stapholz.jack_groupproject

Jack Stapholz

2023-11-29

```r
TrainSAData <- read.csv(file = "../Data/TrainSAData2.csv")
TestSADataNoY <- read.csv(file = "../Data/TestSAData2NoY.csv")
kaggleSampleSolution <- read.csv(file = "../Data/SampleSolution.csv")

TrainSAData$sex <- as.factor(TrainSAData$sex)
TrainSAData$hear_left <- as.factor(TrainSAData$hear_left)
TrainSAData$hear_right <- as.factor(TrainSAData$hear_right)
TrainSAData$BMI.Category <- as.factor(TrainSAData$BMI.Category)
TrainSAData$AGE.Category <- as.factor(TrainSAData$AGE.Category)
TrainSAData$Smoking.Status <- as.factor(TrainSAData$Smoking.Status)

TestSADataNoY$sex <- as.factor(TestSADataNoY$sex)
TestSADataNoY$hear_left <- as.factor(TestSADataNoY$hear_left)
TestSADataNoY$hear_right <- as.factor(TestSADataNoY$hear_right)
TestSADataNoY$BMI.Category <- as.factor(TestSADataNoY$BMI.Category)
TestSADataNoY$AGE.Category <- as.factor(TestSADataNoY$AGE.Category)
TestSADataNoY$Smoking.Status <- as.factor(TestSADataNoY$Smoking.Status)
```

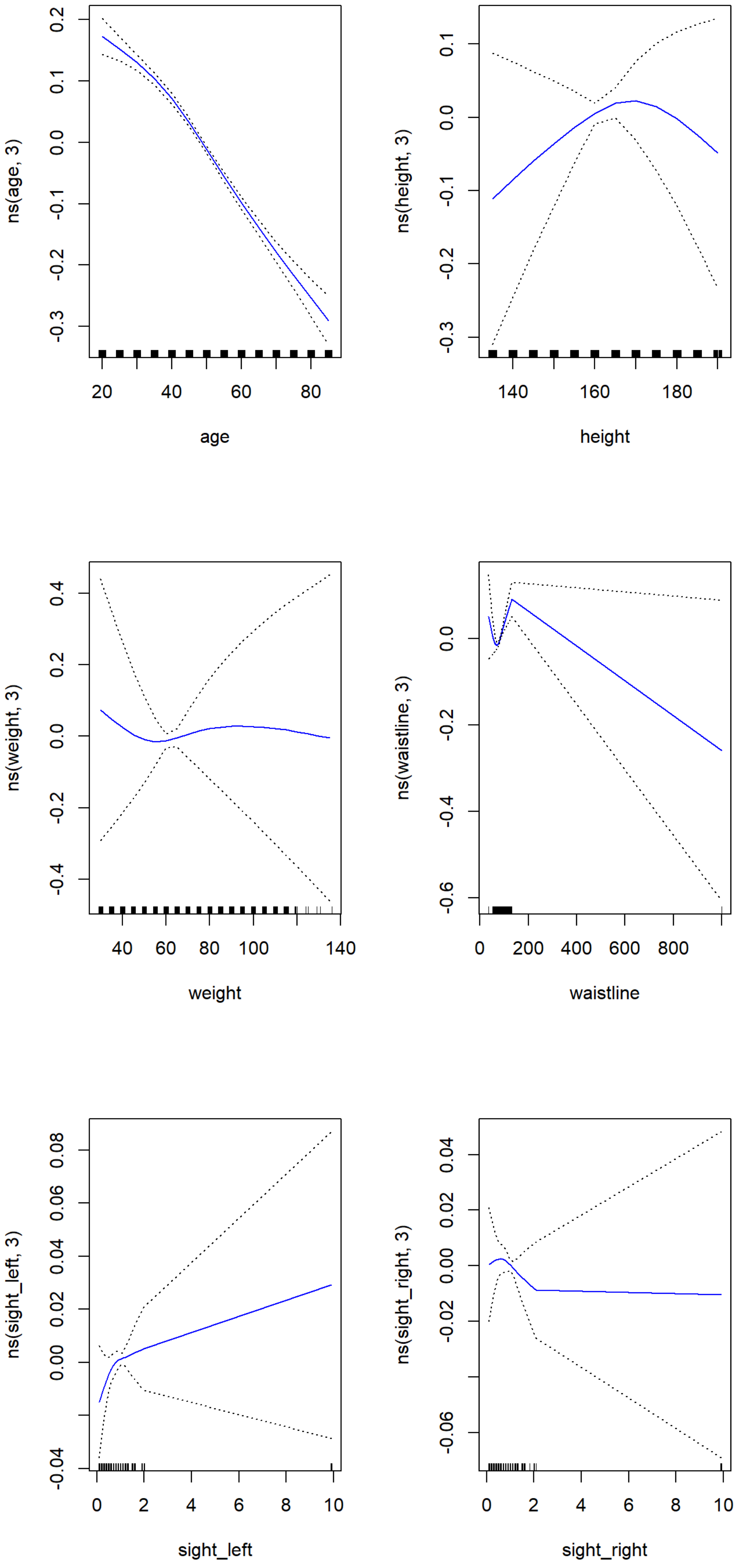
```r
#Removing NAs from Training Data
TrainSAData[sapply(TrainSAData, is.numeric)] <- lapply(TrainSAData[sapply(TrainSAData, is.numeric)], function(x) ifelse(is.na(x), median(x, na.rm = TRUE), x))
TrainSAData[sapply(TrainSAData, is.factor)] <- lapply(TrainSAData[sapply(TrainSAData, is.factor)], function(x) ifelse(is.na(x), Mode(x, na.rm = TRUE), x))

#Removing NAs from Testing Data
TestSADataNoY[sapply(TestSADataNoY, is.numeric)] <- lapply(TestSADataNoY[sapply(TestSADataNoY, is.numeric)], function(x) ifelse(is.na(x), median(x, na.rm = TRUE), x))
TestSADataNoY[sapply(TestSADataNoY, is.factor)] <- lapply(TestSADataNoY[sapply(TestSADataNoY, is.factor)], function(x) ifelse(is.na(x), Mode(x, na.rm = TRUE), x))
```

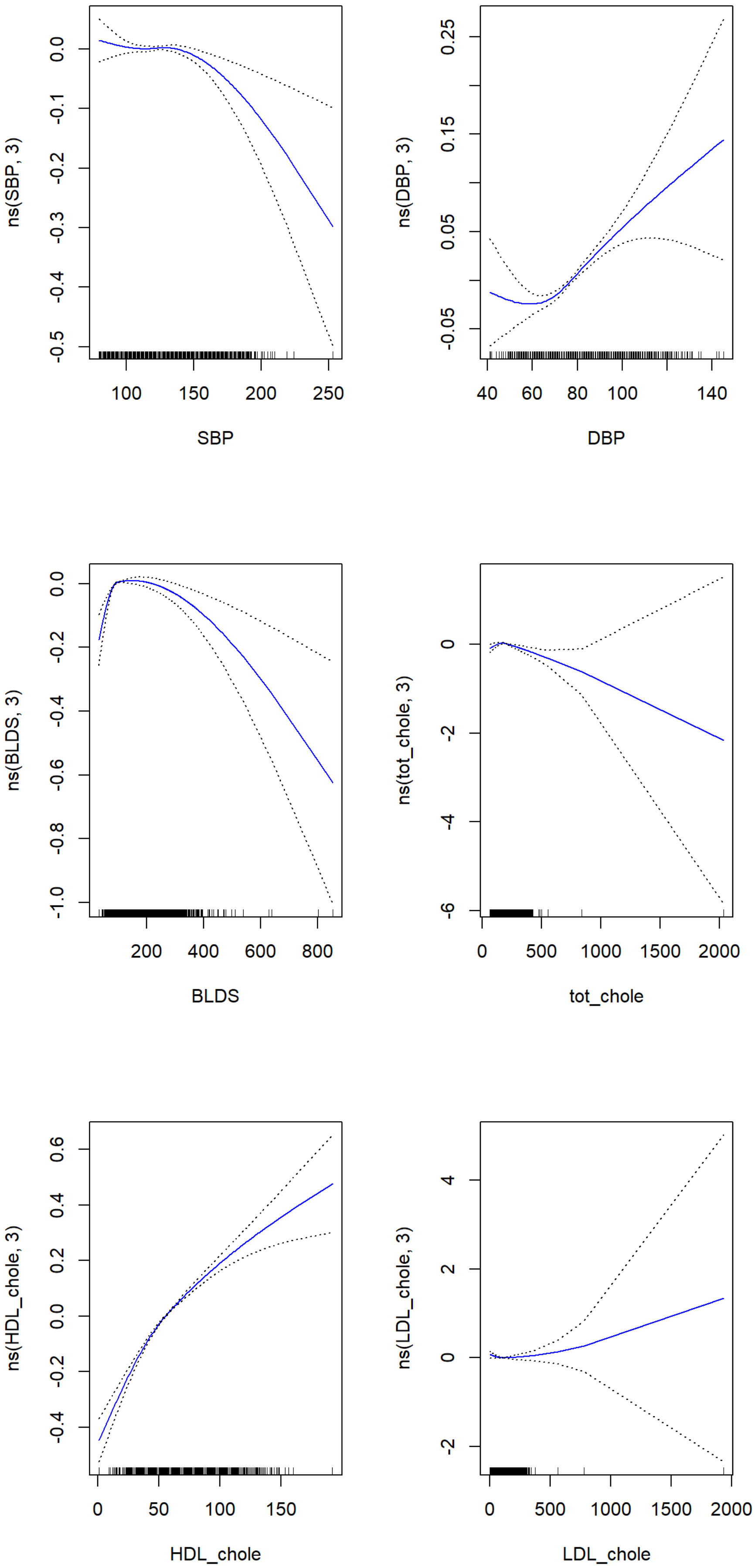
```r
hmiscTrain <- read.csv("../Data/HmiscTrain.csv")
hmiscTest <- read.csv("../Data/HmiscTest.csv")
```

