Heart Disease Prediction

1. Problem Statement

The project aims to predict the probability of heart disease of a patient using a range of demographic, health, nutrition and biochemical variables, such as age, ethnicity, smoking status, blood pressure and perperfluorooctanoic acid(PFOA). The project also aims to find out the set of variables that has the most impact on the occurrence of heart disease.

2. Data Import & Check

Libraries needed

```
# Loading the library
library(knitr)
library(statmod)
library("mgcv")
library(MuMIn)
library(pROC)
```

Dataset

```
# Loading the data
pfoa.df <- read.csv("heart_disease.csv")
head(pfoa.df)
```

```
##   hadcvd pfoa4 DMDEDUC LBXGH RIDAGEEX RIDRETH1 RIAGENDR BPXSAR BPXDAR smoking
## 1  FALSE    Q2       2   5.5      518        4        1    142     95 Current
## 2  FALSE    Q2       1   4.6       NA        4        2    139     60   Never
## 3   TRUE    Q4       2   5.5      745        3        1    124     71  Former
## 4  FALSE    Q1       2   5.5      738        4        2    110     66   Never
## 5  FALSE    Q3       3   4.7      628        1        1    101     75   Never
## 6  FALSE    Q1       1   5.6      494        5        2    102     70   Never
##   LBXTC
## 1   140
## 2   164
## 3   216
## 4   244
## 5   147
## 6   145
```

```
str(pfoa.df)
```

```
## 'data.frame':    1737 obs. of  11 variables:
##  $ hadcvd  : logi  FALSE FALSE TRUE FALSE FALSE FALSE ...
##  $ pfoa4   : chr  "Q2" "Q2" "Q4" "Q1" ...
##  $ DMDEDUC : int  2 1 2 2 3 1 3 1 3 2 ...
##  $ LBXGH   : num  5.5 4.6 5.5 5.5 4.7 5.6 5.1 5.2 5.2 5.1 ...
##  $ RIDAGEEX: int  518 NA 745 738 628 494 719 780 677 487 ...
##  $ RIDRETH1: int  4 4 3 4 1 5 3 3 3 2 ...
##  $ RIAGENDR: int  1 2 1 2 1 2 2 1 1 2 ...
##  $ BPXSAR  : int  142 139 124 110 101 102 119 134 118 92 ...
##  $ BPXDAR  : int  95 60 71 66 75 70 74 52 82 59 ...
##  $ smoking : chr  "Current" "Never" "Former" "Never" ...
##  $ LBXTC   : int  140 164 216 244 147 145 175 211 207 187 ...
```

```
summary(pfoa.df)
```

```
##    hadcvd           pfoa4              DMDEDUC          LBXGH
##  Mode :logical   Length:1737        Min.   :1.000   Min.   : 4.100
##  FALSE:1468      Class :character   1st Qu.:1.000   1st Qu.: 5.300
##  TRUE :269       Mode  :character   Median :2.000   Median : 5.500
##                                     Mean   :2.003   Mean   : 5.856
##                                     3rd Qu.:3.000   3rd Qu.: 5.900
##                                     Max.   :9.000   Max.   :13.700
##                                                     NA's   :36
##     RIDAGEEX         RIDRETH1       RIAGENDR         BPXSAR
##  Min.   : 480.0   Min.   :1.00   Min.   :1.000   Min.   :  0.0
##  1st Qu.: 594.0   1st Qu.:2.00   1st Qu.:1.000   1st Qu.:118.0
##  Median : 742.0   Median :3.00   Median :2.000   Median :130.0
##  Mean   : 735.6   Mean   :2.72   Mean   :1.519   Mean   :133.2
##  3rd Qu.: 862.0   3rd Qu.:3.00   3rd Qu.:2.000   3rd Qu.:145.0
##  Max.   :1019.0   Max.   :5.00   Max.   :2.000   Max.   :248.0
##  NA's   :100                                     NA's   :110
##      BPXDAR         smoking              LBXTC
##  Min.   :  0.00   Length:1737        Min.   : 99.0
##  1st Qu.: 64.00   Class :character   1st Qu.:181.0
##  Median : 72.00   Mode  :character   Median :208.0
##  Mean   : 71.01                      Mean   :209.6
##  3rd Qu.: 80.00                      3rd Qu.:234.0
##  Max.   :113.00                      Max.   :394.0
##  NA's   :110                         NA's   :46
```

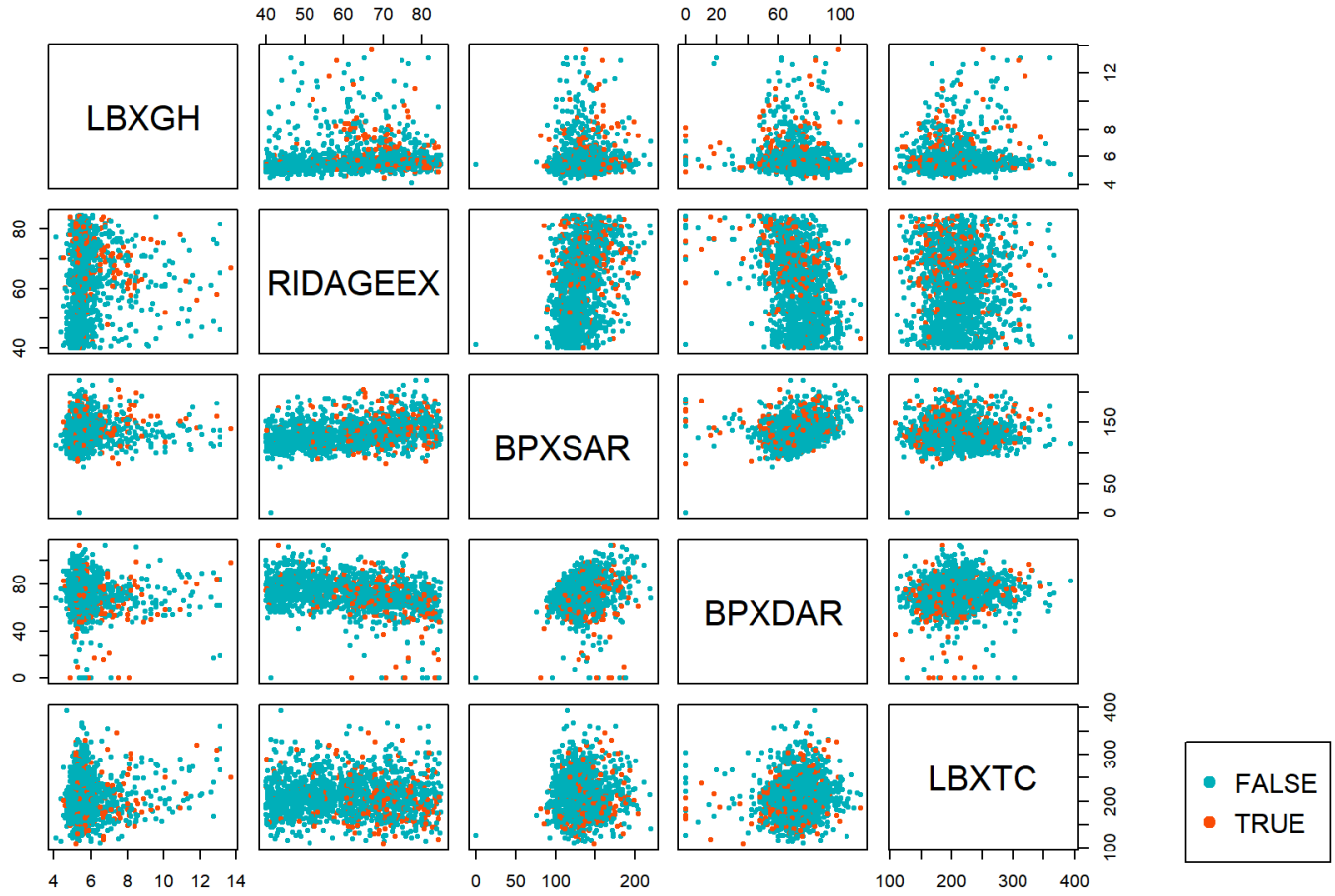
3. Data Pre-processing

```
#Convert pfoa4, DMDEDUC, RIDRETH1, RIAGENDR and smoking as categorical variables
pfoa.df$pfoa4 <- as.factor(pfoa.df$pfoa4)
pfoa.df$DMDEDUC <- as.factor(pfoa.df$DMDEDUC)
pfoa.df$RIDRETH1 <- as.factor(pfoa.df$RIDRETH1)
pfoa.df$RIAGENDR <- as.factor(pfoa.df$RIAGENDR)
pfoa.df$smoking <- as.factor(pfoa.df$smoking)

#Convert RIDAGEEX to age in years
pfoa.df$RIDAGEEX <- pfoa.df$RIDAGEEX/12

#Rename levels for DMDEDUC
levels(pfoa.df$DMDEDUC)[levels(pfoa.df$DMDEDUC)=="1"] <- "< High school"
levels(pfoa.df$DMDEDUC)[levels(pfoa.df$DMDEDUC)=="2"] <- "High school"
levels(pfoa.df$DMDEDUC)[levels(pfoa.df$DMDEDUC)=="3"] <- "> High school"
#Treat 'Refused' and 'Don't know' as NA for DMDEDUC
levels(pfoa.df$DMDEDUC)[levels(pfoa.df$DMDEDUC)=="7"] <- NA
levels(pfoa.df$DMDEDUC)[levels(pfoa.df$DMDEDUC)=="9"] <- NA
pfoa.df$DMDEDUC <- droplevels(pfoa.df$DMDEDUC)

#Rename levels for RIDRETH1
levels(pfoa.df$RIDRETH1)[levels(pfoa.df$RIDRETH1)=="1"] <- "Mexican Hispanic"
levels(pfoa.df$RIDRETH1)[levels(pfoa.df$RIDRETH1)=="2"] <- "Other Hispanic"
levels(pfoa.df$RIDRETH1)[levels(pfoa.df$RIDRETH1)=="3"] <- "non-Hispanic White"
levels(pfoa.df$RIDRETH1)[levels(pfoa.df$RIDRETH1)=="4"] <- "non-Hispanic Black"
levels(pfoa.df$RIDRETH1)[levels(pfoa.df$RIDRETH1)=="5"] <- "other"

#Rename levels for RIAGENDR
levels(pfoa.df$RIAGENDR)[levels(pfoa.df$RIAGENDR)=="1"] <- "Male"
levels(pfoa.df$RIAGENDR)[levels(pfoa.df$RIAGENDR)=="2"] <- "Female"
```

```
#Omit NA values
pfoa.df <- na.omit(pfoa.df)
row.names(pfoa.df) <- 1:nrow(pfoa.df)
```

