Personal Key Indicators of Heart Disease Project Code

Charles Hwang

12/18/2022

Charles Hwang

Dr. Xi

STAT 408-001

2022 December 18

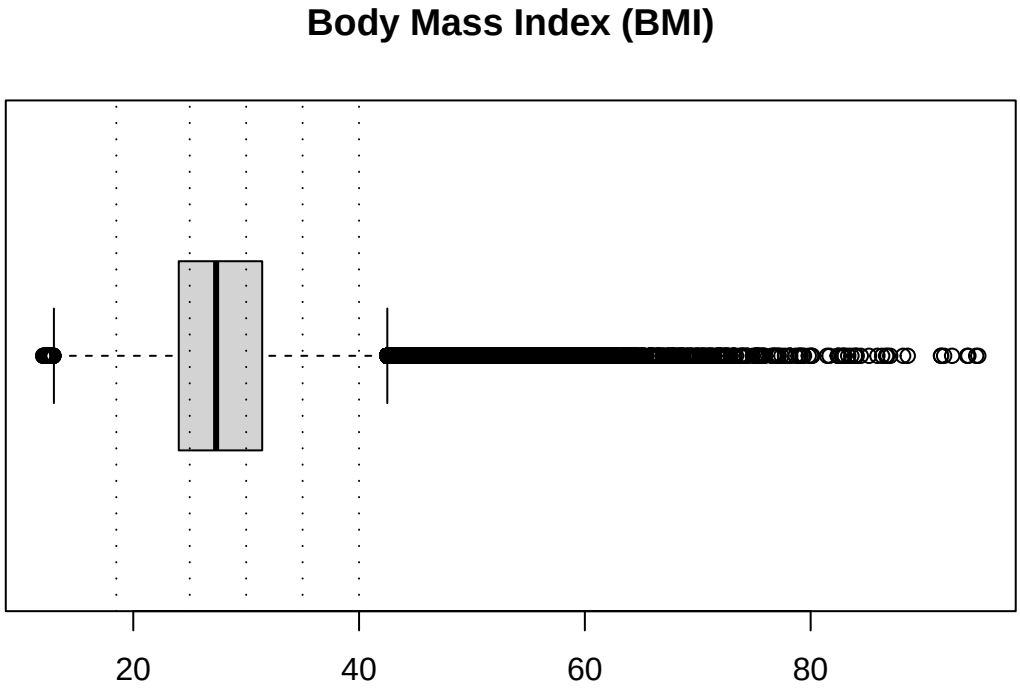
Data

```
rm(list=ls())
heart<-read.csv("~/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/heart_2020_cleaned.csv
heart$HeartDisease<-as.factor(heart$HeartDisease)
heart$Smoking<-as.factor(heart$Smoking)
heart$AlcoholDrinking<-as.factor(heart$AlcoholDrinking)
heart$Stroke<-as.factor(heart$Stroke)
heart$DiffWalking<-as.factor(heart$DiffWalking)
heart$Sex<-as.factor(heart$Sex)
heart$AgeCategory<-as.factor(heart$AgeCategory)
heart$Race<-as.factor(heart$Race)
heart$Diabetic<-factor(heart$Diabetic,labels=c("No","BL","Yes","YesPreg")) # Renaming levels
heart$PhysicalActivity<-as.factor(heart$PhysicalActivity)
heart$GenHealth<-factor(heart$GenHealth,labels=c(5,2,3,1,4)) # Reordering levels to EX, VG, G, F, P
heart$Asthma<-as.factor(heart$Asthma)
heart$KidneyDisease<-as.factor(heart$KidneyDisease)
heart$SkinCancer<-as.factor(heart$SkinCancer)
set.seed(612)
s<-sort(sample(nrow(heart),round(nrow(heart)*.8)))
train<-heart[s,]
test<-heart[-s,]
rm(s)
table(heart$HeartDisease) # p
```

