Predicting Heart Disease Risk Using Clinical Variables

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

## Warning: package 'tidyr' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(glmnet)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1
```

```
library(pwr)
theme_set(theme_bw())
library(mgcv)
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## This is mgcv 1.8-33. For overview type 'help("mgcv-package")'.
```

```
library(ggplot2)
```

```
df <- read.csv(file = 'Heart_Disease_Prediction.csv')

head(df,n=5)
```

```
##   index Age Sex Chest.pain.type  BP Cholesterol FBS.over.120 EKG.results Max.HR
## 1     0  70   1               4 130         322            0           2    109
## 2     1  67   0               3 115         564            0           2    160
## 3     2  57   1               2 124         261            0           0    141
## 4     3  64   1               4 128         263            0           0    105
## 5     4  74   0               2 120         269            0           2    121
##   Exercise.angina ST.depression Slope.of.ST Number.of.vessels.fluro Thallium
## 1               0           2.4           2                       3        3
## 2               0           1.6           2                       0        7
## 3               0           0.3           1                       0        7
## 4               1           0.2           2                       1        7
## 5               1           0.2           1                       1        3
##   Heart.Disease
## 1      Presence
## 2       Absence
## 3      Presence
## 4       Absence
## 5       Absence
```

```
df$Heart.Disease <-ifelse(df$Heart.Disease=="Presence",1,0)
head(df,n=5)
```

```
##   index Age Sex Chest.pain.type  BP Cholesterol FBS.over.120 EKG.results Max.HR
## 1     0  70   1               4 130         322            0           2    109
## 2     1  67   0               3 115         564            0           2    160
## 3     2  57   1               2 124         261            0           0    141
## 4     3  64   1               4 128         263            0           0    105
## 5     4  74   0               2 120         269            0           2    121
##   Exercise.angina ST.depression Slope.of.ST Number.of.vessels.fluro Thallium
```

```
## 1               0           2.4           2                       3        3
## 2               0           1.6           2                       0        7
## 3               0           0.3           1                       0        7
## 4               1           0.2           2                       1        7
## 5               1           0.2           1                       1        3
##   Heart.Disease
## 1             1
## 2             0
## 3             1
## 4             0
## 5             0
```

Power

```
# calculate minimal sample size
pwr.anova.test(k=2,             # 5 groups are compared
               f=.25,           # moderate effect size
               sig.level=.05,  # alpha/sig. level = .05
               n=270)           # n of participants
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 2
##               n = 270
##               f = 0.25
##       sig.level = 0.05
##           power = 0.9999383
##
## NOTE: n is number in each group
```

```
#general linear model
pwrglm <- pwr.f2.test(u = 1,    #the degrees of freedom for numerator (´u´)
                      v = 58,
                      f2 = .02,
                      sig.level = 0.05)
# inspect results
pwrglm
```

```
##
##      Multiple regression power calculation
##
##               u = 1
##               v = 58
##              f2 = 0.02
##       sig.level = 0.05
##           power = 0.1899206
```

Lasso

```
set.seed(123)
training.samples <- df$Heart.Disease %>%
```

```
  createDataPartition(p = 0.8, list = FALSE)

train.data <- df[training.samples, ]
test.data <- df[-training.samples, ]

x <- model.matrix(Heart.Disease~., train.data)[,-1]
y <- train.data$Heart.Disease
```

```
# lambda
set.seed(123)
cv <- cv.glmnet(x, y, alpha = 1)

cv$lambda.min
```

```
## [1] 0.01495783
```

```
model <- glmnet(x, y, alpha = 1, lambda = cv$lambda.min)

coef(model)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                      s0
## (Intercept)             -0.489557936
## index                    .
## Age                      .
## Sex                      0.115030911
## Chest.pain.type          0.062496122
## BP                       0.001824142
## Cholesterol              0.001007854
## FBS.over.120             .
## EKG.results              0.038544639
## Max.HR                  -0.002771325
## Exercise.angina          0.082208761
## ST.depression            0.036176006
## Slope.of.ST              0.086140180
## Number.of.vessels.fluro  0.096402283
## Thallium                 0.058856207
```