Pairs plot of the numeric variables

```
#Pairs plot of the numeric variables
my_cols <- c("#00AFBB", "#FC4E07")
pairs(pfoa.df[,c(4,5,8,9,11)], pch = 20, col=my_cols[factor(pfoa.df$hadcvd)], oma=c(3,3,3,15))
par(xpd = TRUE)
legend("bottomright", col=my_cols[c(1:2)], legend= levels(factor(pfoa.df$hadcvd)), pch=19)
```




```
#Identify the outlier
subset(pfoa.df, BPXSAR == 0)
```

```
##      hadcvd pfoa4     DMDEDUC LBXGH RIDAGEEX           RIDRETH1 RIAGENDR BPXSAR
## 1222  FALSE    Q1 High school   5.4 41.16667 non-Hispanic Black     Male      0
##      BPXDAR smoking LBXTC
## 1222      0 Current   128
```

```
#Delete the outlier
pfoa.new <- pfoa.df[-c(1222),]
row.names(pfoa.new) <- 1:nrow(pfoa.new)
```

The plot indicates some unusual observations for BPXSAR. This observation seems to be a mistake because it has 0 value of BPXSAR which is impossible. It may be useful to delete this.

```
#Split the data set into training and test data
set.seed(330)
N <- 1000
sam<-sample(1:nrow(pfoa.new))
pfoa1.df<-pfoa.new[sam[1:N], ]
row.names(pfoa1.df)<-1:nrow(pfoa1.df)
pfoa2.df<-pfoa.new[sam[(N+1):nrow(pfoa.new)], ]
row.names(pfoa2.df)<-1:nrow(pfoa2.df)

#Summary
summary(pfoa1.df)
```

```
##    hadcvd        pfoa4             DMDEDUC        LBXGH           RIDAGEEX
##  Mode :logical   Q1:337   < High school:397   Min.   : 4.100   Min.   :40.00
##  FALSE:871       Q2:247   High school  :241   1st Qu.: 5.300   1st Qu.:49.83
##  TRUE :129       Q3:232   > High school:362   Median : 5.500   Median :61.54
##                  Q4:184                       Mean   : 5.851   Mean   :61.18
##                                               3rd Qu.: 5.900   3rd Qu.:71.92
##                                               Max.   :13.700   Max.   :84.92
##                RIDRETH1     RIAGENDR       BPXSAR          BPXDAR
##  Mexican Hispanic  :252   Male  :484   Min.   : 76.0   Min.   :  0.00
##  Other Hispanic    : 40   Female:516   1st Qu.:118.0   1st Qu.: 64.00
##  non-Hispanic White:507                Median :130.0   Median : 72.50
##  non-Hispanic Black:180                Mean   :132.4   Mean   : 71.62
##  other             : 21                3rd Qu.:144.2   3rd Qu.: 80.00
##                                        Max.   :220.0   Max.   :113.00
##     smoking        LBXTC
##  Current:207   Min.   :112.0
##  Former :337   1st Qu.:181.8
##  Never  :456   Median :209.0
##                Mean   :210.4
##                3rd Qu.:233.0
##                Max.   :360.0
```

```
#Summary
summary(pfoa2.df)
```

```
##    hadcvd        pfoa4             DMDEDUC        LBXGH           RIDAGEEX
##  Mode :logical   Q1:142   < High school:177   Min.   : 4.40   Min.   :40.08
##  FALSE:378       Q2:125   High school  :108   1st Qu.: 5.30   1st Qu.:49.33
##  TRUE :78        Q3: 99   > High school:171   Median : 5.50   Median :61.67
##                  Q4: 90                       Mean   : 5.87   Mean   :60.81
##                                               3rd Qu.: 5.90   3rd Qu.:70.79
##                                               Max.   :12.90   Max.   :84.42
##                RIDRETH1     RIAGENDR       BPXSAR          BPXDAR
##  Mexican Hispanic  :132   Male  :239   Min.   : 86.0   Min.   :  0.00
##  Other Hispanic    : 20   Female:217   1st Qu.:118.0   1st Qu.: 66.00
##  non-Hispanic White:205                Median :130.0   Median : 73.00
##  non-Hispanic Black: 86                Mean   :132.4   Mean   : 72.78
##  other             : 13                3rd Qu.:142.0   3rd Qu.: 81.00
##                                        Max.   :211.0   Max.   :113.00
##     smoking        LBXTC
##  Current: 68   Min.   :109.0
##  Former :150   1st Qu.:181.8
##  Never  :238   Median :206.5
##                Mean   :209.7
##                3rd Qu.:235.2
##                Max.   :394.0
```

4. Model building: Part 1

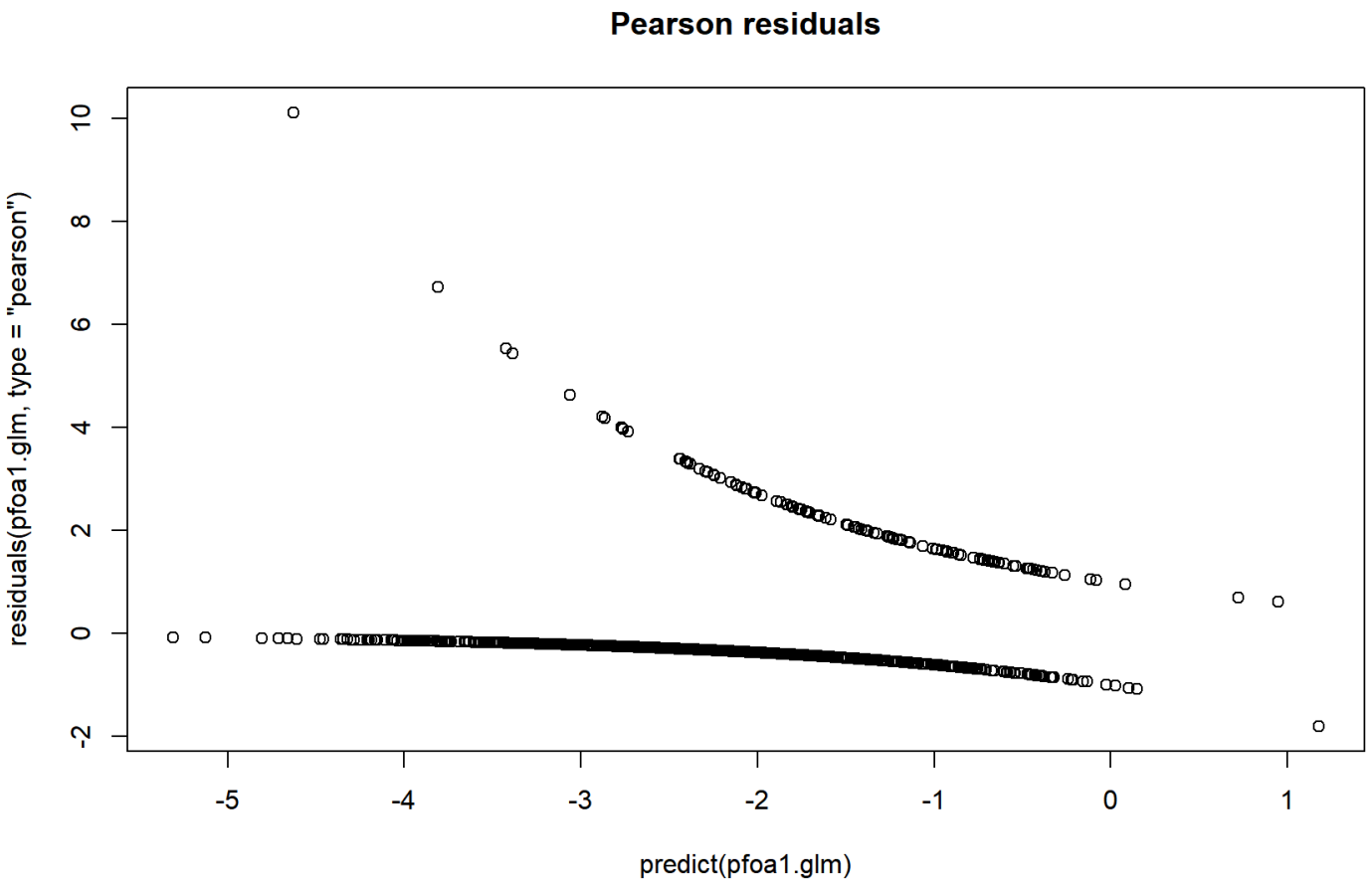
Model 1: GLM Model

The response variable, hadcvd, is binary data, so I will fit a binomial logistic regression model.

