MA4128 Assignment

Colm Kenny 18225012

29/04/2022

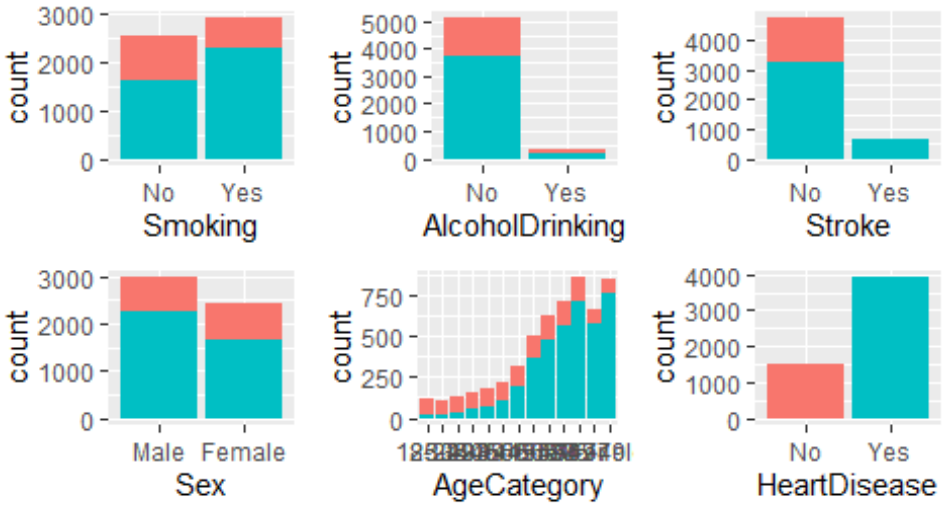
Question (a) - Technical report

The aim of my technical report is to understand which factors influence whether a person has heart disease or not. The dataset in question is courtesy of the CDC with 10 variables with 5420 observations. Discriminant analysis will be used to discriminate between people with heart disease or not. Due to the large number of categorical variables in the data and the binary dependent variable HeartDisease, I will use a logistic regression model. Before I start the formal analysis, I will perform some exploratory analysis on the data.

Exploratory analysis

Visual interpretation is important when analysing data and can produce some easily interpretable insights if executed correctly. These results can reinforce the results from an empirical analysis or provide some contrasting results. We have a large set of variables at our disposal making it tough to graphically interpret data. Especially due to the presence of 6 categorical variables. The first step I will take is to present charts of the categorical variables.

Categorical Variables



There is close to 4000 participants in the data with a heart disease. From looking at the categorical variables, it is easy to see that the older a person gets, the higher chance they have of having a heart disease. There are no clear signs that heart disease discriminates

against a person's sex or if they smoke. There are significantly more people with heart disease that are not heavy drinkers than are heavy drinkers. This has surprised me but like if a person has had a stroke or not, could be explained by the low numbers of people who have not had a stroke or are heavy drinkers. For these two variables it might be helpful to calculate and display the proportions.

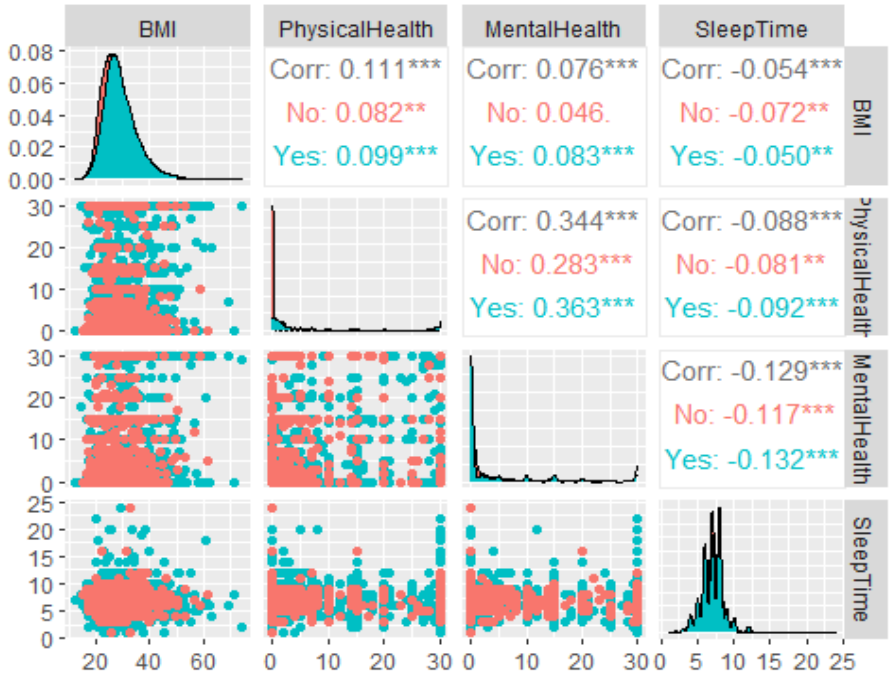
```
##             Stroke
## HeartDisease    No   Yes
##          No  0.309 0.056
##          Yes 0.691 0.944

##             AlcoholDrinking
## HeartDisease    No   Yes
##          No  0.269 0.419
##          Yes 0.731 0.581
```

Looking at the above, we can see that close to 95% of all people that have had a stroke, also have some form of heart disease. It is very rare that a person who has had a stroke before has never had a heart disease. 73% of all non-heavy drinkers have a heart disease. This is significantly greater than the 58% of heavy drinkers that have a heart disease. An explanation to this could be that heavy alcohol consumption could lead to other diseases instead of heart diseases e.g. liver problems.