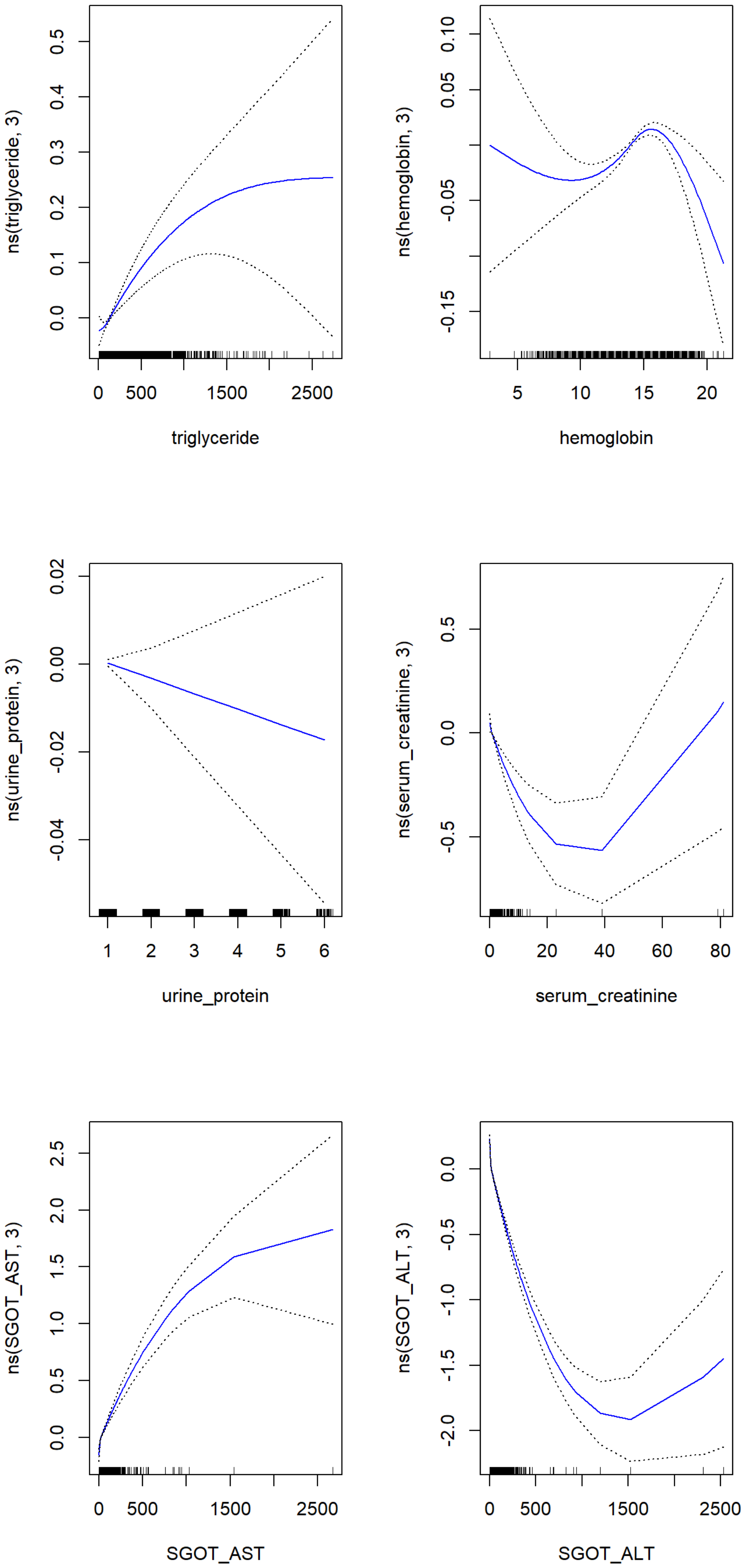
```r
hmisclm <- lm(Alcoholic.Status == "Y" ~ ., data = hmiscTrain)
summary(hmisclm)
```

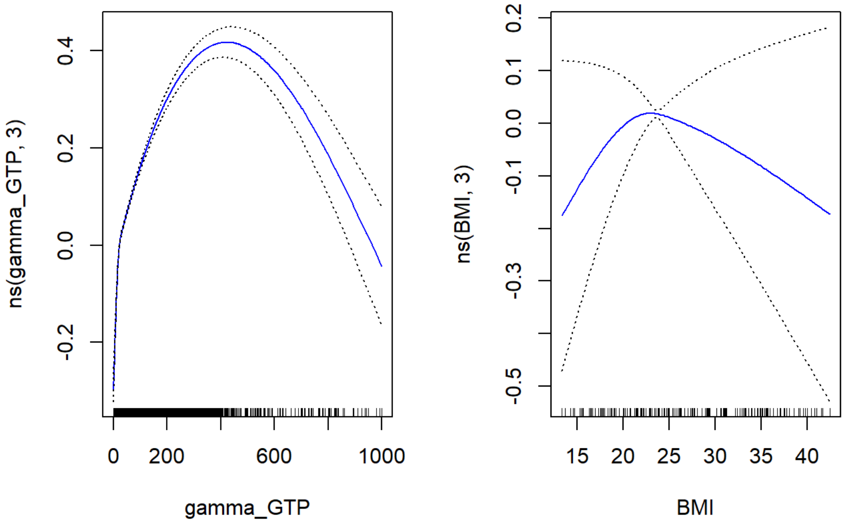
```
##
## Call:
## lm(formula = Alcoholic.Status == "Y" ~ ., data = hmiscTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4637 -0.3492 -0.0060  0.3537  2.1037
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -3.582e-01  1.993e-01  -1.798 0.072255 .
## sexMale                     2.053e-01  6.442e-03  31.874  < 2e-16 ***
## age                        -7.357e-03  3.240e-04 -22.708  < 2e-16 ***
## height                      2.457e-03  1.213e-03   2.027 0.042712 *
## weight                     -2.205e-04  1.497e-03  -0.147 0.882895
## waistline                   1.735e-04  1.675e-04   1.036 0.300275
## sight_left                  4.563e-03  2.832e-03   1.611 0.107093
## sight_right                -1.513e-03  2.868e-03  -0.527 0.597892
## hear_leftNormal            -9.694e-04  1.111e-02  -0.087 0.930483
## hear_rightNormal           -5.882e-03  1.128e-02  -0.522 0.601918
## SBP                        -2.069e-04  1.771e-04  -1.168 0.242909
## DBP                         2.411e-03  2.547e-04   9.469  < 2e-16 ***
## BLDS                        9.351e-05  6.995e-05   1.337 0.181317
## tot_chole                   9.092e-05  2.124e-04   0.428 0.668624
## HDL_chole                   4.867e-03  2.352e-04  20.695  < 2e-16 ***
## LDL_chole                  -5.674e-04  2.138e-04  -2.654 0.007949 **
## triglyceride                1.914e-04  3.612e-05   5.299 1.17e-07 ***
## hemoglobin                  7.236e-03  1.464e-03   4.942 7.74e-07 ***
## urine_protein              -6.451e-03  3.786e-03  -1.704 0.088386 .
## serum_creatinine           -9.503e-03  3.093e-03  -3.073 0.002123 **
## SGOT_AST                    9.456e-04  1.121e-04   8.434  < 2e-16 ***
## SGOT_ALT                   -1.524e-03  9.789e-05 -15.568  < 2e-16 ***
## gamma_GTP                   1.049e-03  3.860e-05  27.172  < 2e-16 ***
## BMI                         5.429e-03  4.041e-03   1.343 0.179119
## BMI.CategoryObese          -6.702e-02  1.291e-02  -5.189 2.12e-07 ***
## BMI.CategoryOverweight     -8.709e-03  6.000e-03  -1.452 0.146610
## BMI.CategoryUnderweight    -2.347e-02  9.238e-03  -2.541 0.011066 *
## AGE.CategoryOld            -4.303e-02  6.431e-03  -6.691 2.24e-11 ***
## AGE.CategoryVery Old       -3.553e-02  1.280e-02  -2.776 0.005505 **
## AGE.CategoryYoung          -2.454e-02  7.020e-03  -3.496 0.000473 ***
## Smoking.StatusStill Smoking  1.679e-01  5.055e-03  33.216  < 2e-16 ***
## Smoking.StatusUsed to Smoke  1.637e-01  5.258e-03  31.131  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4309 on 69968 degrees of freedom
## Multiple R-squared:  0.2576, Adjusted R-squared:  0.2573
## F-statistic: 783.1 on 31 and 69968 DF,  p-value: < 2.2e-16
```

```
hmiscgam3deg <- gam(Alcoholic.Status == "Y" ~ sex + ns(age, 3) + ns(height, 3) + ns(weight, 3) + ns(waistline, 3) + ns(sight
_left, 3) + ns(sight_right, 3) + hear_left + hear_right + ns(SBP, 3) + ns(DBP, 3) + ns(BLDS, 3) + ns(tot_chole, 3) + ns(HDL_
chole, 3) + ns(LDL_chole, 3) + ns(triglyceride, 3) + ns(hemoglobin, 3) + ns(urine_protein, 3) + ns(serum_creatinine, 3) + ns
(SGOT_AST, 3) + ns(SGOT_ALT, 3) + ns(gamma_GTP, 3) + ns(BMI, 3) + BMI.Category + AGE.Category + Smoking.Status, data = hmisc
Train)
par(mfrow = c(1, 2))
plot(hmiscgam3deg, se = TRUE, col = "blue")
```


ns(age, 3)
0.2
0.1
0.0
-0.1
-0.2
-0.3
20
40
60
80
age
ns(height, 3)
0.1
0.0
-0.1
-0.2
-0.3
140
160
180
height
ns(weight, 3)
0.4
0.2
0.0
-0.2
-0.4
40
60
80
100
140
weight
ns(waistline, 3)
0.0
-0.2
-0.4
-0.6
0
200
400
600
800
waistline
ns(sight_left, 3)
0.08
0.06
0.04
0.02
0.00
-0.04
0
2
4
6
8
10
sight_left
ns(sight_right, 3)
0.04
0.02
0.00
-0.02
-0.06
0
2
4
6
8
10
sight_right


ns(SBP, 3)
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
100
150
200
250
SBP
ns(DBP, 3)
0.25
0.15
0.05
-0.05
40
60
80
100
140
DBP
ns(BLDS, 3)
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
200
400
600
800
BLDS
ns(tot_chole, 3)
0
-2
-4
-6
0
500
1000
1500
2000
tot_chole
ns(HDL_chole, 3)
0.6
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
HDL_chole
ns(LDL_chole, 3)
4
2
0
-2
0
500
1000
1500
2000
LDL_chole


ns(triglyceride, 3)
0.0
0.1
0.2
0.3
0.4
0.5
0
500
1500
2500
triglyceride
ns(hemoglobin, 3)
-0.15
-0.05
0.05
0.10
5
10
15
20
hemoglobin
ns(urine_protein, 3)
-0.04
-0.02
0.00
0.02
1
2
3
4
5
6
urine_protein
ns(serum_creatinine, 3)
-0.5
0.0
0.5
0
20
40
60
80
serum_creatinine
ns(SGOT_AST, 3)
0.0
0.5
1.0
1.5
2.0
2.5
0
500
1500
2500
SGOT_AST
ns(SGOT_ALT, 3)
-2.0
-1.5
-1.0
-0.5
0.0
0
500
1500
2500
SGOT_ALT



```
summary(hmiscgam3deg)
```