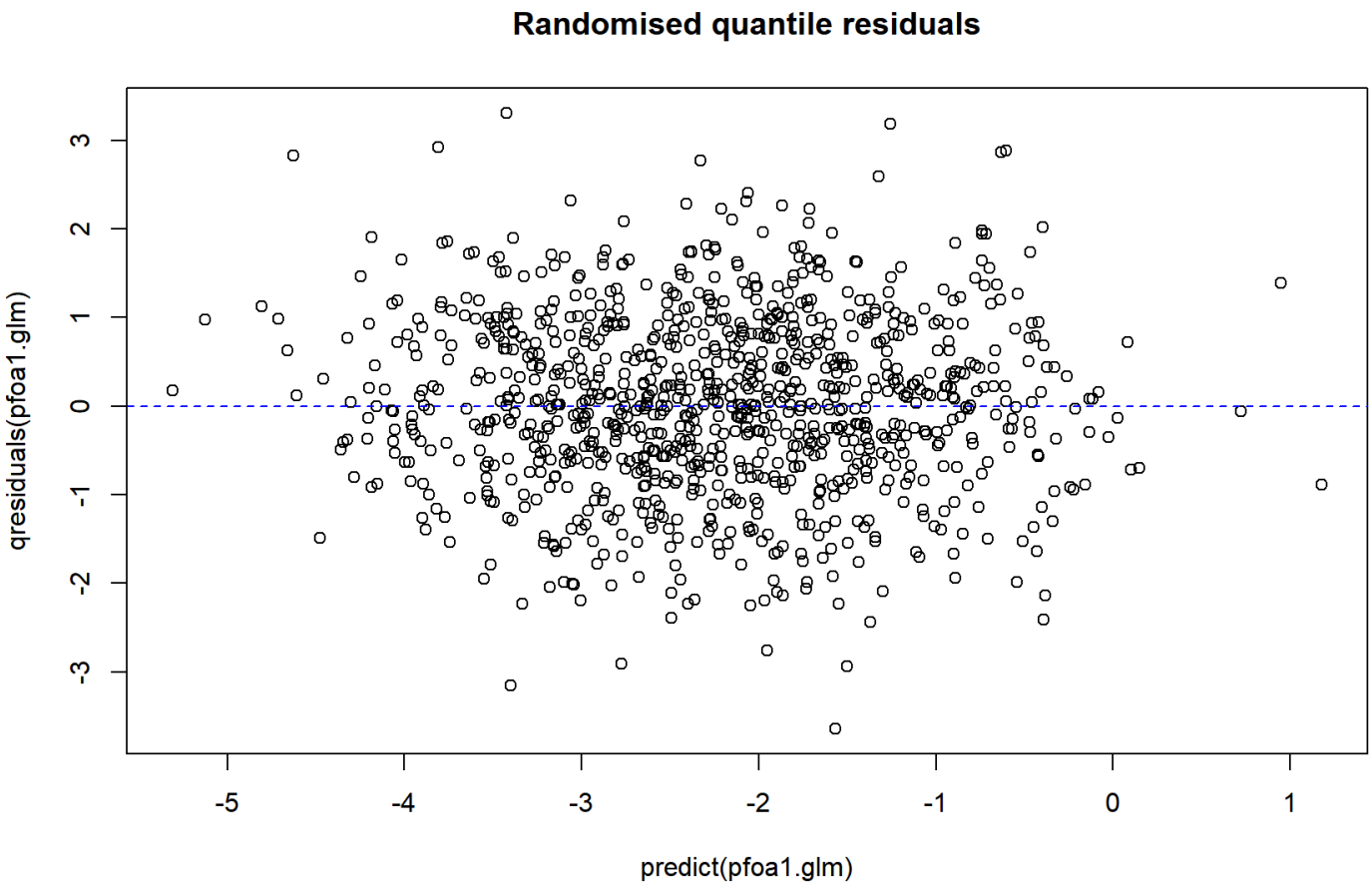
```
#Logistic regression model with all the variables
pfoa1.glm<-glm(hadcvd ~ pfoa4 + DMDEDUC + LBXGH + RIDAGEEX + RIDRETH1 + RIAGENDR + BPXSAR + BPXDAR + smoking + LBX
TC, family="binomial", data=pfoa1.df)
summary(pfoa1.glm)
```

```
##
## Call:
## glm(formula = hadcvd ~ pfoa4 + DMDEDUC + LBXGH + RIDAGEEX + RIDRETH1 +
##     RIAGENDR + BPXSAR + BPXDAR + smoking + LBXTC, family = "binomial",
##     data = pfoa1.df)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -5.205845   1.130194  -4.606 4.10e-06 ***
## pfoa4Q2                     0.246058   0.286644   0.858 0.390666
## pfoa4Q3                     0.666208   0.278133   2.395 0.016608 *
## pfoa4Q4                     0.817121   0.303964   2.688 0.007184 **
## DMDEDUCHigh school         -0.386473   0.272756  -1.417 0.156507
## DMDEDUC> High school       -0.529422   0.251880  -2.102 0.035563 *
## LBXGH                       0.200415   0.076460   2.621 0.008763 **
## RIDAGEEX                    0.048643   0.010769   4.517 6.27e-06 ***
## RIDRETH1Other Hispanic      0.471826   0.599572   0.787 0.431318
## RIDRETH1non-Hispanic White  0.732686   0.317405   2.308 0.020979 *
## RIDRETH1non-Hispanic Black  0.759206   0.349021   2.175 0.029612 *
## RIDRETH1other               2.034333   0.572061   3.556 0.000376 ***
## RIAGENDRFemale             -0.183459   0.216158  -0.849 0.396033
## BPXSAR                      0.004467   0.005142   0.869 0.385013
## BPXDAR                     -0.008439   0.006704  -1.259 0.208139
## smokingFormer              -0.302772   0.281129  -1.077 0.281487
## smokingNever               -0.753554   0.286834  -2.627 0.008611 **
## LBXTC                      -0.006098   0.002538  -2.403 0.016277 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 768.96  on 999  degrees of freedom
## Residual deviance: 676.26  on 982  degrees of freedom
## AIC: 712.26
##
## Number of Fisher Scoring iterations: 5
```

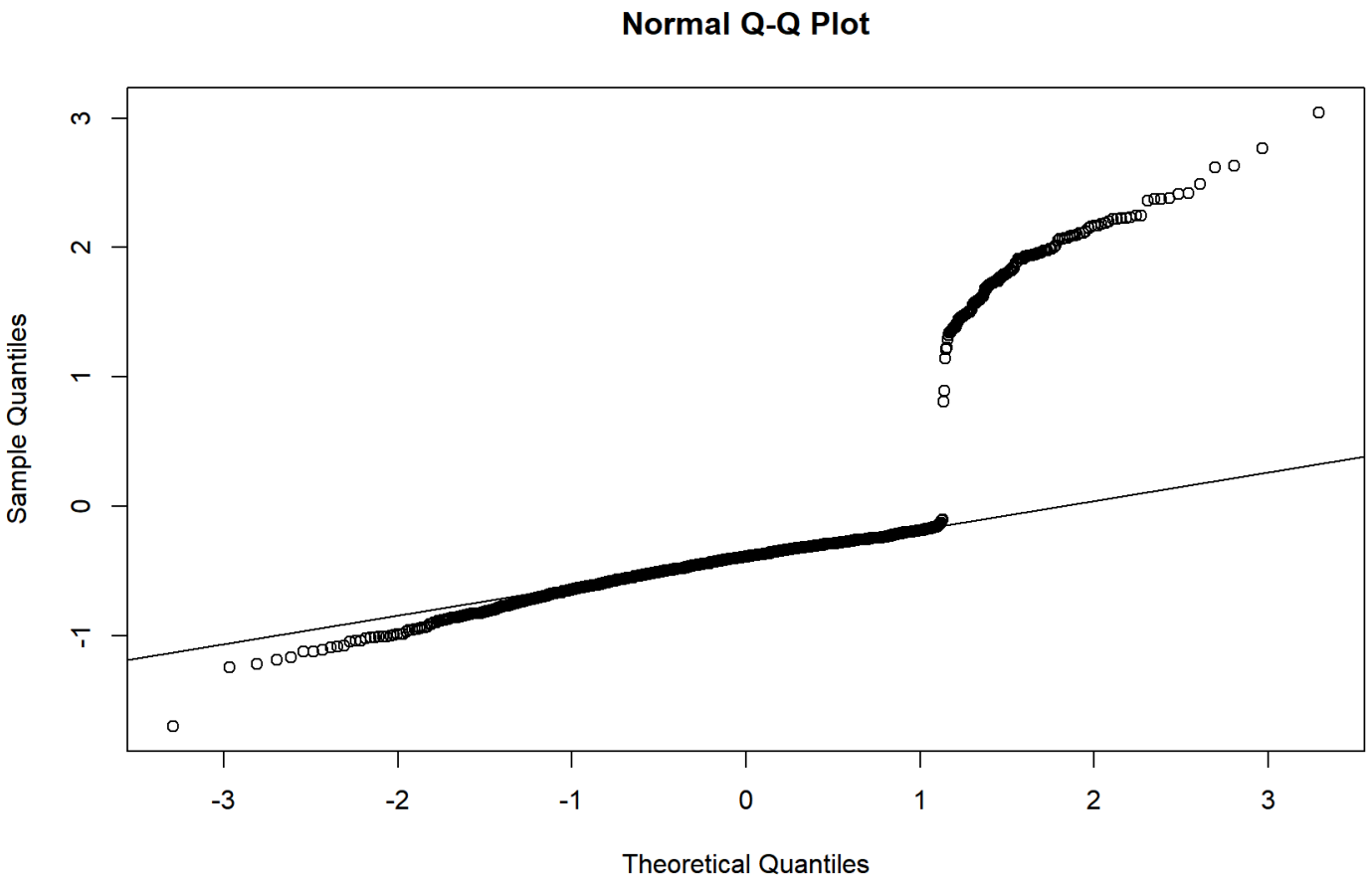
```
#Goodness of fit
plot(predict(pfoa1.glm), residuals(pfoa1.glm, type="pearson"), main="Pearson residuals")
```




```
plot(predict(pfoa1.glm), qresiduals(pfoa1.glm), main="Randomised quantile residuals")
abline(h=0, lty="dashed", col='blue')
```




```
qqnorm(residuals(pfoa1.glm))
qqline(residuals(pfoa1.glm))
```



We have strong evidence to keep RIDAGEEX, but BPXSAR and BPXDAR are not statistically significant. The normal Q-Q plot indicates some non-normality, but this does not seem to be a big issue. Now, I will use dredge() function to search for a suitable predictive model.