```
x.test <- model.matrix(Heart.Disease ~., test.data)[,-1]
predictions <- model %>% predict(x.test) %>% as.vector()

data.frame(
  RMSE = RMSE(predictions, test.data$Heart.Disease),
  Rsquare = R2(predictions, test.data$Heart.Disease)
)
```

```
##        RMSE   Rsquare
## 1 0.3718633 0.4420803
```

Logistic Regression

```r
dfl <- na.omit(df)

sample_n(dfl, 3)
```

```
##   index Age Sex Chest.pain.type  BP Cholesterol FBS.over.120 EKG.results Max.HR
## 1   185  43   1               3 130         315            0           0    162
## 2    25  48   0               3 130         275            0           0    139
## 3    26  46   0               4 138         243            0           2    152
##   Exercise.angina ST.depression Slope.of.ST Number.of.vessels.fluro Thallium
## 1               0           1.9           1                       1        3
## 2               0           0.2           1                       0        3
## 3               1           0.0           2                       0        3
##   Heart.Disease
## 1             0
## 2             0
## 3             0
```

```r
set.seed(123)
training.samples <- dfl$Heart.Disease %>%
  createDataPartition(p = 0.8, list = FALSE)

train.data <- dfl[training.samples, ]
test.data <- dfl[-training.samples, ]
```

```r
#Fit the model

model <- glm(Heart.Disease ~., data = train.data, family = binomial)

summary(model)
```

```
##
## Call:
## glm(formula = Heart.Disease ~ ., family = binomial, data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5748  -0.4770  -0.1547   0.3844   2.4953
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -9.8729577  3.4497236  -2.862 0.004210 **
## index                  -0.0007223  0.0028122  -0.257 0.797307
## Age                    -0.0183268  0.0290462  -0.631 0.528071
## Sex                     1.4782197  0.6178922   2.392 0.016740 *
## Chest.pain.type         0.5537920  0.2416921   2.291 0.021945 *
## BP                      0.0293383  0.0129706   2.262 0.023703 *
## Cholesterol             0.0138029  0.0053877   2.562 0.010410 *
## FBS.over.120           -0.5776730  0.6643345  -0.870 0.384546
## EKG.results             0.3813842  0.2250396   1.695 0.090124 .
## Max.HR                 -0.0252001  0.0119909  -2.102 0.035588 *
## Exercise.angina         0.6999530  0.4855330   1.442 0.149410
## ST.depression           0.2295083  0.2654397   0.865 0.387239
```

```
## Slope.of.ST              0.8576083  0.4695956   1.826 0.067810 .
## Number.of.vessels.fluro  1.0355024  0.3124193   3.314 0.000918 ***
## Thallium                 0.3423278  0.1260026   2.717 0.006591 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 297.59  on 215  degrees of freedom
## Residual deviance: 140.24  on 201  degrees of freedom
## AIC: 170.24
##
## Number of Fisher Scoring iterations: 6
```

```
probabilities <- model %>%
  predict(test.data, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "1", "0")