Numerical Variables

Here I will look at the relationships between the numeric variables and the dependent variable HeartDisease. It is important to look at the correlations between the independent variables before we fit our logistic regression model. If there are strong relationships between some of the variables, then multi-collinearity exists, and we will need to remove some variables from the model.



Looking at the ggpairs plot, it appears that none of the variables significantly impact a persons chance of having heart disease. Each density plot on the diagonal do not show any major differences by response yes or no. Furthermore, none of the numeric independent variables appear to be strongly related. All correlation values are quite low, suggesting there is no multicollinearity in our data. We will not need to emit any numeric variables from our logistic regression model.

Initial Conclusions

From a visual interpretation, it appears that age and stroke play a major role in the chances of having heart disease. It appears that the numerical variables play little to no impact on a person's chances of having heart disease. We will be able to compare these results to our final results once we create a logistic regression model.

Formal analysis

Logistic regression is convenient for modelling binary dependent variables. Here we have a dependent variable with responses "yes" or "no". This perfectly supports the use of a Generalised Linear Model, with having heart disease coded as 1 and not having heart disease coded as 0. This would mean the values in [0,1] estimated by the model are the probabilities that a person has heart disease given certain values for the independent variables. When we fit our model, we are most concerned with the coefficient estimates and the p-values for each variable. Insignificant variables with high p-values could result in an inaccurate model, we will have to consider dropping variables. We will use these estimates to calculate the odds for each variable which are easier to interpret. The odds are used to interpret how the variables impact the chances of having heart disease or not. The odds are the probability of a person having heart disease over the probability of not having it. An odds value greater than 1 means a success (heart disease) is more likely while a value less than 1 means failure is more likely. Values close to or equal to 1 are seen as having no significant impact on the dependent variable as success and failure are equally likely. Now we can fit our logistic regression model. There are no correlated independent variables, therefore we will not have to drop any variables for the model.

```
##
## Call:
## glm(formula = HeartDisease ~ BMI + Smoking + AlcoholDrinking +
##     Stroke + PhysicalHealth + MentalHealth + Sex + AgeCategory +
##     SleepTime, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2167  -0.5354   0.4477   0.6864   2.4316
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -2.962050   0.382747  -7.739 1.00e-14 ***
## BMI                     0.045968   0.006153   7.471 7.96e-14 ***
## SmokingYes              0.542514   0.073734   7.358 1.87e-13 ***
## AlcoholDrinkingYes     -0.492916   0.148932  -3.310 0.000934 ***
## StrokeYes               1.696171   0.180869   9.378  < 2e-16 ***
## PhysicalHealth          0.036624   0.004626   7.917 2.43e-15 ***
## MentalHealth            0.015116   0.004946   3.056 0.002241 **
```

```
## SexFemale              -0.641028   0.073979  -8.665  < 2e-16 ***
## AgeCategory25-29        0.488883   0.392249   1.246 0.212632
## AgeCategory30-34        0.577639   0.366270   1.577 0.114776
## AgeCategory35-39        1.011814   0.348554   2.903 0.003697 **
## AgeCategory40-44        1.209504   0.339770   3.560 0.000371 ***
## AgeCategory45-49        1.368127   0.334315   4.092 4.27e-05 ***
## AgeCategory50-54        2.095564   0.322394   6.500 8.03e-11 ***
## AgeCategory55-59        2.549408   0.316028   8.067 7.20e-16 ***
## AgeCategory60-64        2.765321   0.313430   8.823  < 2e-16 ***
## AgeCategory65-69        3.005873   0.312235   9.627  < 2e-16 ***
## AgeCategory70-74        3.361447   0.312405  10.760  < 2e-16 ***
## AgeCategory75-79        3.736930   0.322722  11.579  < 2e-16 ***
## AgeCategory80 or older  4.205473   0.320935  13.104  < 2e-16 ***
## SleepTime              -0.057022   0.022958  -2.484 0.013002 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6399.8  on 5419  degrees of freedom
## Residual deviance: 4808.4  on 5399  degrees of freedom
## AIC: 4850.4
##
## Number of Fisher Scoring iterations: 5
```