Model 2: GAM model

Now, I will fit a GAM model to see whether non-linear terms should be added.

```
#GAM model
pfoa1.gam <- gam(hadcvd ~ pfoa4 + DMDEDUC + s(LBXGH) + s(RIDAGEEX) + RIDRETH1 + RIAGENDR + s(BPXSAR) + s(BPXDAR) +
smoking + s(LBXTC), family="binomial", data=pfoa1.df)
print(summary(pfoa1.gam)$p.table)
```

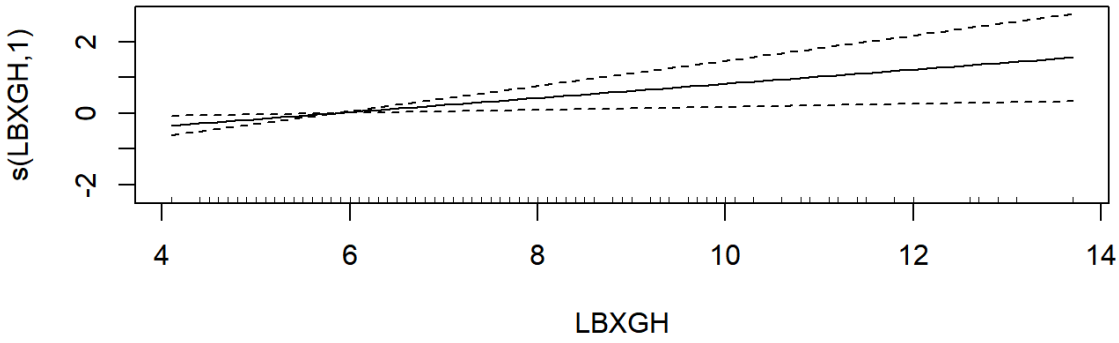
```
##                              Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)                -2.4362209  0.3485703 -6.9891819 2.764937e-12
## pfoa4Q2                     0.2388820  0.2882625  0.8286960 4.072765e-01
## pfoa4Q3                     0.6843501  0.2801903  2.4424479 1.458803e-02
## pfoa4Q4                     0.8222360  0.3054001  2.6923243 7.095592e-03
## DMDEDUCHigh school         -0.3601419  0.2745620 -1.3116958 1.896228e-01
## DMDEDUC> High school       -0.5417997  0.2543243 -2.1303500 3.314273e-02
## RIDRETH1Other Hispanic      0.5073644  0.6052753  0.8382373 4.018974e-01
## RIDRETH1non-Hispanic White  0.8335509  0.3214718  2.5929206 9.516473e-03
## RIDRETH1non-Hispanic Black  0.7901154  0.3528637  2.2391522 2.514602e-02
## RIDRETH1other               2.0227025  0.5920113  3.4166619 6.339395e-04
## RIAGENDRFemale             -0.1991081  0.2182830 -0.9121556 3.616868e-01
## smokingFormer              -0.3510489  0.2817071 -1.2461488 2.127098e-01
## smokingNever               -0.7785688  0.2883505 -2.7000775 6.932333e-03
```

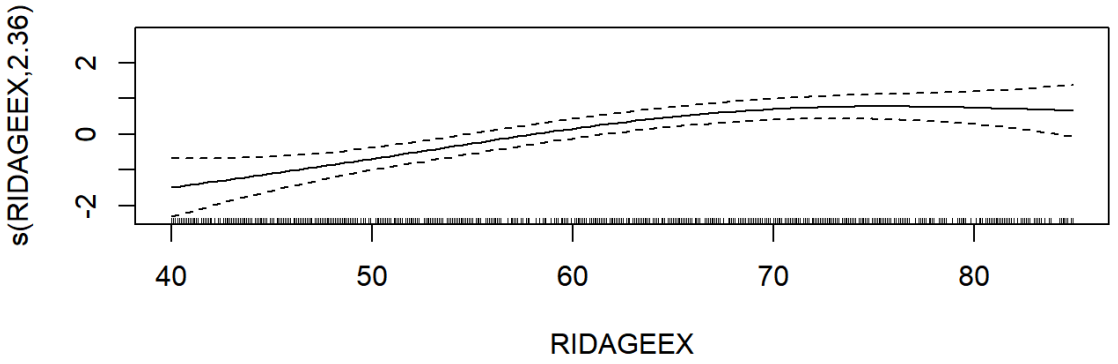
```
print(summary(pfoa1.gam)$s.table)
```

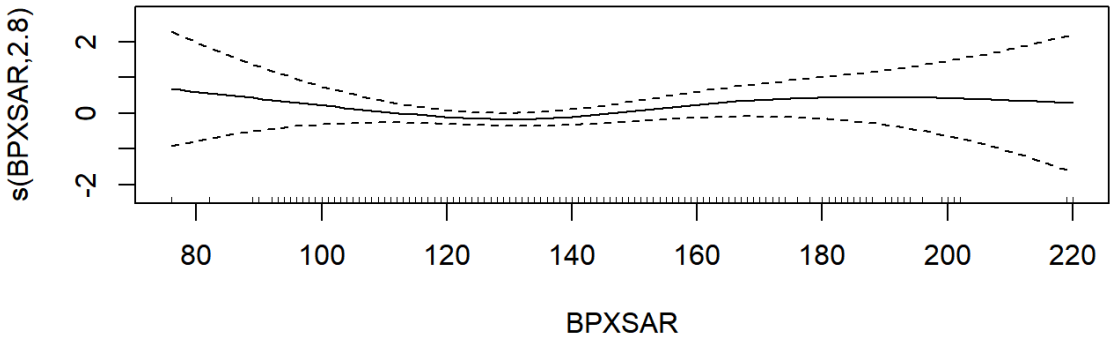
```
##                 edf   Ref.df    Chi.sq      p-value
## s(LBXGH)    1.000583 1.001165  6.477187 1.094376e-02
## s(RIDAGEEX) 2.359405 2.955123 25.956483 1.398722e-05
## s(BPXSAR)   2.800181 3.562782  4.320737 2.567554e-01
## s(BPXDAR)   1.196982 1.368057  2.240458 2.678717e-01
## s(LBXTC)    1.000098 1.000195  6.461897 1.102513e-02
```

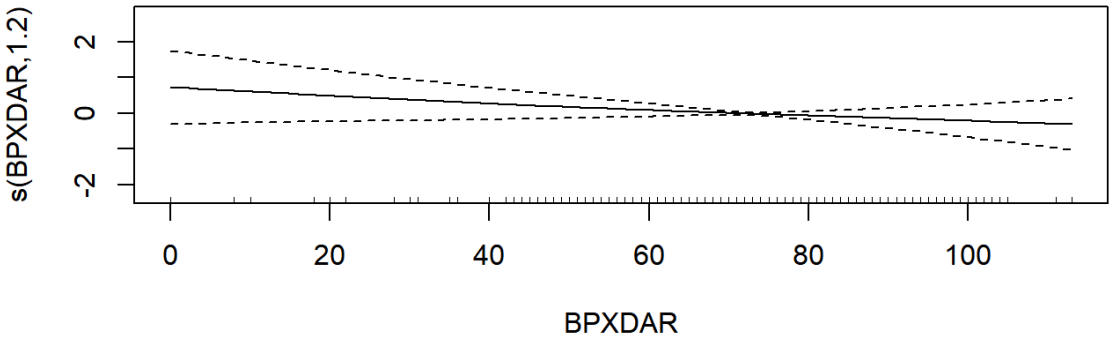
```
par(mfrow=c(2,3))
```

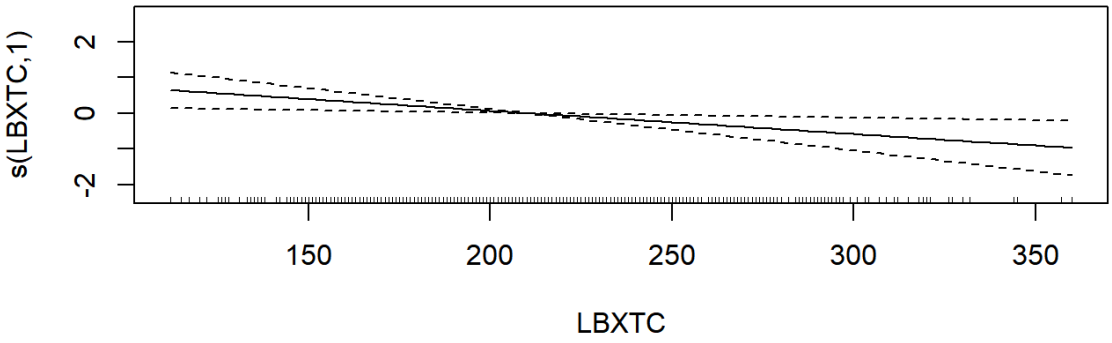
```
#GAM plot
plot(pfoa1.gam)
```


s(LBXGH,1)
-2
0
2
4
6
8
10
12
14
LBXGH


s(RIDAGEEX,2.36)
-2
0
2
40
50
60
70
80
RIDAGEEX


s(BPXSAR,2.8)
-2
0
2
80
100
120
140
160
180
200
220
BPXSAR





The summary output and plots indicate some non-linear effects in RIDAGEEX and BPXSAR. It may be useful to add quadratic terms for these variables.

