MA4128 Assignment

Dara Corr, ID: 22275193

Question (a) - Technical report

The objective of this assignment is to determine what factors influence whether a person has heart disease or not. To determine these factors, a dataset from the Centre for Disease Control was analysed. The contents in this dataset are from survey responses from 5420 American citizens reporting their health condition and background.

The participants were asked a number of questions relating to their current health and their behaviours, such as whether they have heart disease, what their BMI is and have they smoked a lot throughout their life.

Exploratory analysis showed that of the 5420 individuals in this dataset, 72.27% of those sampled said they have heart disease. This statistic seems much higher that what would be expected from a simple random sample of the population but since identifying key factors that influence heart disease is the concern, it is unlikely to be an issue in this analysis. There is a good balance of males and females surveyed in all of the age cohorts in this dataset in general.

```
## # A tibble: 13 x 4
##    AgeCategory  Male Female total
##    <fct>       <int>  <int> <int>
##  1 18-24          70     51   121
##  2 25-29          57     47   104
##  3 30-34          77     57   134
##  4 35-39          86     74   160
##  5 40-44          93     91   184
##  6 45-49         108    108   216
##  7 50-54         165    153   318
##  8 55-59         270    227   497
##  9 60-64         364    253   617
## 10 65-69         432    279   711
## 11 70-74         508    346   854
## 12 75-79         366    288   654
## 13 80 or older   387    463   850
```

One key observation found from the exploratory analysis is that as age increased, the proportion of those with heart disease in each age cohort increased quite significantly.

```
## # A tibble: 13 x 3
##    AgeCategory have_disease total
##    <fct>              <int> <int>
##  1 18-24                 14   121
##  2 25-29                 21   104
##  3 30-34                 33   134
##  4 35-39                 54   160
##  5 40-44                 73   184
##  6 45-49                 99   216
##  7 50-54                197   318
##  8 55-59                358   497
##  9 60-64                469   617
## 10 65-69                556   711
## 11 70-74                706   854
## 12 75-79                574   654
## 13 80 or older          763   850
```

Further analysis will be useful to determine which age groups are most at risk to heart disease.

I used logistic regression to determine what variables contribute most to people developing heart disease. Logistic regression is a suitable model to use here since there are only 2 possible outcomes to having heart disease, “Yes” and “No”.

```
reg <- glm(HeartDisease ~ BMI + Smoking + AlcoholDrinking + Stroke + Physical
Health + MentalHealth + Sex + AgeCategory + SleepTime, data = heartdata, fami
ly = binomial)

summary(reg)

##
## Call:
## glm(formula = HeartDisease ~ BMI + Smoking + AlcoholDrinking +
##     Stroke + PhysicalHealth + MentalHealth + Sex + AgeCategory +
##     SleepTime, family = binomial, data = heartdata)
##
## 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)           -3.603079   0.385988  -9.335  < 2e-16 ***
## 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 **
## SexMale                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
```

From the summary table above, it appears that all of the variables are of significance here.

It appears that those aged 35 years or older may be more at risk to heart disease than those aged under 35 since all the age groups from 35 and over are significant in this model. Those who have previously had a stroke are also flagged at being very likely to have heart disease from the summary output of this logistic regression.

```
round(exp(reg$coefficients),2)
```

```
##            (Intercept)                    BMI             SmokingYes
##                   0.03                   1.05                   1.72
##     AlcoholDrinkingYes              StrokeYes         PhysicalHealth
##                   0.61                   5.45                   1.04
##           MentalHealth                SexMale       AgeCategory25-29
##                   1.02                   1.90                   1.63
##       AgeCategory30-34       AgeCategory35-39       AgeCategory40-44
##                   1.78                   2.75                   3.35
##       AgeCategory45-49       AgeCategory50-54       AgeCategory55-59
##                   3.93                   8.13                  12.80
##       AgeCategory60-64       AgeCategory65-69       AgeCategory70-74
##                  15.88                  20.20                  28.83
##       AgeCategory75-79 AgeCategory80 or older              SleepTime
##                  41.97                  67.05                   0.94
```

The odds ratios above from the logistic regression coefficient outputs provide insight into what conditions and behaviours are more likely to cause heart disease.

Someone who is in the age category of 65-69 years is over 19 times more likely to have heart disease than someone who is not in that age group. We see the age ratios for the age groups increase for the older age groups, which indicates that older people are at a much greater risk of heart disease.