At first glance, the p-values are all quite low. There is significant evidence there is a real association between the independent variables and HeartDisease. We have no reason to drop any variables from the model. To interpret the coefficients, it will be easier to look at the odds values for the IV's.

```
##                             OR  2.5 %  97.5 %
## (Intercept)              0.052  0.024   0.107
## BMI                      1.047  1.035   1.060
## SmokingYes               1.720  1.489   1.988
## AlcoholDrinkingYes       0.611  0.457   0.819
## StrokeYes                5.453  3.877   7.892
## PhysicalHealth           1.037  1.028   1.047
## MentalHealth             1.015  1.006   1.025
## SexFemale                0.527  0.455   0.609
## AgeCategory25-29         1.630  0.762   3.578
## AgeCategory30-34         1.782  0.883   3.744
## AgeCategory35-39         2.751  1.420   5.611
## AgeCategory40-44         3.352  1.765   6.735
## AgeCategory45-49         3.928  2.093   7.820
## AgeCategory50-54         8.130  4.448  15.860
## AgeCategory55-59        12.800  7.101  24.696
## AgeCategory60-64        15.884  8.862  30.509
## AgeCategory65-69        20.204 11.302  38.728
## AgeCategory70-74        28.831 16.124  55.287
## AgeCategory75-79        41.969 22.973  81.983
## AgeCategory80 or older 67.052 36.845 130.574
## SleepTime                0.945  0.903   0.988
```

Most of the odds values are greater than 1, this means that the majority of our IV's have a positive relationship with the chances of having heart disease. When compared to males, females have a lower chance of having a heart disease reducing the odds by a factor of 0.527. If a person is a heavy drinker, the odds of them having a heart disease reduces by a factor of 0.611. The largest odds value is seen for people older than 80. Being 80 or older

increases a person's odds of having a heart disease by a factor of 67.052. The odds values for age in general increase as a person gets older. Variables such as BMI, SleepTime, PhysicalHealth and MentalHealth have odds values close to 1, which are inconclusive and might suggest they do not play a significant role in altering one's chances of having a heart disease. Finally, previously having a stroke increases one's chances of ever having a form of heart disease by a factor of 5.453.

```
##
## pred.class   No  Yes
##        No  0.13 0.04
##        Yes 0.15 0.68

## Classification Rate = 0.81
```

The model above has correctly classified 80% of the observations in the data. It is a better idea to use a cross-validation method in order to assess the accuracy of the logistic model. To do this we will need to split the dataset into two parts. The training data will be used to fit the logistic regression model and the test data will be used to predict from the model and find the classification rate. 75% of the data will go into the training dataset and the remaining data will be used in the test dataset.

```
##
## test.pred.class   No  Yes
##             No  0.12 0.04
##             Yes 0.15 0.69

## Classification Rate = 0.81
```

The two tables are similar. Our logistic model is quite good at predicting if a person has a heart disease or not. The cross-validation method correctly predicted ~81% of observations if they have heart disease or not. The model performs significantly better than random guessing.

Conclusions

The purpose of this technical analysis was to uncover any variables that influenced whether a person has a heart disease or not. The logistic regression model worked very well and can produce accurate predictions of whether a person has ever had a form of heart disease. Backed up by visual analysis, a person's age is the greatest predictor of a person having heart disease or not. Although it does not play a large part when under the age of 30, chances of having heart disease increase drastically among adults year after year. A person who has had a stroke before has a very high chance of also having some form of heart disease.The logistic model uncovered a difference in chances of having heart disease by a person's sex. Females are less likely to have heart disease than males.

Question (b) - Non-technical report

Heart disease is a major issue in today's world that can affect anyone at any time. If you were aware that certain aspects of your life were increasing your chance of heart disease, would you switch up your habits to prevent this? The CDC are concerned for the health of

the American public. They created a large dataset to gather data on a person's health by conducting telephone surveys. Data collected includes whether a respondent has ever had a heart disease or not, do they smoke, their age and sex.

In my technical report I attempted to uncover the extent to which each variable in the data affects a person's chances of having heart disease or not. This would not only help the CDC target certain groups to reduce their risk of getting heart disease but help the public understand what a box of cigarettes or two hours of sleep a night is doing to their body. To draw reasonable conclusions a suitable model must be created and tested. The model type created uses gathered data to return a person's chances of having heart disease given certain characteristics or factors.

The model returned accurate predictions of whether a person has heart disease or not and was significantly more reliable than guessing yes or no. The model was able to successfully classify 8 out of 10 participants, over an expected 50% success rate of guessing yes or no. Our report showed that age is the most important factor to impact a person's chances of having heart disease. This is expected but unfortunately there is not a lot we can do to prevent this. It is important to mind your health as you get older and especially when you get to the age of 80. People that are heavy drinkers surprisingly have less chance of having heart disease. An explanation to this could be that heavy alcohol consumption could lead to other diseases instead of heart diseases e.g., liver problems. Other aspects like smoking over 100 cigarettes in your life and previously having a stroke increase one's chances of having heart disease. Looking to curb numbers of smokers in the US could reduce the number of heart disease events. People who experience a stroke should consider monitoring their health more frequently. Heart disease appears to occur more frequently among males than females. For further study and to increase understanding of the problem, I would suggest the alleyway of looking to uncover the factors which differ by sex to answer the question why is heart disease more common among males than females?