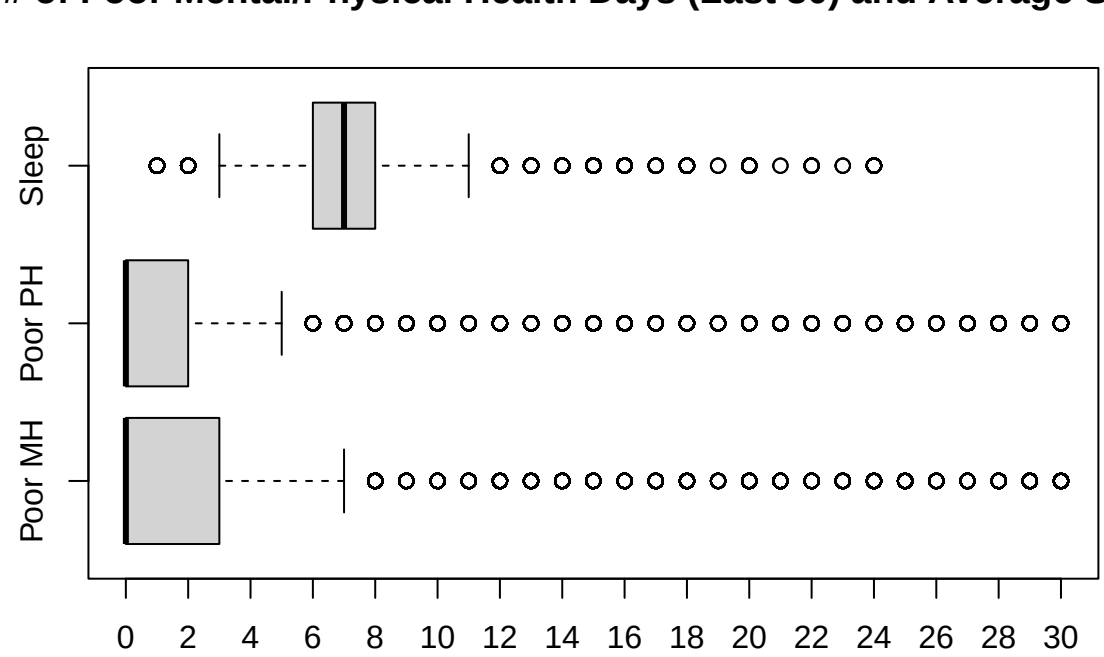
```
##
##     No    Yes
## 292422  27373
```

Data Visualization

```
boxplot(heart$BMI,main="Body Mass Index (BMI)",horizontal=TRUE)
abline(v=c(18.5,25,30,35,40),lty="17") # Underweight, normal, overweight, class I-III obesity
```




```
boxplot(heart$MentalHealth,heart$PhysicalHealth,heart$SleepTime,names=c("Poor MH","Poor PH","Sleep"),ma
```




```
cor(heart[,c("BMI","PhysicalHealth","MentalHealth","SleepTime")])
```

```
##                      BMI PhysicalHealth MentalHealth   SleepTime
## BMI           1.00000000     0.10978754   0.06413057 -0.05182225
## PhysicalHealth 0.10978754     1.00000000   0.28798667 -0.06138663
## MentalHealth   0.06413057     0.28798667   1.00000000 -0.11971679
## SleepTime     -0.05182225    -0.06138663  -0.11971679  1.00000000
```