Someone who has previously had a stroke is also very likely to have heart disease, from the output above they are deemed 4.45 - almost four and a half - times more likely to have heart disease.

From a healthcare perspective, it is important to investigate what behaviours or habits are likely to cause heart disease in the population. The logistic regression model with all variables included - reg - is valuable in assessing what members of the population should have particular attention paid to regards cardiac health issues and heart disease screening. Since aging is impossible to prevent and strokes are likely to be a symptom of heart disease rather than a cause, another logistic regression model with Stroke and AgeCategory variables removed may provide valuable insight into the causes of heart disease, so that they can be identified.

```
reg2 <- glm(HeartDisease ~ BMI + Smoking + AlcoholDrinking + PhysicalHealth +
MentalHealth + Sex + SleepTime, data = heartdata, family = binomial)

summary(reg2)

##
## Call:
## glm(formula = HeartDisease ~ BMI + Smoking + AlcoholDrinking +
##     PhysicalHealth + MentalHealth + Sex + SleepTime, family = binomial,
##     data = heartdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6127  -1.1898   0.6427   0.8601   1.5280
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.795144   0.222618  -3.572 0.000355 ***
## BMI                 0.029061   0.005402   5.379 7.49e-08 ***
## SmokingYes          0.684476   0.064875  10.551  < 2e-16 ***
## AlcoholDrinkingYes -0.815725   0.129585  -6.295 3.08e-10 ***
## PhysicalHealth      0.054156   0.004344  12.466  < 2e-16 ***
## MentalHealth       -0.012493   0.004140  -3.017 0.002549 **
## SexMale             0.422588   0.064435   6.558 5.44e-11 ***
## SleepTime           0.029312   0.020240   1.448 0.147557
## ---
## 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: 5913.0  on 5412  degrees of freedom
## AIC: 5929
##
## Number of Fisher Scoring iterations: 4
```

From the Summary of the regression model with AgeCategory and Stroke variables omitted - reg2 - it appears that heavy drinking and poor mental health do not increase someone's likelihood of having heart disease. This is because these variables have negative log-likelihoods associated with them, meaning that the likelihood of an individual having heart disease decreases if they are a heavy drinker or they have had several poor experiences with poor mental health in the period they were surveyed.

```
coefs <- round(exp(reg2$coefficients),2)
coefs
```

```
##        (Intercept)                BMI         SmokingYes AlcoholDrinkingYe
s
##               0.45               1.03               1.98               0.4
4
##     PhysicalHealth       MentalHealth            SexMale          SleepTim
e
##               1.06               0.99               1.53               1.0
3
```

```
CI <- round(exp(confint(reg2)),2)
```

```
## Waiting for profiling to be done...
```

```
CI_list <- cbind(coefs, CI)
CI_list
```

```
##                    coefs 2.5 % 97.5 %
## (Intercept)         0.45  0.29   0.70
## BMI                 1.03  1.02   1.04
## SmokingYes          1.98  1.75   2.25
## AlcoholDrinkingYes  0.44  0.34   0.57
## PhysicalHealth      1.06  1.05   1.06
## MentalHealth        0.99  0.98   1.00
## SexMale             1.53  1.34   1.73
## SleepTime           1.03  0.99   1.07
```

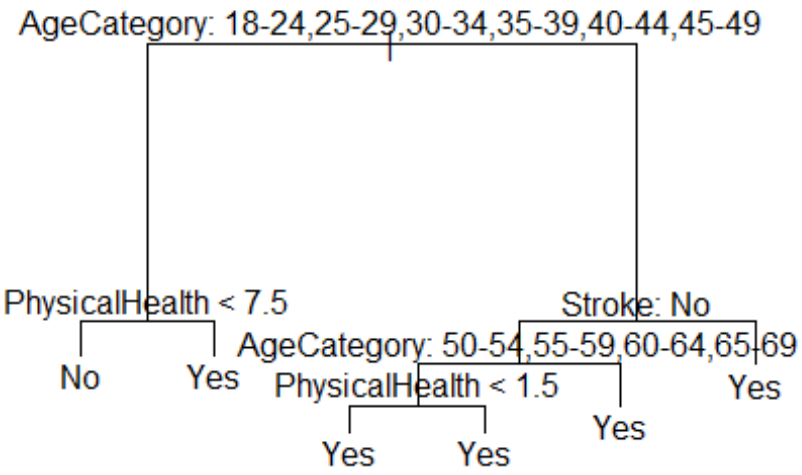
The odds ratios above and their associated 95% confidence intervals give a good idea of what the main causes of heart disease are from those surveyed. Smoking is the largest value here, with a value of 1.98, meaning that someone who smokes is 98% more likely to develop heart disease than someone who does not smoke, making a significant contributor to heart diseases for American adults. Males are also deemed 53% more likely to develop heart disease than women from this model.