Model 3: GLM model with quadratic terms

```
#GLM model with quadratic terms for RIDAGEEX and BPXSAR
pfoa1.glm2 <- glm(hadcvd ~ pfoa4 + DMDEDUC + LBXGH + RIDAGEEX + I(RIDAGEEX^2) + RIDRETH1 + RIAGENDR + BPXSAR + I(B
PXSAR^2) + BPXDAR + smoking + LBXTC, family="binomial", data=pfoa1.df)
summary(pfoa1.glm2)
```

```
##
## Call:
## glm(formula = hadcvd ~ pfoa4 + DMDEDUC + LBXGH + RIDAGEEX + I(RIDAGEEX^2) +
##     RIDRETH1 + RIAGENDR + BPXSAR + I(BPXSAR^2) + BPXDAR + smoking +
##     LBXTC, family = "binomial", data = pfoa1.df)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.021e+01  4.183e+00  -2.440 0.014692 *
## pfoa4Q2                     2.617e-01  2.879e-01   0.909 0.363433
## pfoa4Q3                     6.896e-01  2.800e-01   2.463 0.013782 *
## pfoa4Q4                     8.321e-01  3.056e-01   2.723 0.006471 **
## DMDEDUCHigh school         -3.709e-01  2.744e-01  -1.352 0.176461
## DMDEDUC> High school       -5.216e-01  2.537e-01  -2.056 0.039742 *
## LBXGH                       1.921e-01  7.791e-02   2.466 0.013665 *
## RIDAGEEX                    3.220e-01  1.056e-01   3.051 0.002281 **
## I(RIDAGEEX^2)              -2.133e-03  8.158e-04  -2.615 0.008927 **
## RIDRETH1Other Hispanic      5.090e-01  6.074e-01   0.838 0.402009
## RIDRETH1non-Hispanic White  8.382e-01  3.215e-01   2.607 0.009124 **
## RIDRETH1non-Hispanic Black  8.189e-01  3.529e-01   2.321 0.020294 *
## RIDRETH1other               2.081e+00  5.894e-01   3.530 0.000415 ***
## RIAGENDRFemale             -2.082e-01  2.180e-01  -0.955 0.339632
## BPXSAR                     -4.324e-02  3.932e-02  -1.100 0.271505
## I(BPXSAR^2)                 1.677e-04  1.348e-04   1.244 0.213493
## BPXDAR                     -1.003e-02  6.820e-03  -1.471 0.141387
## smokingFormer              -3.412e-01  2.814e-01  -1.212 0.225330
## smokingNever               -7.664e-01  2.883e-01  -2.659 0.007841 **
## LBXTC                      -6.465e-03  2.552e-03  -2.533 0.011294 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 768.96  on 999  degrees of freedom
## Residual deviance: 667.72  on 980  degrees of freedom
## AIC: 707.72
##
## Number of Fisher Scoring iterations: 6
```

```
anova(pfoa1.glm, pfoa1.glm2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: hadcvd ~ pfoa4 + DMDEDUC + LBXGH + RIDAGEEX + RIDRETH1 + RIAGENDR +
##     BPXSAR + BPXDAR + smoking + LBXTC
## Model 2: hadcvd ~ pfoa4 + DMDEDUC + LBXGH + RIDAGEEX + I(RIDAGEEX^2) +
##     RIDRETH1 + RIAGENDR + BPXSAR + I(BPXSAR^2) + BPXDAR + smoking +
##     LBXTC
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       982     676.26
## 2       980     667.72  2   8.5343  0.01402 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have evidence that adding at least one of these quadratic terms improves the model. BPXSAR is not significant, so I will use a interaction term for only RIDAGEEX.

Model 4: Full model

```
#Full model
pfoa1.full <- glm(hadcvd ~ pfoa4 + DMDEDUC + LBXGH + RIDAGEEX + I(RIDAGEEX^2) + RIDRETH1 + RIAGENDR + BPXSAR + BPX
DAR + smoking + LBXTC, family="binomial", data=pfoa1.df)
summary(pfoa1.full)
```

```
##
## Call:
## glm(formula = hadcvd ~ pfoa4 + DMDEDUC + LBXGH + RIDAGEEX + I(RIDAGEEX^2) +
##     RIDRETH1 + RIAGENDR + BPXSAR + BPXDAR + smoking + LBXTC,
##     family = "binomial", data = pfoa1.df)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.323e+01  3.422e+00  -3.866 0.000111 ***
## pfoa4Q2                     2.474e-01  2.870e-01   0.862 0.388806
## pfoa4Q3                     6.817e-01  2.793e-01   2.441 0.014643 *
## pfoa4Q4                     8.337e-01  3.051e-01   2.732 0.006287 **
## DMDEDUCHigh school         -3.748e-01  2.737e-01  -1.370 0.170840
## DMDEDUC> High school       -5.212e-01  2.531e-01  -2.060 0.039445 *
## LBXGH                       1.876e-01  7.774e-02   2.414 0.015793 *
## RIDAGEEX                    3.132e-01  1.050e-01   2.983 0.002858 **
## I(RIDAGEEX^2)              -2.071e-03  8.117e-04  -2.551 0.010726 *
## RIDRETH1Other Hispanic      4.745e-01  6.056e-01   0.783 0.433371
## RIDRETH1non-Hispanic White  8.193e-01  3.194e-01   2.565 0.010315 *
## RIDRETH1non-Hispanic Black  8.179e-01  3.518e-01   2.325 0.020058 *
## RIDRETH1other               2.113e+00  5.824e-01   3.629 0.000285 ***
## RIAGENDRFemale             -1.964e-01  2.174e-01  -0.903 0.366330
## BPXSAR                      5.197e-03  5.160e-03   1.007 0.313889
## BPXDAR                     -1.059e-02  6.783e-03  -1.561 0.118465
## smokingFormer              -3.350e-01  2.816e-01  -1.190 0.234163
## smokingNever               -7.603e-01  2.881e-01  -2.639 0.008311 **
## LBXTC                      -6.494e-03  2.551e-03  -2.546 0.010904 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 768.96  on 999  degrees of freedom
## Residual deviance: 669.21  on 981  degrees of freedom
## AIC: 707.21
##
## Number of Fisher Scoring iterations: 6
```

The full model I will use is pfoa1.full.

5. Model Selection

AIC and BIC

```
#AICc
options(na.action = "na.fail")
pfoa1.aic <- dredge(pfoa1.full)
options(na.action = "na.omit")
head(pfoa1.aic)
```

```
## Global model call: glm(formula = hadcvd ~ pfoa4 + DMDEDUC + LBXGH + RIDAGEEX + I(RIDAGEEX^2) +
##     RIDRETH1 + RIAGENDR + BPXSAR + BPXDAR + smoking + LBXTC,
##     family = "binomial", data = pfoa1.df)
## ---
## Model selection table
##      (Intrc)     BPXDA    BPXSA DMDED  LBXGH     LBXTC pfoa4 RIAGE  RIDAG
## 1981  -13.24                        + 0.2016 -0.006916     +       0.3033
## 1977  -13.71                          0.2151 -0.006806     +       0.3095
## 1982  -12.79 -0.008151             + 0.1903 -0.006652     +       0.3111
## 1978  -13.29 -0.008047               0.2041 -0.006588     +       0.3179
## 2045  -13.23                        + 0.2009 -0.006721     +     + 0.3032
## 1983  -13.51           0.002537     + 0.2027 -0.007025     +       0.3029
##        RIDAG^2 RIDRE smkng df   logLik  AICc delta weight
## 1981 -0.001929     +     + 16 -336.141 704.8  0.00  0.248
## 1977 -0.001950     +     + 14 -338.395 705.2  0.38  0.205
## 1982 -0.002020     +     + 17 -335.399 705.4  0.59  0.185
## 1978 -0.002046     +     + 15 -337.651 705.8  0.96  0.154
## 2045 -0.001928     +     + 17 -335.954 706.5  1.70  0.106
## 1983 -0.001939     +     + 17 -336.009 706.6  1.81  0.101
## Models ranked by AICc(x)
```

```
#BIC
options(na.action = "na.fail")
pfoa1.bic <- dredge(pfoa1.full, rank="BIC")
options(na.action = "na.omit")
head(pfoa1.bic)
```

```
## Global model call: glm(formula = hadcvd ~ pfoa4 + DMDEDUC + LBXGH + RIDAGEEX + I(RIDAGEEX^2) +
##     RIDRETH1 + RIAGENDR + BPXSAR + BPXDAR + smoking + LBXTC,
##     family = "binomial", data = pfoa1.df)
## ---
## Model selection table
##     (Intrc)  LBXGH     LBXTC   RIDAG    RIDAG^2 df   logLik   BIC delta weight
## 129  -5.340                 0.05344             2 -361.844 737.5  0.00  0.462
## 385 -12.060                 0.26890 -0.0016740  3 -359.344 739.4  1.91  0.178
## 257  -3.620                         0.0004021  2 -363.279 740.4  2.87  0.110
## 145  -4.367        -0.004648 0.05334             3 -359.871 740.5  2.96  0.105
## 137  -6.074 0.1355          0.05227             3 -360.033 740.8  3.29  0.089
## 401 -11.380        -0.005020 0.28110 -0.0017720  4 -357.064 741.8  4.26  0.055
## Models ranked by BIC(x)
```

Estimates of the AUCs

```
#The AIC list
options(width=66)
out=rep(0,30)
for(i in 1:30) {
  newpreds=predict(get.models(pfoa1.aic,i)[[1]],
                   newdata=pfoa2.df, type="response")
  my.roc=roc(response=pfoa2.df$hadcvd,
             predictor=newpreds, ci=TRUE)
  out[i]=my.roc$auc
}
round(out,3)
```

```
##  [1] 0.711 0.699 0.712 0.703 0.714 0.712 0.715 0.705 0.718 0.701
## [11] 0.705 0.708 0.722 0.717 0.713 0.710 0.708 0.706 0.720 0.709
## [21] 0.711 0.712 0.706 0.724 0.709 0.698 0.707 0.697 0.714 0.713
```

```
#The BIC list
options(width=66)
out=rep(0,30)
for(i in 1:30) {
  newpreds=predict(get.models(pfoa1.bic,i)[[1]],
                   newdata=pfoa2.df, type="response")
  my.roc=roc(response=pfoa2.df$hadcvd,
             predictor=newpreds, ci=TRUE)
  out[i]=my.roc$auc
}
round(out, 3)
```

```
##  [1] 0.699 0.696 0.699 0.702 0.711 0.711 0.704 0.700 0.714 0.716
## [11] 0.712 0.707 0.697 0.702 0.703 0.714 0.701 0.722 0.698 0.711
## [21] 0.713 0.715 0.702 0.707 0.701 0.726 0.712 0.713 0.711 0.695
```

24th model (0.724) from the AICc list and 26th model (0.726) from the BIC list have the highest AUC values in each list.