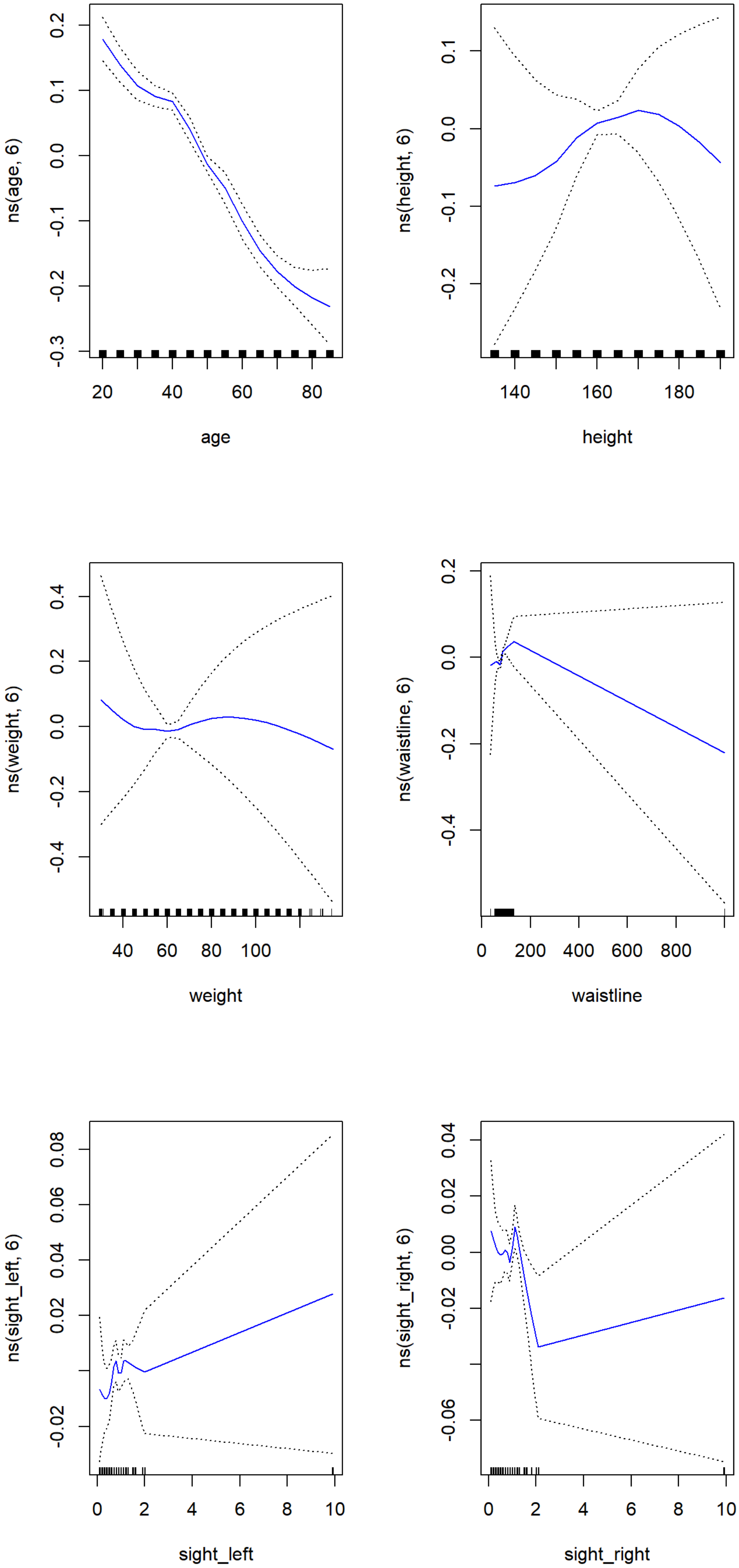
```
##
## Call: gam(formula = Alcoholic.Status == "Y" ~ sex + ns(age, 3) + ns(height,
##     3) + ns(weight, 3) + ns(waistline, 3) + ns(sight_left, 3) +
##     ns(sight_right, 3) + hear_left + hear_right + ns(SBP, 3) +
##     ns(DBP, 3) + ns(BLDS, 3) + ns(tot_chole, 3) + ns(HDL_chole,
##     3) + ns(LDL_chole, 3) + ns(triglyceride, 3) + ns(hemoglobin,
##     3) + ns(urine_protein, 3) + ns(serum_creatinine, 3) + ns(SGOT_AST,
##     3) + ns(SGOT_ALT, 3) + ns(gamma_GTP, 3) + ns(BMI, 3) + BMI.Category +
##     AGE.Category + Smoking.Status, data = hmiscTrain)
## Deviance Residuals:
##       Min        1Q    Median        3Q       Max
## -1.185305 -0.346486  0.001058  0.347492  1.291003
##
## (Dispersion Parameter for gaussian family taken to be 0.1816)
##
##     Null Deviance: 17499.82 on 69999 degrees of freedom
## Residual Deviance: 12702.04 on 69930 degrees of freedom
## AIC: 79322.08
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                            Df  Sum Sq Mean Sq    F value    Pr(>F)
## sex                         1  2371.9 2371.95 13058.5514 < 2.2e-16 ***
## ns(age, 3)                  3  1162.9  387.62  2134.0169 < 2.2e-16 ***
## ns(height, 3)               3    23.7    7.89    43.4549 < 2.2e-16 ***
## ns(weight, 3)               3     3.5    1.18     6.4975 0.0002165 ***
## ns(waistline, 3)            3     8.7    2.90    15.9871 2.188e-10 ***
## ns(sight_left, 3)           3     0.7    0.23     1.2788 0.2796789
## ns(sight_right, 3)          3     0.8    0.28     1.5431 0.2010649
## hear_left                   1     0.1    0.12     0.6842 0.4081520
## hear_right                  1     0.1    0.08     0.4190 0.5174329
## ns(SBP, 3)                  3    34.5   11.49    63.2471 < 2.2e-16 ***
## ns(DBP, 3)                  3    25.0    8.33    45.8339 < 2.2e-16 ***
## ns(BLDS, 3)                 3    16.5    5.50    30.2528 < 2.2e-16 ***
## ns(tot_chole, 3)            3    18.5    6.15    33.8775 < 2.2e-16 ***
## ns(HDL_chole, 3)            3   305.0  101.68   559.7685 < 2.2e-16 ***
## ns(LDL_chole, 3)            3   109.0   36.33   200.0295 < 2.2e-16 ***
## ns(triglyceride, 3)         3    23.1    7.69    42.3596 < 2.2e-16 ***
## ns(hemoglobin, 3)           3     7.7    2.56    14.0894 3.541e-09 ***
## ns(urine_protein, 3)        1     0.2    0.19     1.0400 0.3078309
## ns(serum_creatinine, 3)     3     7.7    2.55    14.0500 3.751e-09 ***
## ns(SGOT_AST, 3)             3     4.6    1.53     8.4377 1.331e-05 ***
## ns(SGOT_ALT, 3)             3    60.2   20.07   110.4910 < 2.2e-16 ***
## ns(gamma_GTP, 3)            3   374.8  124.95   687.9003 < 2.2e-16 ***
## ns(BMI, 3)                  3    16.5    5.51    30.3579 < 2.2e-16 ***
## BMI.Category                3     0.2    0.07     0.3707 0.7741870
## AGE.Category                3     3.5    1.18     6.4689 0.0002255 ***
## Smoking.Status              2   218.4  109.18   601.0795 < 2.2e-16 ***
## Residuals               69930 12702.0    0.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

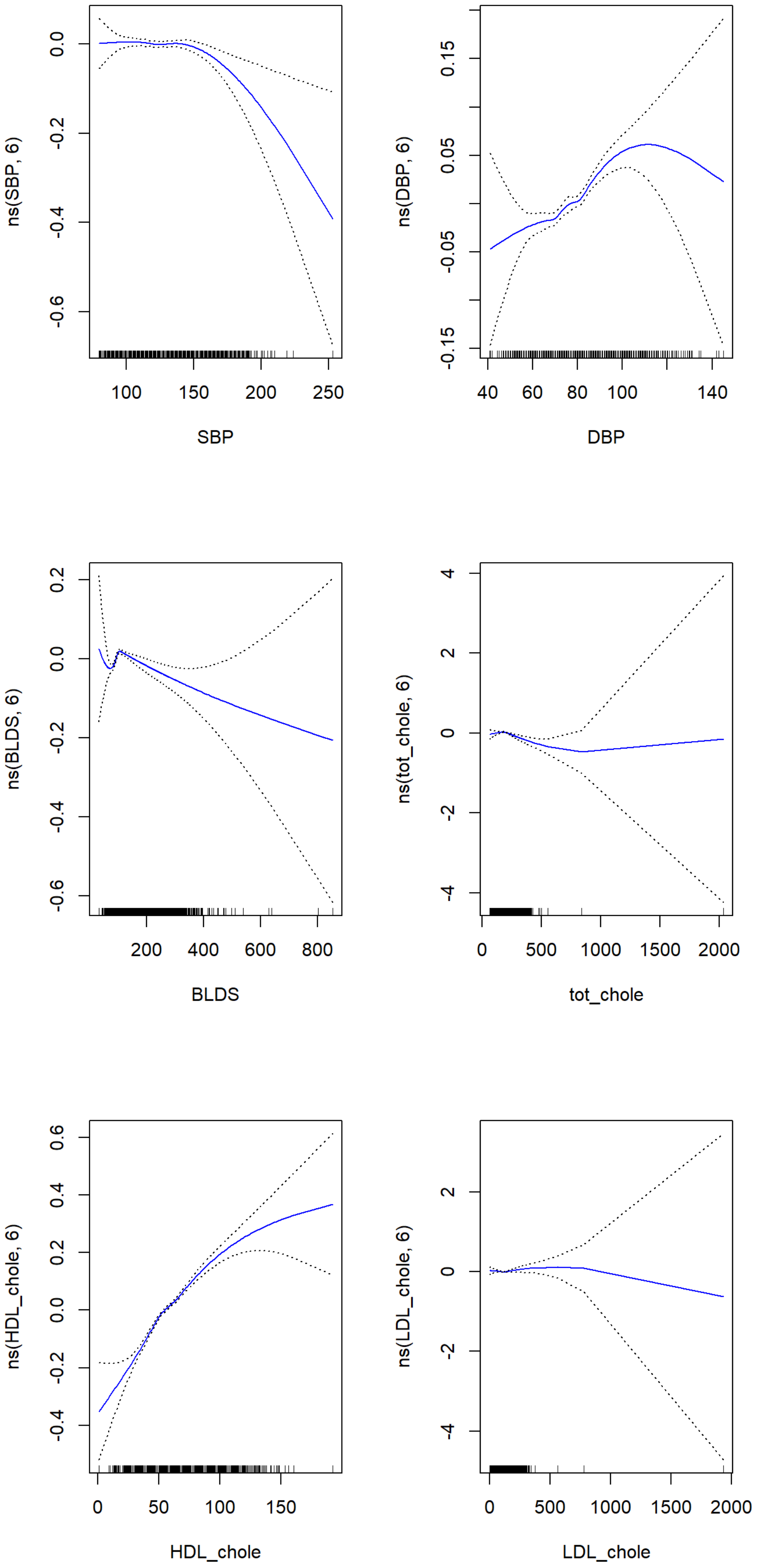
```
hmiscgam3deg.trainPredict <- predict(hmiscgam3deg, hmiscTrain)
hmiscgam3deg.trainPredict[hmiscgam3deg.trainPredict > 0.5] <- 'Y'
hmiscgam3deg.trainPredict[hmiscgam3deg.trainPredict != 'Y'] <- 'N'
sum(hmiscgam3deg.trainPredict == hmiscTrain$Alcoholic.Status) / length(hmiscgam3deg.trainPredict)
```

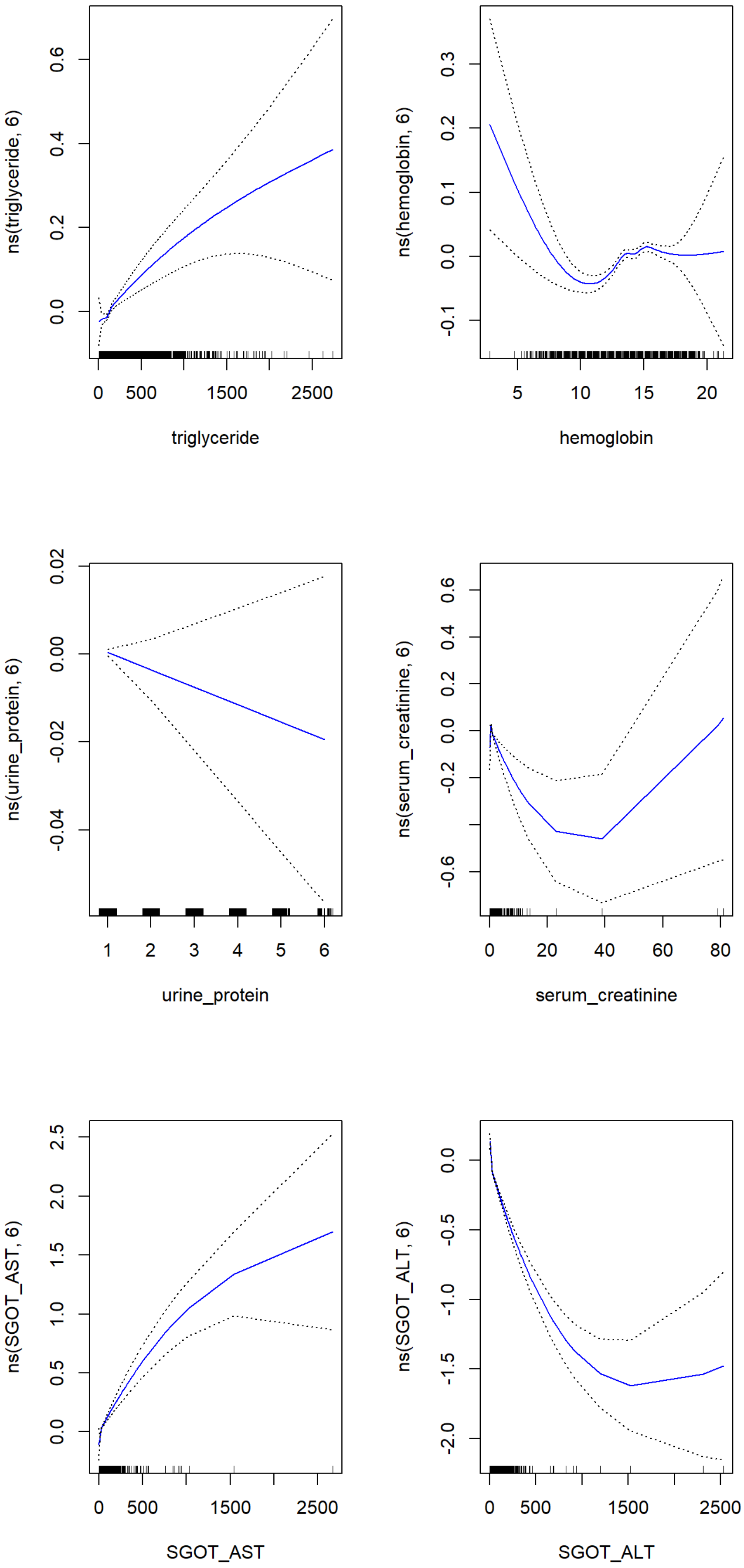
```
## [1] 0.7274
```

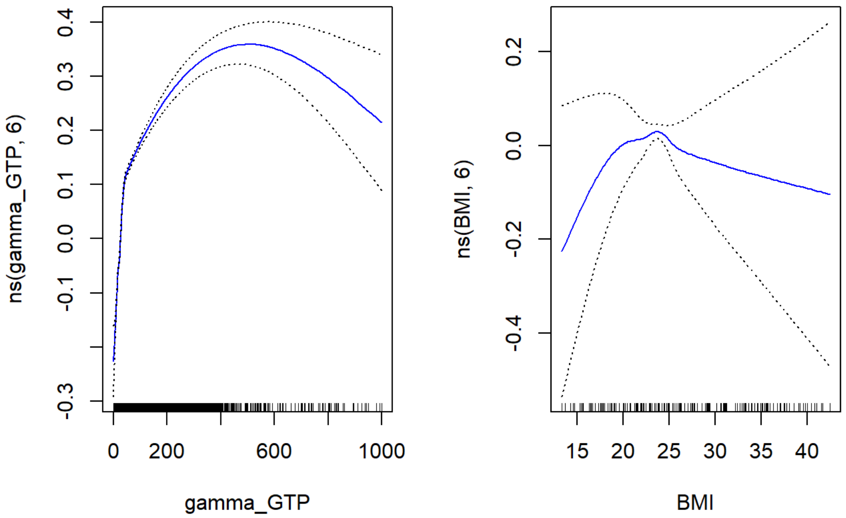
```
hmiscgam3deg.predict <- predict(hmiscgam3deg, hmiscTest)

