pHW2

Ryan Richardson

13/11/2020

Pre-Processing

Start with basic preprocessing and drop the following:

Age is dropped due to being duplicated in agec Bmi is dropped due to being calculated by height and weight dibpat is dropped due to being duplicated by behpat Id is droppepd due to being an identifier field not related to data sbp is droppped due to being contained in lnsbp typchd69 is dropped due to being directly related to our target chd weight and weightcat are dropped due to being contained in lnwght smoking is dropped due to being contained in ncigs Arcus is converted to a categorical variable time169 is removed for being contained in our target variable Records with NA values are dropped as there is a minimal number

```
wcgs <- read.dta('wcgs.dta')
wcgs$arcus = factor(wcgs$arcus)
filtered = wcgs[, -c(1,4,7,8,10,14,15,17,18,20,21)]
filtered <- filtered[complete.cases(filtered),]
summary(filtered)
```

```
##  arcus    behpat   chd69           chol           height          lnsbp    
##  0:2174   A1: 260   No :2850   Min.   :103.0   Min.   :60.00   Min.   
:4.585  
##  1: 927   A2:1304   Yes: 251   1st Qu.:197.0   1st Qu.:68.00   1st 
Qu.:4.787  
##            B3:1194              Median :223.0   Median :70.00   Median 
:4.836  
##            B4: 343              Mean   :226.3   Mean   :69.78   Mean   
:4.850  
##                                 3rd Qu.:253.0   3rd Qu.:72.00   3rd 
Qu.:4.913  
##                                 Max.   :645.0   Max.   :78.00   Max.   
:5.438  
##      lnwght          ncigs             t1                 uni           
##  Min.   :4.357   Min.   : 0.00   Min.   :-47.43147   Min.   :0.0007097  
##  1st Qu.:5.043   1st Qu.: 0.00   1st Qu.: -1.00335   1st Qu.:0.2566030  
##  Median :5.136   Median : 0.00   Median :  0.01191   Median :0.5140467  
##  Mean   :5.128   Mean   :11.56   Mean   : -0.03181   Mean   :0.5045844  
##  3rd Qu.:5.204   3rd Qu.:20.00   3rd Qu.:  0.97947   3rd Qu.:0.7559930  
##  Max.   :5.768   Max.   :99.00   Max.   : 47.01623   Max.   :0.9994496  
##     agec     
##  35-40: 534  
##  41-45:1073
##  46-50: 742
##  51-55: 518
##  56-60: 234
##
```

Backward Elimination Main Effects Model Building

We start with a general linear model using chd69 as our variable of interest, and all other variables as predictors Models are shown with the following variables dropped over time: Behbat t1 height uni arcus

The model we are left with shows a an almost% chance to have some form of CHD, for someone with a Cholesterol of 0, systolic bp of exp(0), a weight of exp(0), smoking 0 cigarretees a day, under age 40.