```
#Model from the AICc list
model24 <- get.models(pfoa1.aic, 24)[[1]]
summary(model24)
```

```
##
## Call:
## glm(formula = hadcvd ~ BPXDAR + DMDEDUC + LBXGH + LBXTC + RIDAGEEX +
##     I(RIDAGEEX^2) + RIDRETH1 + smoking + 1, family = "binomial",
##     data = pfoa1.df)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.226e+01  3.358e+00  -3.652 0.000261
## BPXDAR                     -8.206e-03  6.578e-03  -1.247 0.212219
## DMDEDUCHigh school         -3.842e-01  2.702e-01  -1.422 0.155119
## DMDEDUC> High school       -4.762e-01  2.508e-01  -1.898 0.057643
## LBXGH                       1.517e-01  7.595e-02   1.997 0.045816
## LBXTC                      -5.598e-03  2.487e-03  -2.251 0.024394
## RIDAGEEX                    3.039e-01  1.045e-01   2.909 0.003625
## I(RIDAGEEX^2)              -1.980e-03  8.062e-04  -2.456 0.014056
## RIDRETH10ther Hispanic      5.907e-01  5.960e-01   0.991 0.321613
## RIDRETH1non-Hispanic White  9.365e-01  3.079e-01   3.042 0.002354
## RIDRETH1non-Hispanic Black  9.091e-01  3.460e-01   2.627 0.008609
## RIDRETH1other               2.192e+00  5.751e-01   3.811 0.000139
## smokingFormer              -2.991e-01  2.765e-01  -1.082 0.279461
## smokingNever               -7.558e-01  2.806e-01  -2.693 0.007076
##
## (Intercept)                ***
## BPXDAR
## DMDEDUCHigh school
## DMDEDUC> High school       .
## LBXGH                      *
## LBXTC                      *
## RIDAGEEX                   **
## I(RIDAGEEX^2)              *
## RIDRETH10ther Hispanic
## RIDRETH1non-Hispanic White **
## RIDRETH1non-Hispanic Black **
## RIDRETH1other              ***
## smokingFormer
## smokingNever               **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 768.96  on 999  degrees of freedom
## Residual deviance: 680.72  on 986  degrees of freedom
## AIC: 708.72
##
## Number of Fisher Scoring iterations: 6
```

```
#Model from the BIC list
model26 <- get.models(pfoa1.bic, 26)[[1]]
summary(model26)
```

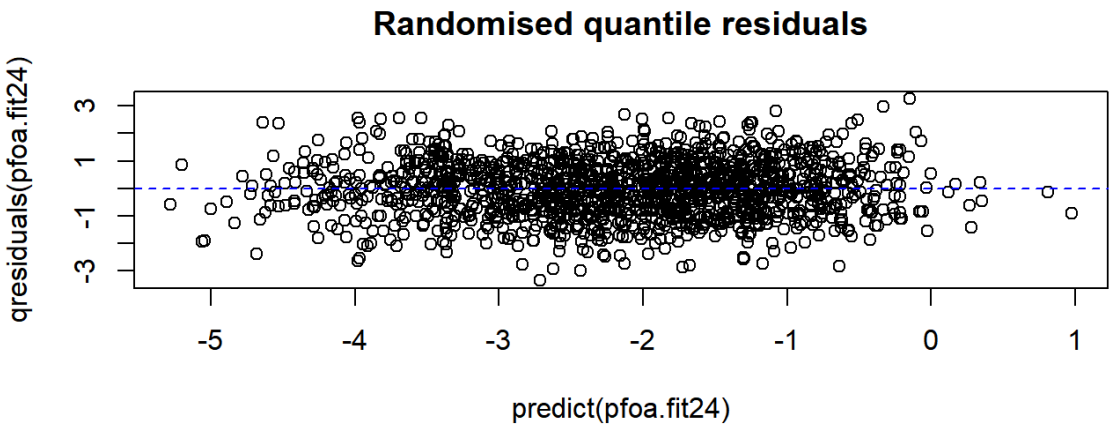
```
##
## Call:
## glm(formula = hadcvd ~ LBXGH + RIAGENDR + RIDAGEEX + 1, family = "binomial",
##     data = pfoa1.df)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.927885   0.698006  -8.493  < 2e-16 ***
## LBXGH           0.134833   0.067891   1.986    0.047 *
## RIAGENDRFemale -0.274599   0.194424  -1.412    0.158
## RIDAGEEX        0.052123   0.008473   6.152 7.65e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 768.96  on 999  degrees of freedom
## Residual deviance: 718.06  on 996  degrees of freedom
## AIC: 726.06
##
## Number of Fisher Scoring iterations: 5
```

Final model

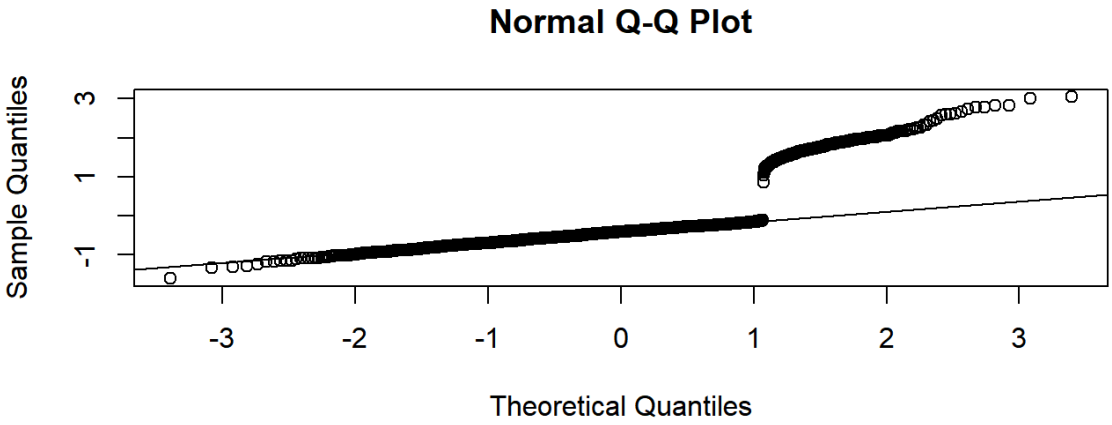
```
#Model 24 from the AICc list
pfoa.fit24 <- glm(hadcvd ~ BPXDAR + I(BPXSAR^2) + DMDEDUC + LBXGH +
    LBXTC + pfoa4 + RIAGENDR + RIDAGEEX + I(RIDAGEEX^2) + RIDRETH1 +
    smoking, family = "binomial",
    data = pfoa.new)
summary(pfoa.fit24)
```

```
##
## Call:
## glm(formula = hadcvd ~ BPXDAR + I(BPXSAR^2) + DMDEDUC + LBXGH +
##     LBXTC + pfoa4 + RIAGENDR + RIDAGEEX + I(RIDAGEEX^2) + RIDRETH1 +
##     smoking, family = "binomial", data = pfoa.new)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.005e+01  2.603e+00  -3.862 0.000113
## BPXDAR                     -1.289e-02  5.641e-03  -2.285 0.022311
## I(BPXSAR^2)                 3.079e-05  1.361e-05   2.262 0.023694
## DMDEDUCHigh school         -2.459e-01  2.160e-01  -1.138 0.254991
## DMDEDUC> High school       -6.406e-01  2.083e-01  -3.075 0.002105
## LBXGH                       1.621e-01  6.156e-02   2.633 0.008458
## LBXTC                      -4.262e-03  1.964e-03  -2.171 0.029964
## pfoa4Q2                     3.536e-01  2.185e-01   1.619 0.105533
## pfoa4Q3                     5.939e-01  2.207e-01   2.691 0.007122
## pfoa4Q4                     5.407e-01  2.472e-01   2.188 0.028703
## RIAGENDRFemale             -4.054e-01  1.734e-01  -2.338 0.019382
## RIDAGEEX                    2.231e-01  8.056e-02   2.770 0.005614
## I(RIDAGEEX^2)              -1.419e-03  6.241e-04  -2.274 0.022948
## RIDRETH1Other Hispanic     -8.830e-02  5.128e-01  -0.172 0.863295
## RIDRETH1non-Hispanic White  6.996e-01  2.389e-01   2.929 0.003405
## RIDRETH1non-Hispanic Black  5.703e-01  2.697e-01   2.114 0.034486
## RIDRETH1other               1.454e+00  4.637e-01   3.135 0.001720
## smokingFormer              -8.136e-03  2.363e-01  -0.034 0.972529
## smokingNever               -3.901e-01  2.402e-01  -1.624 0.104327
##
## (Intercept)                ***
## BPXDAR                     *
## I(BPXSAR^2)                *
## DMDEDUCHigh school
## DMDEDUC> High school       **
## LBXGH                      **
## LBXTC                      *
## pfoa4Q2
## pfoa4Q3                    **
## pfoa4Q4                    *
## RIAGENDRFemale             *
## RIDAGEEX                   **
## I(RIDAGEEX^2)              *
## RIDRETH1Other Hispanic
## RIDRETH1non-Hispanic White **
## RIDRETH1non-Hispanic Black *
## RIDRETH1other              **
## smokingFormer
## smokingNever
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1190.7  on 1455  degrees of freedom
## Residual deviance: 1047.7  on 1437  degrees of freedom
## AIC: 1085.7
##
## Number of Fisher Scoring iterations: 6
```

```
#Goodness of fit
plot(predict(pfoa.fit24), qresiduals(pfoa.fit24), main="Randomised quantile residuals")
abline(h=0, lty="dashed", col='blue')
```




```
qqnorm(residuals(pfoa.fit24))
qqline(residuals(pfoa.fit24))
```



26th model from the BIC list has slightly higher value of AUC than 24th model from the AICc list. But, this model is too simple and has higher AIC value. So, pfoa.fit24 will be the final model.

Confusion matrix