hmiscgam3deg.predict[hmiscgam3deg.predict > 0.5] <- 'Y'
hmiscgam3deg.predict[hmiscgam3deg.predict != 'Y'] <- 'N'

hmiscgam3deg.predict <- as.data.frame(cbind("ID" = 1:30000, "Alcoholic.Status" = hmiscgam3deg.predict))
write.csv(hmiscgam3deg.predict, file = "stapholz.jack_kaggle2.csv", row.names = FALSE)
```

```
hmiscgamfull <- gam(Alcoholic.Status == "Y" ~ sex + ns(age, 6) + ns(height, 6) + ns(weight, 6) + ns(waistline, 6) + ns(sight
_left, 6) + ns(sight_right, 6) + hear_left + hear_right + ns(SBP, 6) + ns(DBP, 6) + ns(BLDS, 6) + ns(tot_chole, 6) + ns(HDL_
chole, 6) + ns(LDL_chole, 6) + ns(triglyceride, 6) + ns(hemoglobin, 6) + ns(urine_protein, 6) + ns(serum_creatinine, 6) + ns
(SGOT_AST, 6) + ns(SGOT_ALT, 6) + ns(gamma_GTP, 6) + ns(BMI, 6) + BMI.Category + AGE.Category + Smoking.Status, data = hmisc
Train)
par(mfrow = c(1, 2))
plot(hmiscgamfull, se = TRUE, col = "blue")
```


ns(age, 6)
0.2
0.1
0.0
-0.1
-0.2
-0.3
20
40
60
80
age
ns(height, 6)
0.1
0.0
-0.1
-0.2
140
160
180
height
ns(weight, 6)
0.4
0.2
0.0
-0.2
-0.4
40
60
80
100
weight
ns(waistline, 6)
0.2
0.0
-0.2
-0.4
0
200
400
600
800
waistline
ns(sight_left, 6)
0.08
0.06
0.04
0.02
-0.02
0
2
4
6
8
10
sight_left
ns(sight_right, 6)
0.04
0.02
0.00
-0.02
-0.06
0
2
4
6
8
10
sight_right


ns(SBP, 6)
0.0
-0.2
-0.4
-0.6
100
150
200
250
SBP
ns(DBP, 6)
0.15
0.05
-0.05
-0.15
40
60
80
100
140
DBP
ns(BLDS, 6)
0.2
0.0
-0.2
-0.4
-0.6
200
400
600
800
BLDS
ns(tot_chole, 6)
4
2
0
-2
-4
0
500
1000
1500
2000
tot_chole
ns(HDL_chole, 6)
0.6
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
HDL_chole
ns(LDL_chole, 6)
2
0
-2
-4
0
500
1000
1500
2000
LDL_chole


ns(triglyceride, 6)
0.0
0.2
0.4
0.6
0
500
1500
2500
triglyceride
ns(hemoglobin, 6)
-0.1
0.0
0.1
0.2
0.3
5
10
15
20
hemoglobin
ns(urine_protein, 6)
-0.04
-0.02
0.00
0.02
1
2
3
4
5
6
urine_protein
ns(serum_creatinine, 6)
-0.6
-0.2
0.0
0.2
0.4
0.6
0
20
40
60
80
serum_creatinine
ns(SGOT_AST, 6)
0.0
0.5
1.0
1.5
2.0
2.5
0
500
1500
2500
SGOT_AST
ns(SGOT_ALT, 6)
-2.0
-1.5
-1.0
-0.5
0.0
0
500
1500
2500
SGOT_ALT



