Heart Disease Analysis

Oscar Monroy & Santiago Rodriguez

#Logistic Regression

Let's load our data and count the duplicated responses. Note the survey responses were not given a unique identifier, which leads to a lot of duplicate entries.

```
hd <- read.csv("heart_2020_cleaned.csv")
attach(hd)

summary(hd)
```

```
##  HeartDisease            BMI          Smoking          AlcoholDrinking
##  Length:319795      Min.   :12.02   Length:319795      Length:319795
##  Class :character   1st Qu.:24.03   Class :character   Class :character
##  Mode  :character   Median :27.34   Mode  :character   Mode  :character
##                     Mean   :28.33
##                     3rd Qu.:31.42
##                     Max.   :94.85
##     Stroke          PhysicalHealth    MentalHealth    DiffWalking
##  Length:319795      Min.   : 0.000   Min.   : 0.000   Length:319795
##  Class :character   1st Qu.: 0.000   1st Qu.: 0.000   Class :character
##  Mode  :character   Median : 0.000   Median : 0.000   Mode  :character
##                     Mean   : 3.372   Mean   : 3.898
##                     3rd Qu.: 2.000   3rd Qu.: 3.000
##                     Max.   :30.000   Max.   :30.000
##      Sex            AgeCategory            Race             Diabetic
##  Length:319795      Length:319795      Length:319795      Length:319795
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  PhysicalActivity    GenHealth            SleepTime         Asthma
##  Length:319795      Length:319795      Min.   : 1.000   Length:319795
##  Class :character   Class :character   1st Qu.: 6.000   Class :character
##  Mode  :character   Mode  :character   Median : 7.000   Mode  :character
##                                        Mean   : 7.097
##                                        3rd Qu.: 8.000
##                                        Max.   :24.000
##  KidneyDisease       SkinCancer
##  Length:319795      Length:319795
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
```

```
##
##
```

```
sum(duplicated(hd))
```

```
## [1] 18078
```

For some reason the ‘Yes’ and ‘No’ replies in the data were not being understood very well by R, so we converted *HeartDisease* into a binary vector.

```
library(dplyr)
```

```
## Warning: package ’dplyr’ was built under R version 4.1.3
```

```
##
## Attaching package: ’dplyr’
```

```
## The following objects are masked from ’package:stats’:
##
##     filter, lag
```

```
## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
hd2 <- hd %>% mutate(HeartDisease = ifelse(HeartDisease == 'Yes', 1, 0))
y <- hd2$HeartDisease
```

Now let’s break up the data into training and test sets. Here we used 60% of the data as our training set.

```
set.seed(1)

# Create training and test sets.
train <- sample(1:nrow(hd2), 0.6*nrow(hd2))
test <- (-train)

y.test <- y[test]
```

Now we apply Logistic Regression to the training set and then attempt to predict the individuals with heart disease. We use every other variable as a predictor.

```
#Apply logistic regression to training set.
model = glm(HeartDisease ~ ., data = hd2[train, ], family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = HeartDisease ~ ., family = binomial, data = hd2[train,
##     ])
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1247  -0.4119  -0.2435  -0.1293   3.6251
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -6.372975   0.151645 -42.026  < 2e-16 ***
## BMI                              0.009917   0.001471   6.743 1.56e-11 ***
## SmokingYes                       0.353749   0.018554  19.066  < 2e-16 ***
## AlcoholDrinkingYes              -0.211623   0.042632  -4.964 6.91e-07 ***
## StrokeYes                        1.036344   0.029155  35.546  < 2e-16 ***
## PhysicalHealth                   0.001949   0.001119   1.742   0.0814 .
## MentalHealth                     0.004936   0.001139   4.335 1.46e-05 ***
## DiffWalkingYes                   0.181376   0.023533   7.707 1.28e-14 ***
## SexMale                          0.713248   0.018816  37.907  < 2e-16 ***
## AgeCategory25-29                 0.228106   0.162578   1.403   0.1606
## AgeCategory30-34                 0.610439   0.145790   4.187 2.83e-05 ***
## AgeCategory35-39                 0.693738   0.140553   4.936 7.98e-07 ***
## AgeCategory40-44                 1.016132   0.134042   7.581 3.44e-14 ***
## AgeCategory45-49                 1.359589   0.128973  10.542  < 2e-16 ***
## AgeCategory50-54                 1.807914   0.124596  14.510  < 2e-16 ***
## AgeCategory55-59                 2.046812   0.122800  16.668  < 2e-16 ***
## AgeCategory60-64                 2.320468   0.121719  19.064  < 2e-16 ***
## AgeCategory65-69                 2.540422   0.121416  20.923  < 2e-16 ***
## AgeCategory70-74                 2.835324   0.121330  23.369  < 2e-16 ***
## AgeCategory75-79                 3.058927   0.121960  25.081  < 2e-16 ***
## AgeCategory80 or older           3.319197   0.121669  27.281  < 2e-16 ***
## RaceAsian                       -0.568826   0.109343  -5.202 1.97e-07 ***
## RaceBlack                       -0.372039   0.074328  -5.005 5.57e-07 ***
## RaceHispanic                    -0.234305   0.075431  -3.106   0.0019 **
## RaceOther                       -0.112870   0.082820  -1.363   0.1729
## RaceWhite                       -0.101819   0.066431  -1.533   0.1253
## DiabeticNo, borderline diabetes  0.132430   0.054126   2.447   0.0144 *
## DiabeticYes                      0.464219   0.021624  21.468  < 2e-16 ***
## DiabeticYes (during pregnancy)   0.198279   0.130973   1.514   0.1301
## PhysicalActivityYes             -0.002650   0.020718  -0.128   0.8982
## GenHealthFair                    1.523867   0.042368  35.967  < 2e-16 ***
## GenHealthGood                    1.049275   0.038107  27.535  < 2e-16 ***
## GenHealthPoor                    1.938582   0.052650  36.820  < 2e-16 ***
## GenHealthVery good               0.464610   0.039172  11.861  < 2e-16 ***
## SleepTime                       -0.025361   0.005628  -4.506 6.60e-06 ***
## AsthmaYes                        0.284644   0.024689  11.529  < 2e-16 ***
## KidneyDiseaseYes                 0.566081   0.031419  18.017  < 2e-16 ***
## SkinCancerYes                    0.117543   0.025153   4.673 2.97e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 112154  on 191876  degrees of freedom
## Residual deviance:  87066  on 191839  degrees of freedom
## AIC: 87142
##
## Number of Fisher Scoring iterations: 7
```

```r
# Predict the responders that are diagnosed with heart disease.
model.pred = predict(model, data = hd2[test, ], type = 'response')
vec = rep(0, length(y))
vec[model.pred >= 0.5] = 1
```

Now let's check how well Logistic Regression predicted $HeartDisease$

```r
mean(vec == y)
```

```
## [1] 0.9007051
```

```r
prop.table(table(vec, y))
```

```
##    y
## vec           0           1
##   0 0.899157273 0.084047593
##   1 0.015247268 0.001547867
```

Notice that our model did very poorly at detecting which individuals said they had heart disease at some point.