Physical health is also an important contributor to heart disease. A BMI increase of 1 for an individual means that they are 3% more likely to develop heart disease. The variable PhysicalHealth is a measure of how many days an individual felt they were in poor physical

health in the past 30 days before being surveyed. An individual who had one day with poor physical health in the 30 days before the survey is 6% more likely to have heart disease. This means that individuals with a BMI of 35 (which is considered obese) are considered 30% more likely to have heart disease than someone with a BMI of 25.

The logistic regression models predicts that elderly people or someone who has had a stroke are the most likely to have heart disease. Smoking, having poor physical health and being over-weight are some of the other key factors that lead to heart disease in US adults according to this model. This model also predicts that males are also more likely than females to have heart disease.

I also used decision trees in my analysis of this dataset.



I used a classification tree, as I want to find out what variables will classify an individual as having heart disease or not. This decision trees predicts that someone who is aged over 50 years old is predicted to have heart disease. It also predicts that someone aged under 50 years old is predicted not to have heart disease, unless they have had some physical illness for more than 7 days in the last 30 days when surveyed.

This means that the decision tree model predicts that people over age 50 are most likely to be at risk of heart disease and people under the age of 50 are not deemed likely to be at risk of heart disease unless they have physical health problems.

This method oversimplifies predictions compared to logistic regression, but it does give a better visualisation of the predictions it makes. While logistic regression provides more insight into each parameter's contribution to the likelihood of an individual getting heart disease, the decision tree model very effectively narrows down the list of parameters to a few that are deemed most important to predicting the outcome and puts it in a visual format which is easy to interpret.

Exploratory analysis

From exploratory analysis of this dataset, it appeared that older generations are more at risk of heart disease. Exploratory analysis showed that 72.27% of the sample had heart disease, which is a large proportion of the dataset. This is deemed not to be a major issue in this analysis due to a sufficiently large dataset of 5420 individuals and due to the classification nature of the problem – trying to identify which individuals are at risk of heart disease.

Formal analysis

From my analysis I conclude that the people who are identified as most at risk of heart disease are individuals who have had a stroke in the past or individuals who are aged over 50. Age and Stroke are deemed the most significant variables in the logistic regression analysis along with sex, smoking, BMI, and physical illness.

The decision tree identified age, stroke, and physical health as the main predictors for having heart disease.

Smoking, having poor physical health (or physical illness) and being over-weight are identified as key factors that cause heart disease in American Adults. Variables associated with these factors have the high odds-ratios from the logistic regression analysis and are deemed very significant.

Conclusions

From my analysis I conclude that the people who are identified as most at risk of heart disease are individuals who have had a stroke in the past or individuals who are aged over 50. Smoking, having poor physical health/illness and being over-weight are key factors that cause heart disease in American Adults.

Males are about 50% more likely than females to be diagnosed with heart disease. Smokers are almost twice as likely to get heart disease than members of the population who do not smoke.

People who have suffered a stroke are more than 5 times more likely to have heart disease than someone who has not had a stroke.

Question (b) - Non-technical report

Using Statistical techniques of Logistic Regression and Decision trees, I found that the people most at risk of heart disease are those who have previously suffered a stroke and those who are aged over 50.

The key factors I identified as causes of heart disease from my analysis are smoking, having physical illness/poor physical health and being over-weight are key factors that cause heart disease in adults.

Males are about 50% more likely than females to be diagnosed with heart disease – although further analysis may be needed to determine if this discrepancy is due to anatomical factors or due to factors related to habits that men tend to adopt more than women do. Smokers are almost twice as likely to get heart disease than members of the population who do not smoke.

People who have suffered a stroke are more than 5 times more likely to have heart disease than someone who has not had a stroke.

Further analysis may be done to determine the reason behind the difference in likelihood of males and females developing heart disease.