```
summary(hmiscgamfull)
```

```
##
## Call: gam(formula = Alcoholic.Status == "Y" ~ sex + ns(age, 6) + ns(height,
##     6) + ns(weight, 6) + ns(waistline, 6) + ns(sight_left, 6) +
##     ns(sight_right, 6) + hear_left + hear_right + ns(SBP, 6) +
##     ns(DBP, 6) + ns(BLDS, 6) + ns(tot_chole, 6) + ns(HDL_chole,
##     6) + ns(LDL_chole, 6) + ns(triglyceride, 6) + ns(hemoglobin,
##     6) + ns(urine_protein, 6) + ns(serum_creatinine, 6) + ns(SGOT_AST,
##     6) + ns(SGOT_ALT, 6) + ns(gamma_GTP, 6) + ns(BMI, 6) + BMI.Category +
##     AGE.Category + Smoking.Status, data = hmiscTrain)
## Deviance Residuals:
##       Min        1Q    Median        3Q       Max
## -1.199937 -0.345898  0.005286  0.343070  1.287322
##
## (Dispersion Parameter for gaussian family taken to be 0.1802)
##
##     Null Deviance: 17499.82 on 69999 degrees of freedom
## Residual Deviance: 12589.35 on 69873 degrees of freedom
## AIC: 78812.32
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                            Df  Sum Sq Mean Sq    F value    Pr(>F)
## sex                         1  2371.9 2371.95 13164.6965 < 2.2e-16 ***
## ns(age, 6)                  6  1178.3  196.38  1089.9411 < 2.2e-16 ***
## ns(height, 6)               6    24.8    4.13    22.9341 < 2.2e-16 ***
## ns(weight, 6)               6     4.8    0.79     4.3963 0.0001896 ***
## ns(waistline, 6)            6    12.0    2.00    11.1034 2.041e-12 ***
## ns(sight_left, 6)           6     1.2    0.21     1.1446 0.3332597
## ns(sight_right, 6)          6     2.0    0.33     1.8315 0.0887314 .
## hear_left                   1     0.1    0.11     0.5982 0.4392862
## hear_right                  1     0.1    0.05     0.2940 0.5876536
## ns(SBP, 6)                  6    35.6    5.93    32.9087 < 2.2e-16 ***
## ns(DBP, 6)                  6    24.4    4.07    22.5967 < 2.2e-16 ***
## ns(BLDS, 6)                 6    28.7    4.78    26.5043 < 2.2e-16 ***
## ns(tot_chole, 6)            6    18.7    3.11    17.2687 < 2.2e-16 ***
## ns(HDL_chole, 6)            6   306.9   51.15   283.8656 < 2.2e-16 ***
## ns(LDL_chole, 6)            6   107.3   17.89    99.2814 < 2.2e-16 ***
## ns(triglyceride, 6)         6    25.4    4.23    23.4623 < 2.2e-16 ***
## ns(hemoglobin, 6)           6    10.7    1.78     9.8813 6.349e-11 ***
## ns(urine_protein, 6)        1     0.2    0.16     0.8687 0.3513219
## ns(serum_creatinine, 6)     6    10.7    1.78     9.8519 6.893e-11 ***
## ns(SGOT_AST, 6)             6     5.7    0.95     5.2722 1.924e-05 ***
## ns(SGOT_ALT, 6)             6    73.5   12.25    68.0091 < 2.2e-16 ***
## ns(gamma_GTP, 6)            6   449.7   74.94   415.9492 < 2.2e-16 ***
## ns(BMI, 6)                  6    15.3    2.54    14.1213 3.848e-16 ***
## BMI.Category                3     0.5    0.15     0.8433 0.4699143
## AGE.Category                3     1.2    0.41     2.2935 0.0758202 .
## Smoking.Status              2   201.1  100.55   558.0515 < 2.2e-16 ***
## Residuals               69873 12589.4    0.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
hmiscgamfull.trainPredict <- predict(hmiscgamfull, hmiscTrain)
hmiscgamfull.trainPredict[hmiscgamfull.trainPredict > 0.5] <- 'Y'
hmiscgamfull.trainPredict[hmiscgamfull.trainPredict != 'Y'] <- 'N'
sum(hmiscgamfull.trainPredict == hmiscTrain$Alcoholic.Status) / length(hmiscgamfull.trainPredict)
```

```
## [1] 0.7296857
```

```
caret::confusionMatrix(table(hmiscgamfull.trainPredict, hmiscTrain$Alcoholic.Status))
```

```
## Confusion Matrix and Statistics
## 
## 
## hmiscgamfull.trainPredict     N     Y
##                         N 25714  9523
##                         Y  9399 25364
## 
##                Accuracy : 0.7297
##                  95% CI : (0.7264, 0.733)
##     No Information Rate : 0.5016
##     P-Value [Acc > NIR] : <2e-16
## 
##                   Kappa : 0.4594
## 
##  Mcnemar's Test P-Value : 0.3712
## 
##             Sensitivity : 0.7323
##             Specificity : 0.7270
##          Pos Pred Value : 0.7297
##          Neg Pred Value : 0.7296
##              Prevalence : 0.5016
##          Detection Rate : 0.3673
##    Detection Prevalence : 0.5034
##       Balanced Accuracy : 0.7297
## 
##        'Positive' Class : N
## 
```

```
hmiscgamfull.predict <- predict(hmiscgamfull, hmiscTest)

hmiscgamfull.predict[hmiscgamfull.predict > 0.5] <- 'Y'
hmiscgamfull.predict[hmiscgamfull.predict != 'Y'] <- 'N'

hmiscgamfull.predict <- as.data.frame(cbind("ID" = 1:30000, "Alcoholic.Status" = hmiscgamfull.predict))
write.csv(hmiscgamfull.predict, file = "stapholz.jack_kaggle5.csv", row.names = FALSE)
```

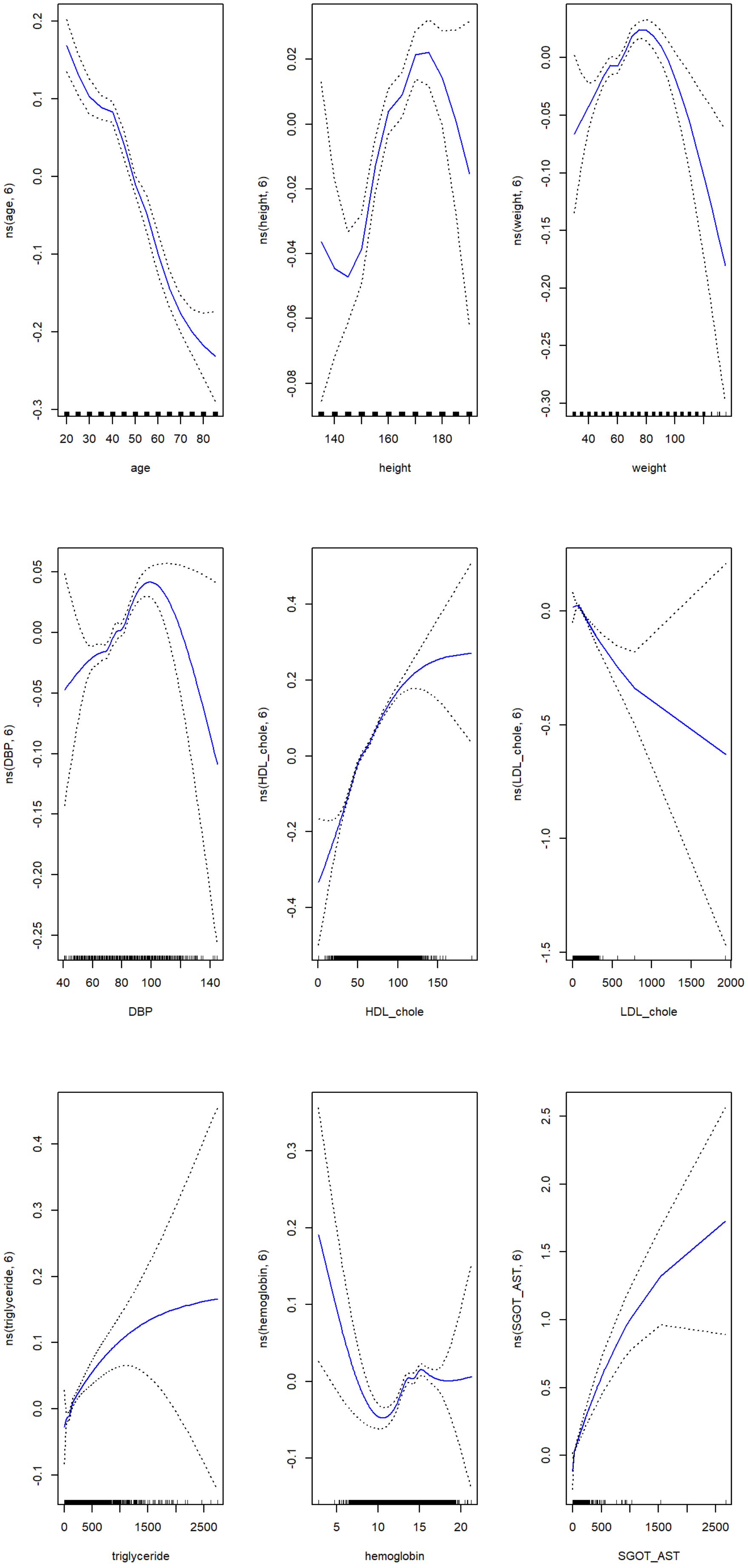
```
stepBIC.n <- dim(TrainSAData)[1]
stepBIC.mFull <- lm(Alcoholic.Status == "Y" ~ ., data = TrainSAData)
stepBIC.step <- step(stepBIC.mFull, direction = "backward", k = log(stepBIC.n))
```