```
#Confusion matrix
table(actual=pfoa.new$hadcvd, pred=round(fitted(pfoa.fit24)))
```

```
##        pred
## actual     0    1
##   FALSE 1244    5
##   TRUE   203    4
```

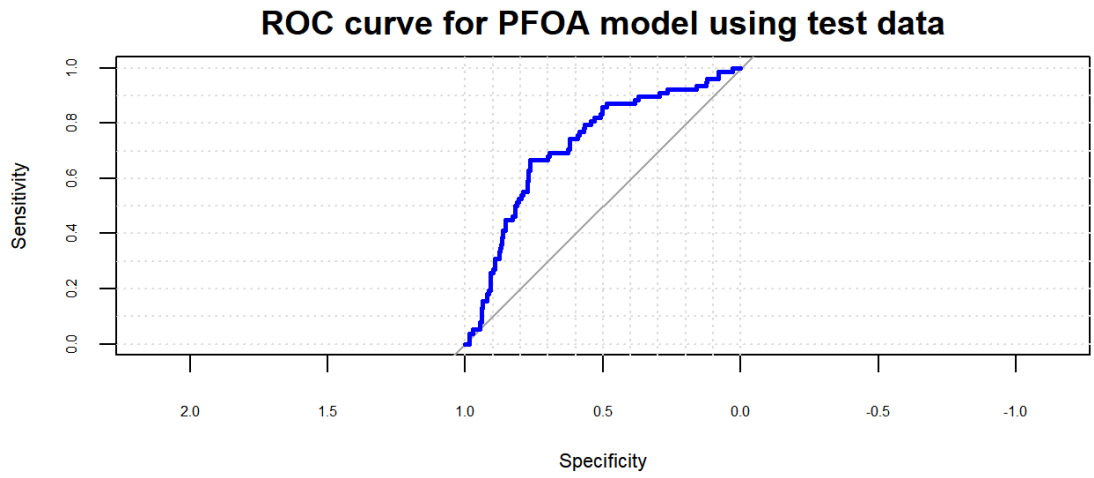
The estimated specificity is very high, but the estimated sensitivity is very low. Estimated prediction error is 0.167. The model is good at detecting those who don't have heart disease, but performs bad in detecting those who have heart disease.

Interpretation of coefficients

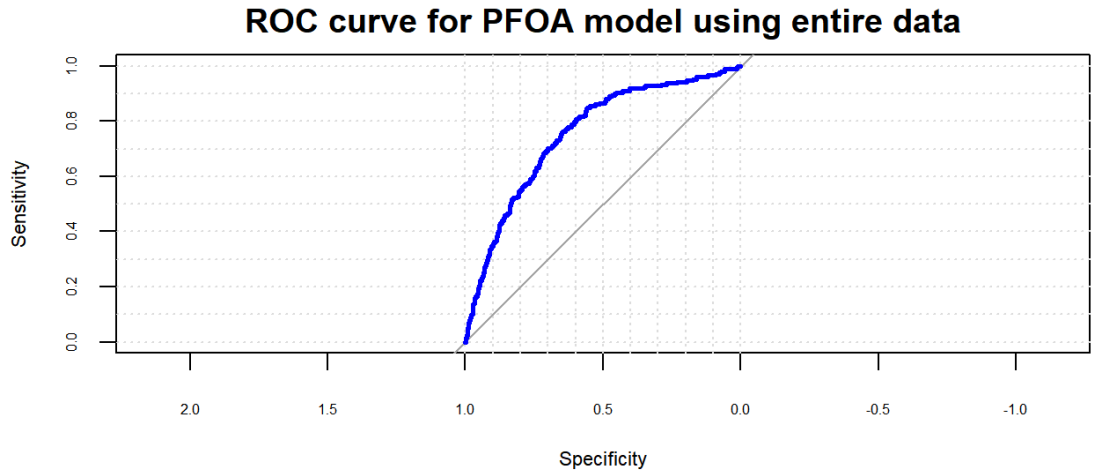
- According to this model, we have some evidence that the increase in diastolic blood pressure decreases the probability of having heart disease. Holding other variables as constant, we estimate that, for every 1 unit increase in diastolic blood pressure, the odds of having heart disease is multiplied by about 0.99. The model also suggests that, at lower values, the increase in systolic blood pressure increases the odds of having heart disease, but at higher values, the odds of having heart disease decreases as systolic blood pressure increases.
- We have some evidence that high level of education decreases the probability of having heart disease. Holding other variables as constant, we estimate that, the odds of patients whose highest education completed is more than high school having heart disease is about 0.53 times those of patients whose highest education completed is less than high school.
- We have some evidence that an increase in blood concentration of glycosylated haemoglobin increases the probability of having heart disease. Holding other variables as constant, we estimate that, for every 1 unit increase in glycosylated haemoglobin, the odds of having heart disease is multiplied by about 1.18.
- We have some evidence that an increase in blood cholesterol levels decreases the probability of having heart disease. Holding other variables as constant, we estimate that, for every 1 unit increase in cholestrol levels, the odds of having heart disease is multiplied by about 0.10.
- There is some evidence to suggest that higher blood concentration level of PFOA increases the probability of having heart disease. Holding other variables as constant, we estimate that the odds of a patient whose level of PFOA is Q3 having heart disease is about 1.81 times those of a patient whose level of PFOA is Q1. For a patient whose level of PFOA is Q4, we estimate that the odds of having heart disease is about 1.72 times those of a patient whose level of PFOA is Q1.
- There is some evidence that the probability of having heart disease decreases when a patient is female. Holding other variables as constant, we estimate that the odds of a female patient having heart disease is about 0.67 times those of a male patient.
- There is evidence that the probability of having heart disease increases as a patient ages. Holding other variables as constant, we estimate that the odds of having heart disease is multiplied by 1.25 every 1 year increase in age. But, at some point of age, the odds start to decrease.
- There is evidence that the probability of having heart disease increases when a patient is non-hispanic. Holding other variables as constant, we estimate that the odds of a non-hispanic white patient having heart disease is about 2.01 times those of a mexican hispanic patient. For a non-hispanic black patient, we estimate that the odds of having heart disease is about 1.77 times those of a mexican hispanic patient. For a patient in other racial/ethnic group, we estimate that the odds of having heart disease is about 4.28 times those of a mexican hispanic patient.
- There is no evidence that smoking affects the probability of having heart disease.

ROC curve showing the performance of the model

```
#ROC curve using test data
preds <- predict(model24, newdata=pfoa2.df, type="response")
test.roc <- roc(response=pfoa2.df$hadcvd, predictor=preds)
plot(test.roc, col="blue", grid=TRUE, lwd=2.5, cex.lab=0.7, cex.axis=0.5, main="ROC curve for PFOA model using tes
t data")
```




```
#refit the model using the entire dataset
pfoa.roc <- roc(response=pfoa.new$hadcvd,
                predictor=fitted.values(pfoa.fit24))
plot(pfoa.roc, col="blue", grid=TRUE, lwd=2.5, cex.lab=0.7, cex.axis=0.5, main="ROC curve for PFOA model using ent
ire data")
```



The area under the ROC curve is greater than 0.5 for both test data and entire dataset. The model seems to have good performance in predicting the probability of heart disease.

6. Model Building: Part 2

To estimate the total effect of pfoa4 on hadcvd, we need to include pfoa4, education(DMDEDUC), ethnicity(RIDRETH1), age(RIDAGEEX), gender(RIAGENDR) and smoking.

Blood pressure(BPXSAR, BPXDAR), cholesterol(LBXTC) and diabetes(LBXGH) should be excluded because they are on indirect causal pathways from pfoa4 to hadcvd.

All confounding pathways should be closed, so we need to include education(DMDEDUC), ethnicity(RIDRETH1) and age(RIDAGEEX) which are confounders for pfoa4.

We also need to include age(RIDAGEEX), smoking and gender(RIAGENDR) as they have direct effects on hadcvd.

Fitting a GLM model using these variables

```
pfoa.glm2 <- glm(hadcvd ~ pfoa4 + DMDEDUC + RIDRETH1 + RIDAGEEX + smoking + RIAGENDR, family="binomial", data=pfo
a.new)
summary(pfoa.glm2)
```