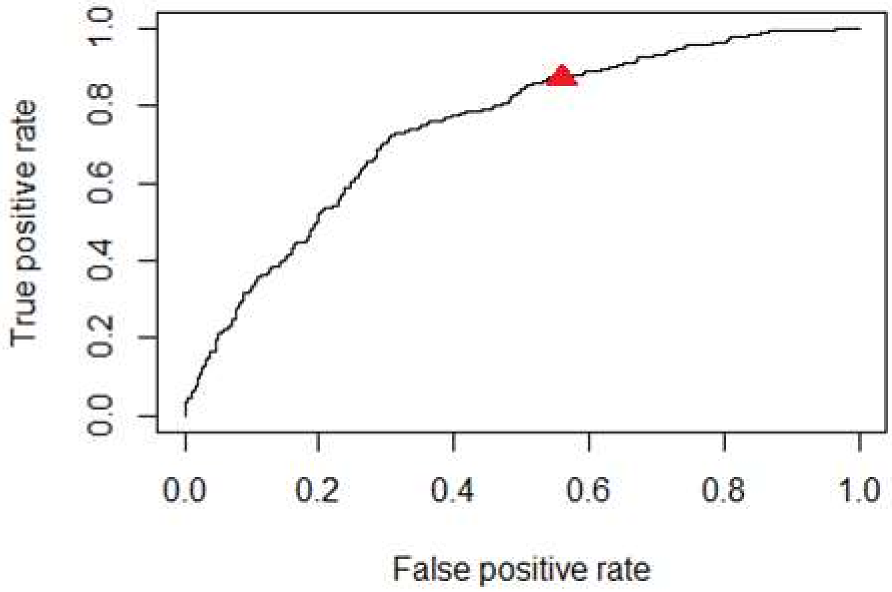
For every 1 point in cholesterol a person is 1% more likely to experience CHD, for every std.dev they are above the mean sbp they 14% more likely to experience CHD. For every std.dev they are above the mean in weight they are 5.5% more likely to experience CHD). For every cigarette they smoke per day they are 1% more likely to experience CHD. Someone in the age category 41-45 is 0.8% more likely to experience CHD, someone 51-55 is 1.9% more likely to experience CHD, and someone from 56-60 is 2.6% more likely to experience CHD than the baseline.

Once we plot the ROC Curve, we get the best AUC of 0.959 when using a threshold of 0.88

Based on the threshold values needed to attain this AUC, I would be worried about the accuracy of the model. Primarily because such a high threshold is dangerously close to simply saying that every person will have coronary heart disease. Even so, we have a high AUC value, which tells me that this model is missing something very significant, and that some effects are missing from it.

The Hosmer-Lemeshow GoF test gies an extremely small p-value, close to 0, indicating a poorly fit model

```
##
## Call:
## glm(formula = chd69 ~ ., family = "binomial", data = mbest)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2935  -0.4370  -0.3178  -0.2269   2.9258
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -27.374056   3.581997  -7.642 2.14e-14 ***
## chol          0.011156   0.001517   7.355 1.92e-13 ***
## lnsbp         2.641480   0.586591   4.503 6.70e-06 ***
## lnwght        1.715942   0.569533   3.013 0.002588 **
## ncigs         0.023486   0.004242   5.537 3.08e-08 ***
## agec41-45    -0.223400   0.241490  -0.925 0.354920
## agec46-50     0.444348   0.233242   1.905 0.056768 .
## agec51-55     0.659128   0.241218   2.732 0.006286 **
## agec56-60     0.960250   0.273035   3.517 0.000437 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1743.2  on 3100  degrees of freedom
## Residual deviance: 1562.4  on 3092  degrees of freedom
## AIC: 1580.4
##
## Number of Fisher Scoring iterations: 6
```



```
## Warning in Ops.factor(1, y): '-' not meaningful for factors

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  mbest$chd69, fitted(mainModelBest)
## X-squared = 3101, df = 98, p-value < 2.2e-16

## [1] "Main Effects Model Best AUC"

## [1] 0.9596774

## [1] "Main Effects Model Best AUC Threshold"

## [1] 0.8881105
```

Backward Elimination Main & 2-Way Interaction Effects Model Building

Starting again with our main model using chd69 as our variable of interest Lnsbp was left in as was weight due to being close to the 0.05 significance level Agec was retained as well for having significant interactions Models are shown with the following dropped over time: uni lnsbp t1 behbat height ncigs arcus