```
## Start:  AIC=-116240
## Alcoholic.Status == "Y" ~ ID + sex + age + height + weight + 
##     waistline + sight_left + sight_right + hear_left + hear_right + 
##     SBP + DBP + BLDS + tot_chole + HDL_chole + LDL_chole + triglyceride + 
##     hemoglobin + urine_protein + serum_creatinine + SGOT_AST + 
##     SGOT_ALT + gamma_GTP + BMI + BMI.Category + AGE.Category + 
##     Smoking.Status
## 
##                    Df Sum of Sq   RSS     AIC
## - tot_chole         1      0.00 13243 -116251
## - hear_right        1      0.02 13243 -116251
## - sight_right       1      0.05 13243 -116251
## - hear_left         1      0.06 13243 -116251
## - waistline         1      0.07 13243 -116251
## - SBP               1      0.10 13243 -116251
## - BMI               1      0.22 13243 -116250
## - BLDS              1      0.23 13243 -116250
## - BMI.Category      1      0.40 13243 -116249
## - ID                1      0.63 13244 -116248
## - urine_protein     1      0.65 13244 -116248
## - serum_creatinine  1      0.75 13244 -116247
## - sight_left        1      0.86 13244 -116247
## <none>                          13243 -116240
## - weight            1      2.55 13246 -116238
## - LDL_chole         1      3.47 13246 -116233
## - AGE.Category      1      7.99 13251 -116209
## - SGOT_AST          1      8.42 13251 -116207
## - DBP               1     17.14 13260 -116161
## - height            1     17.58 13260 -116158
## - triglyceride      1     18.15 13261 -116155
## - hemoglobin        1     28.02 13271 -116103
## - SGOT_ALT          1     45.54 13288 -116011
## - HDL_chole         1    178.04 13421 -115316
## - gamma_GTP         1    178.23 13421 -115315
## - sex               1    189.53 13432 -115256
## - Smoking.Status    1    262.59 13506 -114877
## - age               1    530.12 13773 -113504
## 
## Step:  AIC=-116251.1
## Alcoholic.Status == "Y" ~ ID + sex + age + height + weight + 
##     waistline + sight_left + sight_right + hear_left + hear_right + 
##     SBP + DBP + BLDS + HDL_chole + LDL_chole + triglyceride + 
##     hemoglobin + urine_protein + serum_creatinine + SGOT_AST + 
##     SGOT_ALT + gamma_GTP + BMI + BMI.Category + AGE.Category + 
##     Smoking.Status
## 
##                    Df Sum of Sq   RSS     AIC
## - hear_right        1      0.02 13243 -116262
## - sight_right       1      0.05 13243 -116262
## - hear_left         1      0.06 13243 -116262
## - waistline         1      0.07 13243 -116262
## - SBP               1      0.10 13243 -116262
## - BMI               1      0.22 13243 -116261
## - BLDS              1      0.23 13243 -116261
## - BMI.Category      1      0.40 13243 -116260
## - ID                1      0.63 13244 -116259
## - urine_protein     1      0.65 13244 -116259
## - serum_creatinine  1      0.76 13244 -116258
## - sight_left        1      0.86 13244 -116258
## <none>                          13243 -116251
## - weight            1      2.55 13246 -116249
## - AGE.Category      1      7.99 13251 -116220
## - SGOT_AST          1      8.43 13251 -116218
## - DBP               1     17.14 13260 -116172
## - height            1     17.59 13260 -116169
## - LDL_chole         1     18.08 13261 -116167
## - triglyceride      1     26.88 13270 -116120
## - hemoglobin        1     28.05 13271 -116114
## - SGOT_ALT          1     45.57 13288 -116022
## - gamma_GTP         1    178.34 13421 -115326
## - sex               1    189.67 13433 -115267
## - HDL_chole         1    260.54 13503 -114898
## - Smoking.Status    1    262.59 13506 -114888
## - age               1    530.14 13773 -113515
## 
## Step:  AIC=-116262.2
## Alcoholic.Status == "Y" ~ ID + sex + age + height + weight + 
##     waistline + sight_left + sight_right + hear_left + SBP + 
##     DBP + BLDS + HDL_chole + LDL_chole + triglyceride + hemoglobin +
##     urine_protein + serum_creatinine + SGOT_AST + SGOT_ALT +
##     gamma_GTP + BMI + BMI.Category + AGE.Category + Smoking.Status
##
##                    Df Sum of Sq   RSS     AIC
## - hear_left         1      0.04 13243 -116273
## - sight_right       1      0.05 13243 -116273
## - waistline         1      0.07 13243 -116273
## - SBP               1      0.10 13243 -116273
## - BMI               1      0.22 13243 -116272
## - BLDS              1      0.23 13243 -116272
## - BMI.Category      1      0.40 13243 -116271
## - ID                1      0.63 13244 -116270
## - urine_protein     1      0.65 13244 -116270
## - serum_creatinine  1      0.76 13244 -116269
## - sight_left        1      0.85 13244 -116269
## <none>                          13243 -116262
## - weight            1      2.55 13246 -116260
## - AGE.Category      1      7.97 13251 -116231
## - SGOT_AST          1      8.43 13251 -116229
## - DBP               1     17.12 13260 -116183
## - height            1     17.58 13260 -116180
## - LDL_chole         1     18.09 13261 -116178
## - triglyceride      1     26.88 13270 -116131
## - hemoglobin        1     28.03 13271 -116125
## - SGOT_ALT          1     45.56 13288 -116033
## - gamma_GTP         1    178.36 13421 -115337
## - sex               1    189.83 13433 -115277
## - HDL_chole         1    260.53 13504 -114910
## - Smoking.Status    1    262.57 13506 -114899
## - age               1    533.85 13777 -113507
##
## Step:  AIC=-116273.1
## Alcoholic.Status == "Y" ~ ID + sex + age + height + weight +
##     waistline + sight_left + sight_right + SBP + DBP + BLDS +
##     HDL_chole + LDL_chole + triglyceride + hemoglobin + urine_protein +
##     serum_creatinine + SGOT_AST + SGOT_ALT + gamma_GTP + BMI +
##     BMI.Category + AGE.Category + Smoking.Status
##
##                    Df Sum of Sq   RSS     AIC
## - sight_right       1      0.05 13243 -116284
## - waistline         1      0.07 13243 -116284
## - SBP               1      0.10 13243 -116284
## - BMI               1      0.22 13243 -116283
## - BLDS              1      0.23 13243 -116283
## - BMI.Category      1      0.40 13243 -116282
## - ID                1      0.63 13244 -116281
## - urine_protein     1      0.65 13244 -116281
## - serum_creatinine  1      0.76 13244 -116280
## - sight_left        1      0.85 13244 -116280
## <none>                          13243 -116273
## - weight            1      2.55 13246 -116271
## - AGE.Category      1      8.07 13251 -116242
## - SGOT_AST          1      8.43 13251 -116240
## - DBP               1     17.14 13260 -116194
## - height            1     17.60 13261 -116191
## - LDL_chole         1     18.06 13261 -116189
## - triglyceride      1     26.90 13270 -116142
## - hemoglobin        1     28.06 13271 -116136
## - SGOT_ALT          1     45.55 13288 -116044
## - gamma_GTP         1    178.35 13421 -115348
## - sex               1    189.84 13433 -115288
## - HDL_chole         1    260.61 13504 -114920
## - Smoking.Status    1    262.53 13506 -114910
## - age               1    548.45 13791 -113444
##
## Step:  AIC=-116284
## Alcoholic.Status == "Y" ~ ID + sex + age + height + weight +
##     waistline + sight_left + SBP + DBP + BLDS + HDL_chole + LDL_chole +
##     triglyceride + hemoglobin + urine_protein + serum_creatinine +
##     SGOT_AST + SGOT_ALT + gamma_GTP + BMI + BMI.Category + AGE.Category +
##     Smoking.Status
##
##                    Df Sum of Sq   RSS     AIC
## - waistline         1      0.07 13243 -116295
## - SBP               1      0.10 13243 -116295
## - BMI               1      0.22 13243 -116294
## - BLDS              1      0.23 13243 -116294
## - BMI.Category      1      0.40 13243 -116293
## - ID                1      0.63 13244 -116292
## - urine_protein     1      0.65 13244 -116292
## - serum_creatinine  1      0.75 13244 -116291
## - sight_left        1      1.01 13244 -116290
## <none>                         13243 -116284
## - weight            1      2.56 13246 -116282
## - AGE.Category      1      8.12 13251 -116252
## - SGOT_AST          1      8.42 13251 -116251
## - DBP               1     17.15 13260 -116205
## - height            1     17.63 13261 -116202
## - LDL_chole         1     18.06 13261 -116200
## - triglyceride      1     26.88 13270 -116153
## - hemoglobin        1     28.10 13271 -116147
## - SGOT_ALT          1     45.54 13288 -116055
## - gamma_GTP         1    178.32 13421 -115359
## - sex               1    190.15 13433 -115297
## - HDL_chole         1    260.59 13504 -114931
## - Smoking.Status    1    262.52 13506 -114921
## - age               1    553.86 13797 -113427
##
## Step:  AIC=-116294.8
## Alcoholic.Status == "Y" ~ ID + sex + age + height + weight +
##     sight_left + SBP + DBP + BLDS + HDL_chole + LDL_chole + triglyceride +
##     hemoglobin + urine_protein + serum_creatinine + SGOT_AST +
##     SGOT_ALT + gamma_GTP + BMI + BMI.Category + AGE.Category +
##     Smoking.Status
##
##                    Df Sum of Sq   RSS     AIC
## - SBP               1      0.10 13243 -116305
## - BMI               1      0.18 13243 -116305
## - BLDS              1      0.23 13243 -116305
## - BMI.Category      1      0.39 13244 -116304
## - ID                1      0.63 13244 -116303
## - urine_protein     1      0.65 13244 -116303
## - serum_creatinine  1      0.76 13244 -116302
## - sight_left        1      1.00 13244 -116301
## <none>                         13243 -116295
## - weight            1      2.70 13246 -116292
## - AGE.Category      1      8.09 13251 -116263
## - SGOT_AST          1      8.41 13252 -116262
## - DBP               1     17.13 13260 -116215
## - height            1     17.70 13261 -116212
## - LDL_chole         1     18.05 13261 -116211
## - triglyceride      1     26.98 13270 -116164
## - hemoglobin        1     28.12 13271 -116158
## - SGOT_ALT          1     45.49 13289 -116066
## - gamma_GTP         1    178.61 13422 -115368
## - sex               1    190.62 13434 -115306
## - HDL_chole         1    260.59 13504 -114942
## - Smoking.Status    1    262.63 13506 -114931
## - age               1    566.42 13810 -113374
##
## Step:  AIC=-116305.5
## Alcoholic.Status == "Y" ~ ID + sex + age + height + weight +
##     sight_left + DBP + BLDS + HDL_chole + LDL_chole + triglyceride +
##     hemoglobin + urine_protein + serum_creatinine + SGOT_AST +
##     SGOT_ALT + gamma_GTP + BMI + BMI.Category + AGE.Category +
##     Smoking.Status
##
##                    Df Sum of Sq   RSS     AIC
## - BMI               1      0.20 13243 -116316
## - BLDS              1      0.22 13243 -116315
## - BMI.Category      1      0.39 13244 -116315
## - ID                1      0.63 13244 -116313
## - urine_protein     1      0.66 13244 -116313
## - serum_creatinine  1      0.76 13244 -116313
## - sight_left        1      1.01 13244 -116311
## <none>                         13243 -116305
## - weight            1      2.68 13246 -116302
## - AGE.Category      1      8.31 13252 -116273
## - SGOT_AST          1      8.40 13252 -116272
## - height            1     17.74 13261 -116223
## - LDL_chole         1     18.00 13261 -116222
## - DBP               1     26.78 13270 -116175
## - triglyceride      1     26.93 13270 -116174
## - hemoglobin        1     28.16 13271 -116168
## - SGOT_ALT          1     45.48 13289 -116077
## - gamma_GTP         1    178.51 13422 -115379
## - sex               1    190.53 13434 -115317
## - HDL_chole         1    260.53 13504 -114953
## - Smoking.Status    1    262.72 13506 -114942
## - age               1    598.34 13842 -113223
## 
## Step:  AIC=-116315.6
## Alcoholic.Status == "Y" ~ ID + sex + age + height + weight + 
##     sight_left + DBP + BLDS + HDL_chole + LDL_chole + triglyceride + 
##     hemoglobin + urine_protein + serum_creatinine + SGOT_AST + 
##     SGOT_ALT + gamma_GTP + BMI.Category + AGE.Category + Smoking.Status
## 
##                    Df Sum of Sq   RSS     AIC
## - BLDS              1      0.20 13244 -116326
## - BMI.Category      1      0.47 13244 -116324
## - ID                1      0.63 13244 -116323
## - urine_protein     1      0.66 13244 -116323
## - serum_creatinine  1      0.76 13244 -116323
## - sight_left        1      1.01 13244 -116321
## <none>                          13243 -116316
## - weight            1      5.31 13249 -116299
## - AGE.Category      1      8.29 13252 -116283
## - SGOT_AST          1      8.43 13252 -116282
## - LDL_chole         1     18.10 13262 -116231
## - DBP               1     26.60 13270 -116186
## - triglyceride      1     26.87 13270 -116185
## - hemoglobin        1     28.09 13272 -116178
## - height            1     29.43 13273 -116171
## - SGOT_ALT          1     45.72 13289 -116085
## - gamma_GTP         1    178.44 13422 -115390
## - sex               1    192.05 13435 -115319
## - HDL_chole         1    262.77 13506 -114951
## - Smoking.Status    1    262.86 13506 -114951
## - age               1    602.43 13846 -113213
## 
## Step:  AIC=-116325.7
## Alcoholic.Status == "Y" ~ ID + sex + age + height + weight + 
##     sight_left + DBP + HDL_chole + LDL_chole + triglyceride + 
##     hemoglobin + urine_protein + serum_creatinine + SGOT_AST + 
##     SGOT_ALT + gamma_GTP + BMI.Category + AGE.Category + Smoking.Status
## 
##                    Df Sum of Sq   RSS     AIC
## - BMI.Category      1      0.46 13244 -116334
## - urine_protein     1      0.61 13244 -116334
## - ID                1      0.63 13244 -116333
## - serum_creatinine  1      0.76 13244 -116333
## - sight_left        1      1.00 13245 -116332
## <none>                          13244 -116326
## - weight            1      5.44 13249 -116308
## - AGE.Category      1      8.35 13252 -116293
## - SGOT_AST          1      8.38 13252 -116293
## - LDL_chole         1     18.32 13262 -116240
## - DBP               1     26.81 13270 -116195
## - triglyceride      1     27.74 13271 -116190
## - hemoglobin        1     28.25 13272 -116188
## - height            1     29.35 13273 -116182
## - SGOT_ALT          1     45.60 13289 -116096
## - gamma_GTP         1    180.31 13424 -115390
## - sex               1    192.17 13436 -115328
## - HDL_chole         1    262.60 13506 -114962
## - Smoking.Status    1    262.93 13506 -114961
## - age               1    611.84 13855 -113175
## 
## Step:  AIC=-116334.4
## Alcoholic.Status == "Y" ~ ID + sex + age + height + weight + 
##     sight_left + DBP + HDL_chole + LDL_chole + triglyceride + 
##     hemoglobin + urine_protein + serum_creatinine + SGOT_AST + 
##     SGOT_ALT + gamma_GTP + AGE.Category + Smoking.Status
## 
##                    Df Sum of Sq   RSS     AIC
## - urine_protein     1      0.62 13245 -116342
## - ID                1      0.63 13245 -116342
## - serum_creatinine  1      0.76 13245 -116342
## - sight_left        1      1.00 13245 -116340
## <none>                          13244 -116334
## - weight            1      4.98 13249 -116319
## - SGOT_AST          1      8.39 13252 -116301
## - AGE.Category      1      8.47 13252 -116301
## - LDL_chole         1     18.16 13262 -116250
## - DBP               1     26.69 13271 -116205
## - triglyceride      1     27.57 13272 -116200
## - hemoglobin        1     28.24 13272 -116196
## - height            1     30.84 13275 -116183
## - SGOT_ALT          1     45.72 13290 -116104
## - gamma_GTP         1    180.00 13424 -115401
## - sex               1    192.31 13436 -115336
## - HDL_chole         1    262.67 13507 -114971
## - Smoking.Status    1    262.68 13507 -114971
## - age               1    611.54 13856 -113186
## 
## Step:  AIC=-116342.2
## Alcoholic.Status == "Y" ~ ID + sex + age + height + weight + 
##     sight_left + DBP + HDL_chole + LDL_chole + triglyceride + 
##     hemoglobin + serum_creatinine + SGOT_AST + SGOT_ALT + gamma_GTP + 
##     AGE.Category + Smoking.Status
## 
##                    Df Sum of Sq   RSS     AIC
## - ID                1      0.63 13245 -116350
## - serum_creatinine  1      0.86 13246 -116349
## - sight_left        1      1.01 13246 -116348
## <none>                          13245 -116342
## - weight            1      4.97 13250 -116327
## - SGOT_AST          1      8.29 13253 -116310
## - AGE.Category      1      8.51 13253 -116308
## - LDL_chole         1     18.11 13263 -116258
## - DBP               1     26.50 13271 -116213
## - triglyceride      1     27.43 13272 -116209
## - hemoglobin        1     28.40 13273 -116203
## - height            1     30.84 13276 -116191
## - SGOT_ALT          1     45.64 13290 -116113
## - gamma_GTP         1    179.59 13424 -115411
## - sex               1    192.21 13437 -115345
## - Smoking.Status    1    262.74 13507 -114978
## - HDL_chole         1    262.85 13508 -114978
## - age               1    612.98 13858 -113186
## 
## Step:  AIC=-116350.1
## Alcoholic.Status == "Y" ~ sex + age + height + weight + sight_left + 
##     DBP + HDL_chole + LDL_chole + triglyceride + hemoglobin + 
##     serum_creatinine + SGOT_AST + SGOT_ALT + gamma_GTP + AGE.Category + 
##     Smoking.Status
## 
##                    Df Sum of Sq   RSS     AIC
## - serum_creatinine  1      0.86 13246 -116357
## - sight_left        1      1.02 13246 -116356
## <none>                          13245 -116350
## - weight            1      4.95 13250 -116335
## - SGOT_AST          1      8.34 13254 -116317
## - AGE.Category      1      8.50 13254 -116316
## - LDL_chole         1     18.13 13263 -116265
## - DBP               1     26.47 13272 -116221
## - triglyceride      1     27.49 13273 -116216
## - hemoglobin        1     28.37 13274 -116211
## - height            1     30.84 13276 -116198
## - SGOT_ALT          1     45.72 13291 -116120
## - gamma_GTP         1    179.64 13425 -115418
## - sex               1    192.25 13438 -115352
## - Smoking.Status    1    262.66 13508 -114987
## - HDL_chole         1    262.78 13508 -114986
## - age               1    613.15 13858 -113194
## 
## Step:  AIC=-116356.7
## Alcoholic.Status == "Y" ~ sex + age + height + weight + sight_left + 
##     DBP + HDL_chole + LDL_chole + triglyceride + hemoglobin + 
##     SGOT_AST + SGOT_ALT + gamma_GTP + AGE.Category + Smoking.Status
## 
##                  Df Sum of Sq   RSS     AIC
## - sight_left      1      1.02 13247 -116362
## <none>                        13246 -116357
## - weight          1      4.91 13251 -116342
## - SGOT_AST        1      8.35 13254 -116324
## - AGE.Category    1      8.57 13255 -116323
## - LDL_chole       1     18.10 13264 -116272
## - DBP             1     26.43 13273 -116228
## - triglyceride    1     27.50 13274 -116223
## - hemoglobin      1     28.50 13275 -116217
## - height          1     30.63 13277 -116206
## - SGOT_ALT        1     45.82 13292 -116126
## - gamma_GTP       1    179.53 13426 -115425
## - sex             1    191.65 13438 -115362
## - Smoking.Status  1    262.41 13509 -114995
## - HDL_chole       1    263.13 13509 -114991
## - age             1    615.86 13862 -113187
## 
## Step:  AIC=-116362.4
## Alcoholic.Status == "Y" ~ sex + age + height + weight + DBP + 
##     HDL_chole + LDL_chole + triglyceride + hemoglobin + SGOT_AST + 
##     SGOT_ALT + gamma_GTP + AGE.Category + Smoking.Status
## 
##                  Df Sum of Sq   RSS     AIC
## <none>                        13247 -116362
## - weight          1      4.94 13252 -116347
## - SGOT_AST        1      8.34 13256 -116330
## - AGE.Category    1      8.85 13256 -116327
## - LDL_chole       1     18.08 13265 -116278
## - DBP             1     26.38 13274 -116234
## - triglyceride    1     27.53 13275 -116228
## - hemoglobin      1     28.64 13276 -116222
## - height          1     30.95 13278 -116210
## - SGOT_ALT        1     45.80 13293 -116132
## - gamma_GTP       1    179.32 13426 -115432
## - sex             1    192.36 13440 -115364
## - Smoking.Status  1    262.37 13510 -115001
## - HDL_chole       1    263.29 13510 -114996
## - age             1    633.39 13881 -113104
```

