slightly to approximately 53.66 percent. Although the accuracy rate improved considerably, the increase in false negative rate was concerning, especially given the context of the problem, and I chose to use the original full model as the final model for the "Results" section. I included the output for the AIC and confusion matrix of the reduced model, but suppressed the output of the manual backward selection process as comments to conserve space.

Results

We can see the accuracy rate of the full logistic regression model is approximately 91.41 percent. The sensitivity (true positive rate) is approximately 10.11 percent and the specificity (true negative rate) is approximately 99.18 percent. The precision (positive predictive value) is approximately 53.97 percent and subsequently the false discovery rate is approximately 46.03 percent. Finally, the false positive rate is approximately 0.82 percent and the false negative rate is approximately 89.89 percent. All univariate and bivariate analyses, boxplots, output, summaries, classification tables, and results can be found in the *R Markdown* (.Rmd) and .pdf files attached alongside this report in the Sakai submission.

Conclusion/Future Improvements

It appears a 91.4 percent accuracy rate is relatively good in practice. Of course, the purpose of prediction has influence on the model. If we wanted to minimize the false positive rate, we would simply say no respondent has heart disease, and conversely, if we wanted to minimize the false negative rate, we would simply say every respondent has heart disease. The

tradeoff between the two rates is similar to the bias-variance tradeoff encountered when fitting models. This is a common issue that arises in classification problems like binary logistic regression when p is very close to 0 or 1, which is the case with this dataset and many similar science-related topics like mammograms, epidemiology, pathology, meteorology, etc. When p is very small for binary response variables, a high false negative rate seems to be unavoidable, even in robust models, and similarly, a high false positive rate can be difficult to avoid when p is high. This makes sense intuitively when looking at how each are calculated.

There are many different improvements that I could make if given additional time. One of them is using other statistical methods outside the scope of this class to predict heart disease, including (but not limited to) classification trees, random forests, and gradient boosted models (GBM). Another is continuing to experiment with different combinations of variables or introducing interaction terms to see if they are significant in the model. A third is to review and potentially impute values for or remove the outliers and influential points from the quantitative variables as well as "unusual" responses collected in the survey to further infer the intention behind them and the reason(s) they were recorded a certain way. Influential points would still have undue influence in the model, but the large sample size of the dataset makes it statistically probable that at least some of these observations truly occurred and were not a result of data entry errors, unserious responses, or other human errors. This analysis could suggest a transformation of one or more of the quantitative variables. Finally, it may be helpful to review the results of the model among subsets of the data (e.g., only respondents with certain levels of diabetes, certain age groups, etc.) to see if the model had a considerably lower accuracy rate in prediction among certain levels of variables. This may suggest one or more variables are interfering with the model's prediction accuracy and justify collapsing some variables into a smaller number of levels or even creating a separate model solely to predict on the subset.

In conclusion, it appears this project was a success in using regression analysis to predict the presence of heart disease in United States residents from real-world data. This analysis utilized several topics discussed in class in a practical manner.

References/Bibliography

"Alcohol Use and Your Health." Centers for Disease Control and Prevention, 14 Apr. 2022, https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm.

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.

"Heart Disease Facts." *Centers for Disease Control and Prevention*, 14 Oct. 2022, https://www.cdc.gov/heartdisease/facts.htm.

Leung, Kenneth. "Assumptions of Logistic Regression, Clearly Explained." *Medium*, Towards Data Science, 13 Sept. 2022, https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290.

Pytlak, Kamil. "Personal Key Indicators of Heart Disease." Kaggle, 16 Feb. 2022, https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.