Model

```
summary(glm(HeartDisease~.,family=binomial,data=train))
```

```
##
## Call:
## glm(formula = HeartDisease ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1133  -0.4104  -0.2428  -0.1270   3.6221
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -6.3097610  0.1299880 -48.541  < 2e-16 ***
## BMI                       0.0090913  0.0012816   7.094 1.30e-12 ***
## SmokingYes                0.3602878  0.0161066  22.369  < 2e-16 ***
## AlcoholDrinkingYes       -0.2507708  0.0378075  -6.633 3.29e-11 ***
## StrokeYes                 1.0453483  0.0252846  41.343  < 2e-16 ***
## PhysicalHealth            0.0034601  0.0009667   3.579 0.000345 ***
## MentalHealth              0.0046946  0.0009894   4.745 2.08e-06 ***
## DiffWalkingYes            0.1958066  0.0203425   9.626  < 2e-16 ***
## SexMale                   0.7183172  0.0163257  43.999  < 2e-16 ***
## AgeCategory25-29          0.1345623  0.1419976   0.948 0.343314
## AgeCategory30-34          0.4766847  0.1272601   3.746 0.000180 ***
## AgeCategory35-39          0.6005253  0.1217014   4.934 8.04e-07 ***
## AgeCategory40-44          1.0230442  0.1142758   8.952  < 2e-16 ***
## AgeCategory45-49          1.3435366  0.1103239  12.178  < 2e-16 ***
## AgeCategory50-54          1.7801817  0.1064447  16.724  < 2e-16 ***
## AgeCategory55-59          2.0062542  0.1048671  19.131  < 2e-16 ***
## AgeCategory60-64          2.2845783  0.1039234  21.983  < 2e-16 ***
## AgeCategory65-69          2.5199432  0.1036303  24.317  < 2e-16 ***
## AgeCategory70-74          2.8095883  0.1035570  27.131  < 2e-16 ***
## AgeCategory75-79          3.0046664  0.1041518  28.849  < 2e-16 ***
## AgeCategory80 or older    3.2566497  0.1038827  31.349  < 2e-16 ***
## RaceAsian                -0.5125952  0.0925055  -5.541 3.00e-08 ***
## RaceBlack                -0.3968159  0.0643837  -6.163 7.12e-10 ***
## RaceHispanic             -0.2985811  0.0655792  -4.553 5.29e-06 ***
## RaceOther                -0.1165961  0.0713955  -1.633 0.102448
## RaceWhite                -0.1187634  0.0573349  -2.071 0.038321 *
## DiabeticBL                0.1431830  0.0468454   3.057 0.002239 **
## DiabeticYes               0.4823947  0.0187059  25.788  < 2e-16 ***
## DiabeticYesPreg           0.0544857  0.1198488   0.455 0.649382
## PhysicalActivityYes       0.0298776  0.0180095   1.659 0.097117 .
## GenHealth2                1.5092180  0.0366381  41.193  < 2e-16 ***
## GenHealth3                1.0323838  0.0329548  31.327  < 2e-16 ***
## GenHealth1                1.8878382  0.0457246  41.287  < 2e-16 ***
## GenHealth4                0.4475175  0.0338513  13.220  < 2e-16 ***
## SleepTime                -0.0263790  0.0048746  -5.412 6.25e-08 ***
## AsthmaYes                 0.2709828  0.0215686  12.564  < 2e-16 ***
## KidneyDiseaseYes          0.5520620  0.0274444  20.116  < 2e-16 ***
## SkinCancerYes             0.1037718  0.0218499   4.749 2.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 149034  on 255835  degrees of freedom
## Residual deviance: 115592  on 255798  degrees of freedom
## AIC: 115668
##
## Number of Fisher Scoring iterations: 7
```

```
step(glm(HeartDisease~.,family=binomial,data=train),direction="both") # Same as full model
```

```
## Start:  AIC=115667.7
## HeartDisease ~ BMI + Smoking + AlcoholDrinking + Stroke + PhysicalHealth +
##     MentalHealth + DiffWalking + Sex + AgeCategory + Race + Diabetic +
##     PhysicalActivity + GenHealth + SleepTime + Asthma + KidneyDisease +
##     SkinCancer
##
##                     Df Deviance    AIC
## <none>                   115592 115668
## - PhysicalActivity   1   115594 115668
## - PhysicalHealth     1   115604 115678
## - MentalHealth       1   115614 115688
## - SkinCancer         1   115614 115688
## - SleepTime          1   115621 115695
## - AlcoholDrinking    1   115638 115712
## - BMI                1   115641 115715
## - DiffWalking        1   115683 115757
## - Race               5   115721 115787
## - Asthma             1   115745 115819
## - KidneyDisease      1   115977 116051
## - Smoking            1   116094 116168
## - Diabetic           3   116238 116308
## - Stroke             1   117183 117257
## - Sex                1   117579 117653
## - GenHealth          4   118494 118562
## - AgeCategory       12   123337 123389
```

```