```
stepBIC.reducedModel <- lm(stepBIC.step$call, data = TrainSAData)
anova(stepBIC.reducedModel, stepBIC.mFull)
```

```
## Analysis of Variance Table
## 
## Model 1: Alcoholic.Status == "Y" ~ sex + age + height + weight + DBP + 
##     HDL_chole + LDL_chole + triglyceride + hemoglobin + SGOT_AST + 
##     SGOT_ALT + gamma_GTP + AGE.Category + Smoking.Status
## Model 2: Alcoholic.Status == "Y" ~ ID + sex + age + height + weight + 
##     waistline + sight_left + sight_right + hear_left + hear_right + 
##     SBP + DBP + BLDS + tot_chole + HDL_chole + LDL_chole + triglyceride + 
##     hemoglobin + urine_protein + serum_creatinine + SGOT_AST + 
##     SGOT_ALT + gamma_GTP + BMI + BMI.Category + AGE.Category + 
##     Smoking.Status
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)  
## 1  69985 13247                              
## 2  69972 13243 13    4.2677 1.7346 0.04744 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
hmiscgamBIC <- gam(Alcoholic.Status == "Y" ~ sex + ns(age, 6) + ns(height, 6) + ns(weight, 6) + ns(DBP, 6) + ns(HDL_chole, 
6) + ns(LDL_chole, 6) + ns(triglyceride, 6) + ns(hemoglobin, 6) + ns(SGOT_AST, 6) + ns(SGOT_ALT, 6) + ns(gamma_GTP, 6) + AG
E.Category + Smoking.Status, data = hmiscTrain)
par(mfrow = c(1, 3))
plot(hmiscgamBIC, se = TRUE, col = "blue")
```


ns(age, 6)
0.2
0.1
0.0
-0.1
-0.2
-0.3
20 30 40 50 60 70 80
age
ns(height, 6)
0.02
0.00
-0.02
-0.04
-0.06
-0.08
140 160 180
height
ns(weight, 6)
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
40 60 80 100
weight
ns(DBP, 6)
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
40 60 80 100 140
DBP
ns(HDL_chole, 6)
0.4
0.2
0.0
-0.2
-0.4
0 50 100 150
HDL_chole
ns(LDL_chole, 6)
0.0
-0.5
-1.0
-1.5
0 500 1000 1500 2000
LDL_chole
ns(triglyceride, 6)
0.4
0.3
0.2
0.1
0.0
-0.1
0 500 1500 2500
triglyceride
ns(hemoglobin, 6)
0.3
0.2
0.1
0.0
-0.1
5 10 15 20
hemoglobin
ns(SGOT_AST, 6)
2.5
2.0
1.5
1.0
0.5
0.0
0 500 1500 2500
SGOT_AST

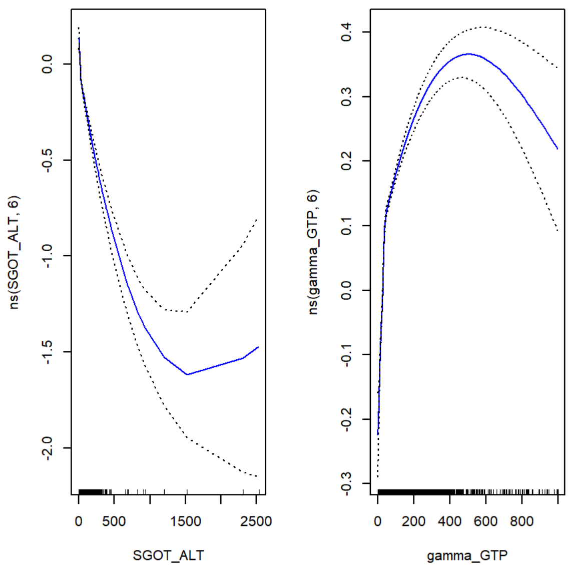
```
summary(hmiscgamBIC)
```