mean(predicted.classes == test.data$Heart.Disease)
```

```
## [1] 0.8148148
```

```
model2 <- glm(Heart.Disease ~ Number.of.vessels.fluro, data = train.data, family = binomial)
summary(model2)$coef
```

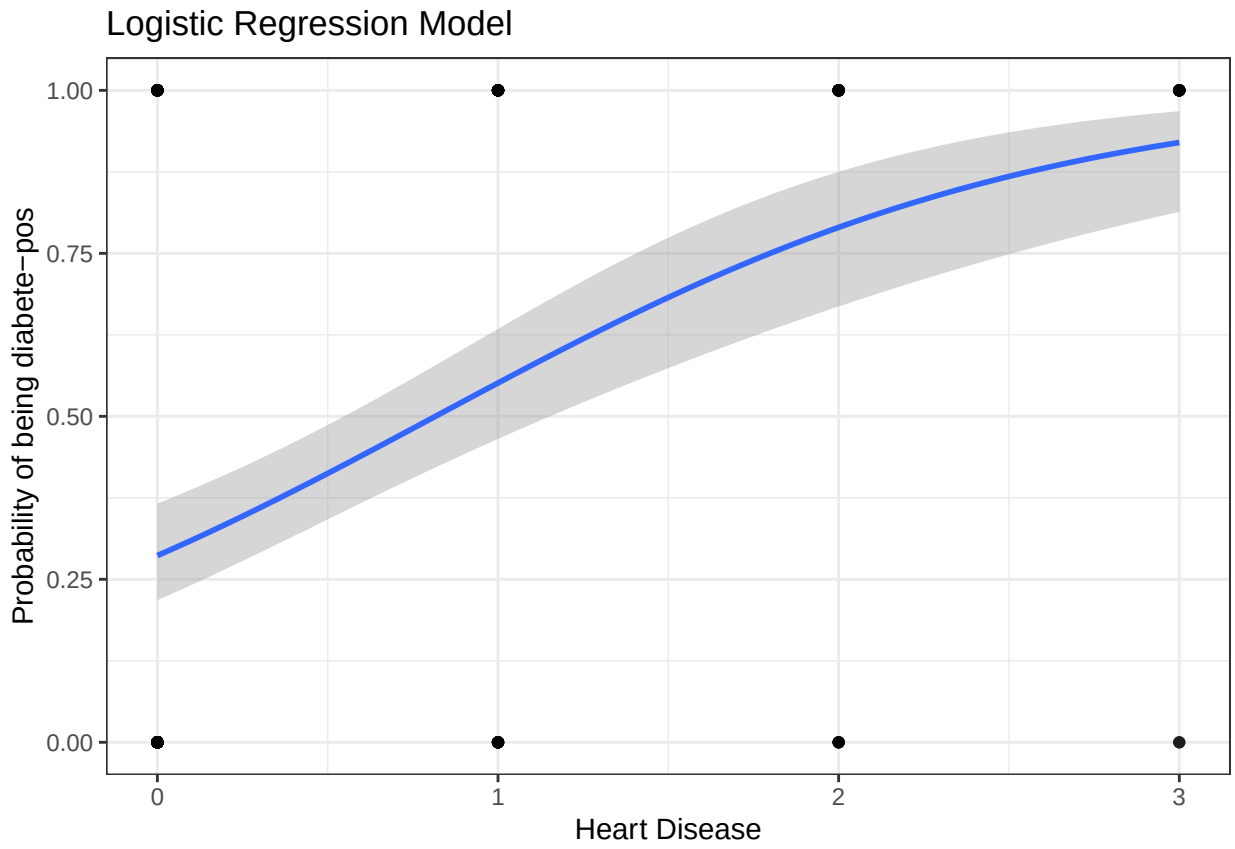
```
##                            Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)              -0.9131103  0.1859757 -4.909837 9.115203e-07
## Number.of.vessels.fluro  1.1184349  0.1919346  5.827166 5.637646e-09
```

```
newdata <- data.frame(Number.of.vessels.fluro = c(20, 180))
probabilities <- model2 %>% predict(newdata, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, 1,0)
predicted.classes
```

```
## 1 2
## 1 1
```

```
train.data %>%
  mutate(prob = Heart.Disease) %>%
  ggplot(aes(Number.of.vessels.fluro, prob)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +
  labs(
    title = "Logistic Regression Model",
    x = "Heart Disease",
    y = "Probability of being diabete-pos"
  )
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
gam.model <- gam(Heart.Disease ~ Number.of.vessels.fluro + Thallium + Sex, data = train.data, family =

summary(gam.model)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## Heart.Disease ~ Number.of.vessels.fluro + Thallium + Sex
##
## Parametric coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -3.78205    0.52655  -7.183 6.84e-13 ***
## Number.of.vessels.fluro  1.00169    0.20792   4.818 1.45e-06 ***
## Thallium                 0.52682    0.09713   5.424 5.83e-08 ***
## Sex                      0.55055    0.43037   1.279    0.201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.396   Deviance explained =   32%
## UBRE = -0.02652  Scale est. = 1         n = 216
```

```r
probabilities <- gam.model %>% predict(test.data, type = "response")
predicted.classes <- ifelse(probabilities> 0.5, "pos", "neg")

mean(predicted.classes == test.data$Heart.Disease)
```

```
## [1] 0
```