##
## Call:  glm(formula = HeartDisease ~ BMI + Smoking + AlcoholDrinking +
##     Stroke + PhysicalHealth + MentalHealth + DiffWalking + Sex +
##     AgeCategory + Race + Diabetic + PhysicalActivity + GenHealth +
##     SleepTime + Asthma + KidneyDisease + SkinCancer, family = binomial,
##     data = train)
##
## Coefficients:
##             (Intercept)                      BMI              SmokingYes
##               -6.309761                 0.009091                 0.360288
##     AlcoholDrinkingYes               StrokeYes          PhysicalHealth
##               -0.250771                 1.045348                 0.003460
##           MentalHealth          DiffWalkingYes                  SexMale
##                0.004695                 0.195807                 0.718317
##       AgeCategory25-29        AgeCategory30-34        AgeCategory35-39
##                0.134562                 0.476685                 0.600525
##       AgeCategory40-44        AgeCategory45-49        AgeCategory50-54
##                1.023044                 1.343537                 1.780182
```

```
##       AgeCategory55-59       AgeCategory60-64       AgeCategory65-69
##               2.006254               2.284578               2.519943
##       AgeCategory70-74       AgeCategory75-79  AgeCategory80 or older
##               2.809588               3.004666               3.256650
##              RaceAsian              RaceBlack              RaceHispanic
##              -0.512595              -0.396816              -0.298581
##              RaceOther              RaceWhite              DiabeticBL
##              -0.116596              -0.118763               0.143183
##            DiabeticYes        DiabeticYesPreg    PhysicalActivityYes
##               0.482395               0.054486               0.029878
##             GenHealth2             GenHealth3             GenHealth1
##               1.509218               1.032384               1.887838
##             GenHealth4              SleepTime              AsthmaYes
##               0.447518              -0.026379               0.270983
##       KidneyDiseaseYes          SkinCancerYes
##               0.552062               0.103772
##
## Degrees of Freedom: 255835 Total (i.e. Null);  255798 Residual
## Null Deviance:       149000
## Residual Deviance: 115600    AIC: 115700
```

```
table(train[train$AgeCategory=="25-29","HeartDisease"]) # Non-significant levels/variables
```

```
##
##    No   Yes
## 13429   102
```

```
table(train[train$Diabetic=="YesPreg","HeartDisease"])
```

```
##
##   No  Yes
## 1990   83
```

```
table(train[train$Race=="Other","HeartDisease"])
```

```
##
##   No  Yes
## 8031  700
```

```
table(train$PhysicalActivity,train$HeartDisease)
```

```
##
##           No    Yes
##   No   49730   7834
##   Yes 184311  13961
```

```
summary(glm(HeartDisease~.,family=binomial,data=train))$aic # Akaike information criterion (AIC)
```

```
## [1] 115667.7
```

```
pred<-predict(glm(HeartDisease~.,family=binomial,data=train),test,type="response")
table(round(pred))                # Predictions
```

```
##
##     0     1
## 62914  1045
```

```
table(test$HeartDisease)          # Actual values
```

```
##
##    No   Yes
## 58381  5578
```

```
glmt<-table(test$HeartDisease,round(pred))
glmt                              # Confusion matrix
```

```
##
##           0     1
##   No  57900   481
##   Yes  5014   564
```

```
sum(diag(glmt))/nrow(test)        # Accuracy rate
```

```
## [1] 0.9140856
```

```
glmt["Yes","1"]/sum(glmt["Yes",]) # True positive rate (sensitivity)
```

```
## [1] 0.1011115
```

```
glmt["No","0"]/sum(glmt["No",])   # True negative rate (specificity)
```

```
## [1] 0.991761
```

```
glmt["No","1"]/sum(glmt["No",])   # False positive rate
```

```
## [1] 0.008238982
```

```
glmt["Yes","0"]/sum(glmt["Yes",]) # False negative rate
```

```
## [1] 0.8988885
```

```
glmt["Yes","1"]/sum(glmt[,"1"])   # Positive predictive value (precision)
```

```
## [1] 0.5397129
```

```
glmt["No","1"]/sum(glmt[,"1"])    # False discovery rate
```

```
## [1] 0.4602871
```

```
#summary(glm(HeartDisease~.-PhysicalActivity,family=binomial,data=train)) # Manual backward selection
#summary(glm(HeartDisease~.-PhysicalActivity-Diabetic,family=binomial,data=train))
#summary(glm(HeartDisease~.-PhysicalActivity-Diabetic-Race,family=binomial,data=train))
#summary(glm(HeartDisease~.-PhysicalActivity-Diabetic-Race-AgeCategory,family=binomial,data=train))
#summary(glm(HeartDisease~.-PhysicalActivity-Diabetic-Race-AgeCategory-BMI,family=binomial,data=train))
summary(glm(HeartDisease~.-PhysicalActivity-Diabetic-Race-AgeCategory-BMI,family=binomial,data=train))$a
```

```
## [1] 125422.9
```

```
table(test$HeartDisease,round(predict(glm(HeartDisease~.-PhysicalActivity-Diabetic-Race-AgeCategory-BMI
```

```
##
##           0     1
##   No  58008   373
##   Yes  5146   432
```