```
##
## Call: gam(formula = Alcoholic.Status == "Y" ~ sex + ns(age, 6) + ns(height,
##     6) + ns(weight, 6) + ns(DBP, 6) + ns(HDL_chole, 6) + ns(LDL_chole,
##     6) + ns(triglyceride, 6) + ns(hemoglobin, 6) + ns(SGOT_AST,
##     6) + ns(SGOT_ALT, 6) + ns(gamma_GTP, 6) + AGE.Category +
##     Smoking.Status, data = hmiscTrain)
## Deviance Residuals:
##       Min        1Q    Median        3Q       Max
## -1.191821 -0.346908  0.006664  0.343676  1.312965
##
## (Dispersion Parameter for gaussian family taken to be 0.1809)
##
##     Null Deviance: 17499.82 on 69999 degrees of freedom
## Residual Deviance: 12648.79 on 69927 degrees of freedom
## AIC: 79034.05
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                        Df  Sum Sq Mean Sq    F value    Pr(>F)
## sex                     1  2371.9 2371.95 13112.9574 < 2.2e-16 ***
## ns(age, 6)              6  1178.3  196.38  1085.6575 < 2.2e-16 ***
## ns(height, 6)           6    24.8    4.13    22.8440 < 2.2e-16 ***
## ns(weight, 6)           6     4.8    0.79     4.3790 0.0001983 ***
## ns(DBP, 6)              6    60.0   10.00    55.2900 < 2.2e-16 ***
## ns(HDL_chole, 6)        6   303.5   50.58   279.6512 < 2.2e-16 ***
## ns(LDL_chole, 6)        6    34.8    5.81    32.0949 < 2.2e-16 ***
## ns(triglyceride, 6)     6   114.1   19.01   105.1004 < 2.2e-16 ***
## ns(hemoglobin, 6)       6    11.9    1.99    10.9741 2.940e-12 ***
## ns(SGOT_AST, 6)         6     6.2    1.03     5.6675 6.734e-06 ***
## ns(SGOT_ALT, 6)         6    64.8   10.81    59.7485 < 2.2e-16 ***
## ns(gamma_GTP, 6)        6   471.4   78.56   434.3219 < 2.2e-16 ***
## AGE.Category            3     1.2    0.40     2.2169 0.0839168 .
## Smoking.Status          2   203.4  101.68   562.1040 < 2.2e-16 ***
## Residuals           69927 12648.8    0.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
hmiscgamBIC.trainPredict <- predict(hmiscgamBIC, hmiscTrain)
hmiscgamBIC.trainPredict[hmiscgamBIC.trainPredict > 0.5] <- 'Y'
hmiscgamBIC.trainPredict[hmiscgamBIC.trainPredict != 'Y'] <- 'N'
sum(hmiscgamBIC.trainPredict == hmiscTrain$Alcoholic.Status) / length(hmiscgamBIC.trainPredict)
```

```
## [1] 0.7276857
```

```
hmiscgamBIC.predict <- predict(hmiscgamBIC, hmiscTest)

hmiscgamBIC.predict[hmiscgamBIC.predict > 0.5] <- 'Y'
hmiscgamBIC.predict[hmiscgamBIC.predict != 'Y'] <- 'N'

hmiscgamBIC.predict <- as.data.frame(cbind("ID" = 1:30000, "Alcoholic.Status" = hmiscgamBIC.predict))
write.csv(hmiscgamBIC.predict, file = "stapholz.jack_kaggle8.csv", row.names = FALSE)
```



```
anova(hmisclm, hmiscgamfull)
```

```
## Analysis of Variance Table
##
## Model 1: Alcoholic.Status == "Y" ~ sex + age + height + weight + waistline +
##     sight_left + sight_right + hear_left + hear_right + SBP +
##     DBP + BLDS + tot_chole + HDL_chole + LDL_chole + triglyceride +
##     hemoglobin + urine_protein + serum_creatinine + SGOT_AST +
##     SGOT_ALT + gamma_GTP + BMI + BMI.Category + AGE.Category +
##     Smoking.Status
## Model 2: Alcoholic.Status == "Y" ~ sex + ns(age, 6) + ns(height, 6) +
##     ns(weight, 6) + ns(waistline, 6) + ns(sight_left, 6) + ns(sight_right,
##     6) + hear_left + hear_right + ns(SBP, 6) + ns(DBP, 6) + ns(BLDS,
##     6) + ns(tot_chole, 6) + ns(HDL_chole, 6) + ns(LDL_chole,
##     6) + ns(triglyceride, 6) + ns(hemoglobin, 6) + ns(urine_protein,
##     6) + ns(serum_creatinine, 6) + ns(SGOT_AST, 6) + ns(SGOT_ALT,
##     6) + ns(gamma_GTP, 6) + ns(BMI, 6) + BMI.Category + AGE.Category +
##     Smoking.Status
##   Res.Df   RSS Df Sum of Sq     F    Pr(>F)
## 1  69968 12992
## 2  69873 12589 95    402.59 23.52 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(hmiscgam3deg, hmiscgamfull)
```

```
## Analysis of Deviance Table
##
## Model 1: Alcoholic.Status == "Y" ~ sex + ns(age, 3) + ns(height, 3) +
##     ns(weight, 3) + ns(waistline, 3) + ns(sight_left, 3) + ns(sight_right,
##     3) + hear_left + hear_right + ns(SBP, 3) + ns(DBP, 3) + ns(BLDS,
##     3) + ns(tot_chole, 3) + ns(HDL_chole, 3) + ns(LDL_chole,
##     3) + ns(triglyceride, 3) + ns(hemoglobin, 3) + ns(urine_protein,
##     3) + ns(serum_creatinine, 3) + ns(SGOT_AST, 3) + ns(SGOT_ALT,
##     3) + ns(gamma_GTP, 3) + ns(BMI, 3) + BMI.Category + AGE.Category +
##     Smoking.Status
## Model 2: Alcoholic.Status == "Y" ~ sex + ns(age, 6) + ns(height, 6) +
##     ns(weight, 6) + ns(waistline, 6) + ns(sight_left, 6) + ns(sight_right,
##     6) + hear_left + hear_right + ns(SBP, 6) + ns(DBP, 6) + ns(BLDS,
##     6) + ns(tot_chole, 6) + ns(HDL_chole, 6) + ns(LDL_chole,
##     6) + ns(triglyceride, 6) + ns(hemoglobin, 6) + ns(urine_protein,
##     6) + ns(serum_creatinine, 6) + ns(SGOT_AST, 6) + ns(SGOT_ALT,
##     6) + ns(gamma_GTP, 6) + ns(BMI, 6) + BMI.Category + AGE.Category +
##     Smoking.Status
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     69930      12702
## 2     69873      12589 57   112.69 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(hmiscgamBIC, hmiscgamfull)
```

```
## Analysis of Deviance Table
##
## Model 1: Alcoholic.Status == "Y" ~ sex + ns(age, 6) + ns(height, 6) +
##     ns(weight, 6) + ns(DBP, 6) + ns(HDL_chole, 6) + ns(LDL_chole,
##     6) + ns(triglyceride, 6) + ns(hemoglobin, 6) + ns(SGOT_AST,
##     6) + ns(SGOT_ALT, 6) + ns(gamma_GTP, 6) + AGE.Category +
##     Smoking.Status
## Model 2: Alcoholic.Status == "Y" ~ sex + ns(age, 6) + ns(height, 6) +
##     ns(weight, 6) + ns(waistline, 6) + ns(sight_left, 6) + ns(sight_right,
##     6) + hear_left + hear_right + ns(SBP, 6) + ns(DBP, 6) + ns(BLDS,
##     6) + ns(tot_chole, 6) + ns(HDL_chole, 6) + ns(LDL_chole,
##     6) + ns(triglyceride, 6) + ns(hemoglobin, 6) + ns(urine_protein,
##     6) + ns(serum_creatinine, 6) + ns(SGOT_AST, 6) + ns(SGOT_ALT,
##     6) + ns(gamma_GTP, 6) + ns(BMI, 6) + BMI.Category + AGE.Category +
##     Smoking.Status
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     69927      12649
## 2     69873      12589 54   59.441 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
{R BIC TREE} # stepBIC.n <- dim(TrainSAData)[1] # stepBIC.mF

{R ANN} # ann.train <- TrainSAData[, c(3:8, 11:24, 28)] # an

{R} # hstepAIC.n <- dim(hmiscTrain)[1] # hstepAIC.mFull <- l

{R Hmisc2} # hstepglm <- glm(hstepAIC.step$call, data = hmis

{R Hmisc3} # hstepgam <- gam(Alcoholic.Status == "Y" ~ sex +

{R Hmisc 5} # hstepgamfull <- gam(Alcoholic.Status == "Y" ~

{R Hmisc6} # hstepgam <- gam(Alcoholic.Status == "Y" ~ sex +

{R Hmisc 7} # hmiscgamfull <- gam(Alcoholic.Status == "Y" ~

{R Hmisc 8} # hstepgamfull <- gam(Alcoholic.Status == "Y" ~

{R Hmisc 9} # hmiscTrainNumerical <- hmiscTrain[, c(2:7, 10:

{R} # kag5 <- read.csv("../Group Project/stapholz.jack_kaggl
```