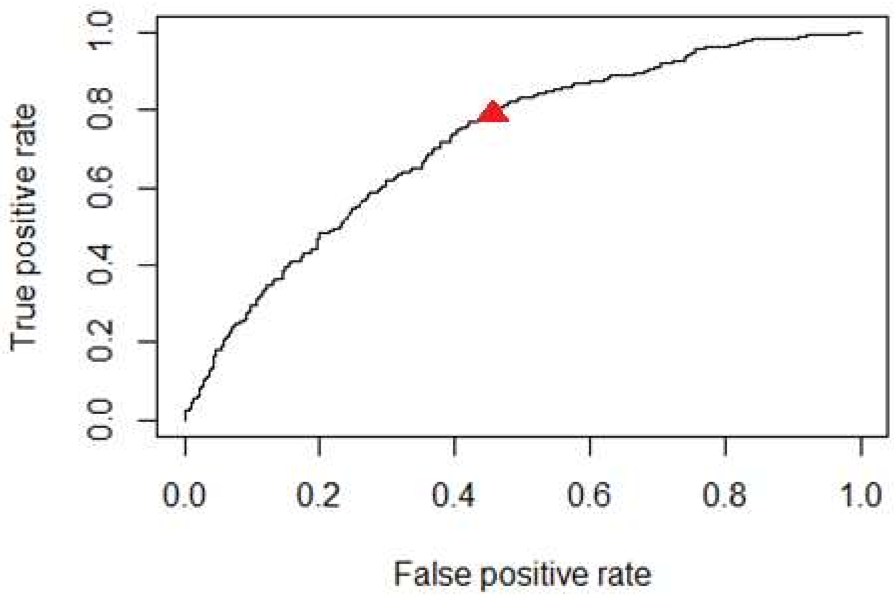
This model is looking at someone with a choelsterol of 0, of average weight, age less than 40 to begin with. There are significant interactions between weight and cholesterol seen in this model. For most of the included effects, every point increase will result in a substantial increase in CHD risk rate. There is a surprising drop in CHD chance for peopl age 56-60 off the bat, but that seems to be partly due to the interaction between lnwght and that age group causing a substantial increase in risk. Beacuse of this, it may have been a good idea to remove the age category from the model and replace it with some other age measure.

When we plot the ROC curve we get the best AUC of 0.96042 at a threshold of 0.5

I’m much happier with the threshold values giving us the AUC in this model. This indicates not only a decent prediction, but the lower threshold indicates a much lower chance that we classify every as having a CHD event or not.

Unfortunately, the Hosmer-Lemeshow goodness of fit still indicates a poorly fit model. This may be due to missing factors, high covariance in the data, or some other mispecification caused by the target variables used. Such as the agec.

```
##
## Call:
## glm(formula = chd69 ~ .^2, family = binomial(link = "logit"),
##     data = best)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1138  -0.4563  -0.3363  -0.2399   3.0561
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -52.012727  15.660815  -3.321 0.000896 ***
## chol               0.152331   0.057127   2.667 0.007664 **
## lnwght             8.848626   3.011998   2.938 0.003306 **
## agec41-45          1.705204  10.074625   0.169 0.865594
## agec46-50          7.143203   9.683172   0.738 0.460701
## agec51-55          2.948586  10.250834   0.288 0.773620
## agec56-60         -6.555316  10.966641  -0.598 0.550006
## chol:lnwght       -0.026583   0.011031  -2.410 0.015959 *
## chol:agec41-45    -0.001060   0.004634  -0.229 0.819121
## chol:agec46-50    -0.003257   0.004982  -0.654 0.513263
## chol:agec51-55    -0.003162   0.005038  -0.628 0.530262
## chol:agec56-60    -0.011342   0.005783  -1.961 0.049833 *
## lnwght:agec41-45  -0.320431   1.911079  -0.168 0.866843
## lnwght:agec46-50  -1.129358   1.833501  -0.616 0.537923
## lnwght:agec51-55  -0.274795   1.932960  -0.142 0.886951
## lnwght:agec56-60   2.007556   2.090160   0.960 0.336814
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1743.2  on 3100  degrees of freedom
## Residual deviance: 1599.1  on 3085  degrees of freedom
## AIC: 1631.1
##
## Number of Fisher Scoring iterations: 6
```



```
## Warning in Ops.factor(1, y): '-' not meaningful for factors

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  best$chd69, fitted(interactionModelBest)
## X-squared = 3101, df = 8, p-value < 2.2e-16

## [1] "Interaction Effects Model Best AUC"

## [1] 0.9598258
## [1] "Interaction Effects Model Best AUC Threshold"

## [1] 0.5041135
```