```
##
## Call:
## glm(formula = hadcvd ~ pfoa4 + DMDEDUC + RIDRETH1 + RIDAGEEX +
##     smoking + RIAGENDR, family = "binomial", data = pfoa.new)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -5.292660   0.519633 -10.185  < 2e-16
## pfoa4Q2                     0.282694   0.214696   1.317   0.1879
## pfoa4Q3                     0.492916   0.215775   2.284   0.0223
## pfoa4Q4                     0.412096   0.240440   1.714   0.0865
## DMDEDUCHigh school         -0.274827   0.212296  -1.295   0.1955
## DMDEDUC> High school       -0.644113   0.205224  -3.139   0.0017
## RIDRETH1Other Hispanic     -0.211536   0.509763  -0.415   0.6782
## RIDRETH1non-Hispanic White  0.512212   0.229981   2.227   0.0259
## RIDRETH1non-Hispanic Black  0.521207   0.263646   1.977   0.0481
## RIDRETH1other               1.433545   0.454285   3.156   0.0016
## RIDAGEEX                    0.053664   0.007378   7.274 3.49e-13
## smokingFormer               0.006849   0.232724   0.029   0.9765
## smokingNever               -0.417742   0.236499  -1.766   0.0773
## RIAGENDRFemale             -0.349141   0.166541  -2.096   0.0360
##
## (Intercept)                ***
## pfoa4Q2
## pfoa4Q3                    *
## pfoa4Q4                    .
## DMDEDUCHigh school
## DMDEDUC> High school       **
## RIDRETH1Other Hispanic
## RIDRETH1non-Hispanic White *
## RIDRETH1non-Hispanic Black *
## RIDRETH1other              **
## RIDAGEEX                   ***
## smokingFormer
## smokingNever               .
## RIAGENDRFemale             *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1190.7  on 1455  degrees of freedom
## Residual deviance: 1073.2  on 1442  degrees of freedom
## AIC: 1101.2
##
## Number of Fisher Scoring iterations: 5
```

RIDAGEEX is very statistically significant so we should not drop it. There are some evidence that suggest keeping DMDEDUC, RIDRETH1 and RIAGENDR in the model.

Fitting a GLM model using these variables

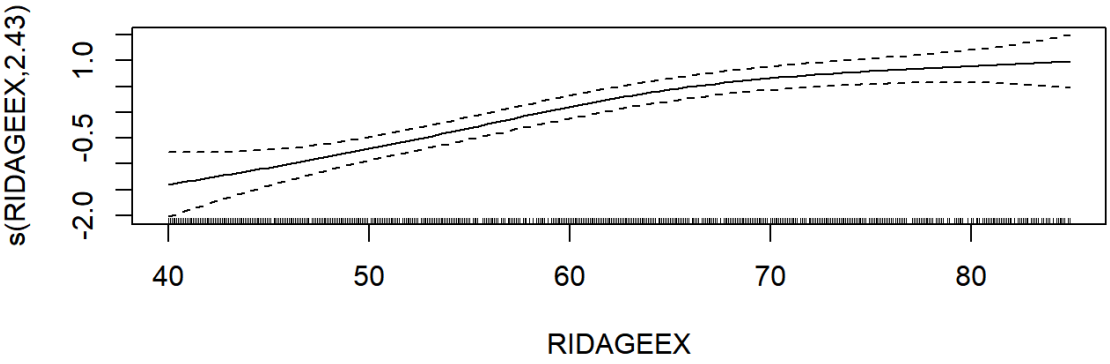
```
pfoa.gam2 <- gam(hadcvd ~ pfoa4 + DMDEDUC + RIDRETH1 + s(RIDAGEEX) + smoking + RIAGENDR, family="binomial", data=pfoa.new)
print(summary(pfoa.gam2)$p.table)
```

```
##                               Estimate Std. Error     z value
## (Intercept)                -2.090075816  0.2738827 -7.63128165
## pfoa4Q2                     0.300628729  0.2147377  1.39998137
## pfoa4Q3                     0.507015733  0.2158655  2.34875708
## pfoa4Q4                     0.419542885  0.2407692  1.74251068
## DMDEDUCHigh school         -0.266754487  0.2123312 -1.25631328
## DMDEDUC> High school       -0.640055701  0.2052720 -3.11808541
## RIDRETH1Other Hispanic     -0.197018703  0.5104650 -0.38595930
## RIDRETH1non-Hispanic White  0.562616368  0.2308524  2.43712598
## RIDRETH1non-Hispanic Black  0.541482570  0.2645928  2.04647537
## RIDRETH1other               1.468838311  0.4567516  3.21583646
## smokingFormer              -0.009143161  0.2331956 -0.03920812
## smokingNever               -0.407108267  0.2373401 -1.71529527
## RIAGENDRFemale             -0.356210362  0.1666264 -2.13777917
##                                  Pr(>|z|)
## (Intercept)                2.324313e-14
## pfoa4Q2                    1.615189e-01
## pfoa4Q3                    1.883619e-02
## pfoa4Q4                    8.141912e-02
## DMDEDUCHigh school         2.090024e-01
## DMDEDUC> High school       1.820300e-03
## RIDRETH1Other Hispanic     6.995268e-01
## RIDRETH1non-Hispanic White 1.480452e-02
## RIDRETH1non-Hispanic Black 4.070962e-02
## RIDRETH1other              1.300649e-03
## smokingFormer              9.687245e-01
## smokingNever               8.629110e-02
## RIAGENDRFemale             3.253467e-02
```

```
print(summary(pfoa.gam2)$s.table)
```

```
##                  edf   Ref.df   Chi.sq p-value
## s(RIDAGEEX) 2.432501 3.042903 52.36661       0
```

```
plot(pfoa.gam2)
```



RIDAGEEX has a edf value greater than 2 and there is some indication of non-linearity in the plot. I will consider adding a quadratic term for RIDAGEEX.

Fitting a GLM model with quadratic terms

```
pfoa.glm3 <- glm(hadcvd ~ pfoa4 + DMDEDUC + RIDRETH1 + RIDAGEEX + I(RIDAGEEX^2) + smoking + RIAGENDR, family="bino
mial", data=pfoa.new)
anova(pfoa.glm2, pfoa.glm3, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: hadcvd ~ pfoa4 + DMDEDUC + RIDRETH1 + RIDAGEEX + smoking + RIAGENDR
## Model 2: hadcvd ~ pfoa4 + DMDEDUC + RIDRETH1 + RIDAGEEX + I(RIDAGEEX^2) +
##     smoking + RIAGENDR
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1442     1073.2
## 2      1441     1068.1  1   5.0773  0.02424 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have some evidence to include the quadratic term for RIDAGEEX.

Interaction terms

Now, I will test whether adding any interaction term improves the model.

```
pfoa.glm4 <- glm(hadcvd ~ pfoa4 + DMDEDUC * RIDRETH1 * (RIDAGEEX + I(RIDAGEEX^2)) * smoking * RIAGENDR, family="bi
nomial", data=pfoa.new)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1
## occurred
```

```
anova(pfoa.glm2, pfoa.glm4, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: hadcvd ~ pfoa4 + DMDEDUC + RIDRETH1 + RIDAGEEX + smoking + RIAGENDR
## Model 2: hadcvd ~ pfoa4 + DMDEDUC * RIDRETH1 * (RIDAGEEX + I(RIDAGEEX^2)) *
##     smoking * RIAGENDR
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1      1442     1073.2
## 2      1232    21193.7 210   -20121
```

There is no evidence to suggest that including any interaction term would improve the model.

7. Final model

```
summary(pfoa.glm3)
```

```
##
## Call:
## glm(formula = hadcvd ~ pfoa4 + DMDEDUC + RIDRETH1 + RIDAGEEX +
##     I(RIDAGEEX^2) + smoking + RIAGENDR, family = "binomial",
##     data = pfoa.new)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.070e+01  2.551e+00  -4.195 2.72e-05
## pfoa4Q2                     3.034e-01  2.147e-01   1.413  0.15758
## pfoa4Q3                     5.060e-01  2.157e-01   2.346  0.01900
## pfoa4Q4                     4.234e-01  2.406e-01   1.760  0.07848
## DMDEDUCHigh school         -2.681e-01  2.125e-01  -1.262  0.20692
## DMDEDUC> High school       -6.410e-01  2.053e-01  -3.123  0.00179
## RIDRETH1Other Hispanic     -1.924e-01  5.106e-01  -0.377  0.70631
## RIDRETH1non-Hispanic White  5.578e-01  2.308e-01   2.417  0.01566
## RIDRETH1non-Hispanic Black  5.426e-01  2.646e-01   2.051  0.04029
## RIDRETH1other               1.463e+00  4.573e-01   3.200  0.00137
## RIDAGEEX                    2.266e-01  7.945e-02   2.852  0.00435
## I(RIDAGEEX^2)              -1.347e-03  6.129e-04  -2.197  0.02802
## smokingFormer              -9.110e-03  2.331e-01  -0.039  0.96882
## smokingNever               -4.045e-01  2.372e-01  -1.705  0.08814
## RIAGENDRFemale             -3.576e-01  1.666e-01  -2.147  0.03180
##
## (Intercept)                ***
## pfoa4Q2
## pfoa4Q3                    *
## pfoa4Q4                    .
## DMDEDUCHigh school
## DMDEDUC> High school       **
## RIDRETH1Other Hispanic
## RIDRETH1non-Hispanic White *
## RIDRETH1non-Hispanic Black *
## RIDRETH1other              **
## RIDAGEEX                   **
## I(RIDAGEEX^2)              *
## smokingFormer
## smokingNever               .
## RIAGENDRFemale             *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1190.7  on 1455  degrees of freedom
## Residual deviance: 1068.1  on 1441  degrees of freedom
## AIC: 1098.1
##
## Number of Fisher Scoring iterations: 6
```

According to this model, there is some evidence to suggest that higher blood concentration level of PFOA increases the probability of having heart disease.

- We do not have evidence that a patient whose level of PFOA is Q2 have higher probability of having heart disease than a patient whose level of PFOA is Q1.
- However, we have some evidence that the odds of a patient whose level of PFOA is Q3 having heart disease is about 1.66 times those of a patient whose level of PFOA is Q1.

- For a patient whose level of PFOA is Q4, we have moderate evidence that the odds of having heart disease is about 1.53 times those of a patient whose level of PFOA is Q1.
- The probability of having heart disease seems to be highest when a patient's level of PFOA is Q3.

Confusion Matrix

```
predicted <- predict(pfoa.glm3, newdata = pfoa2.df, type="response")
predicted.f <- ifelse(predicted > 0.4, TRUE, FALSE)
confusion_matrix <- table(Predicted = predicted.f, Actual = pfoa2.df$hadcvd)
print(confusion_matrix)
```

```
##          Actual
## Predicted FALSE TRUE
##     FALSE   372   76
##     TRUE      6    2
```

```
# Calculate the accuracy
accuracy <- (2 + 372) / (2 + 372 + 76 + 6)
accuracy
```

```
## [1] 0.8201754
```

8. Conclusion

Our final model is pfoa.glm3 with 82% accuracy.

We were interested in building a model to predict the probability of heart disease using variables, such as perperfluorooctanoic acid(PFOA) and a range of health, nutrition, and biochemical variables.

we conclude that our final model is GLM model with quadratic terms. According to this model, there is some evidence to suggest that higher blood concentration level of PFOA increases the probability of having heart disease.

- We do not have evidence that a patient whose level of PFOA is Q2 have higher probability of having heart disease than a patient whose level of PFOA is Q1.
- However, we have some evidence that the odds of a patient whose level of PFOA is Q3 having heart disease is about 1.66 times those of a patient whose level of PFOA is Q1.
- For a patient whose level of PFOA is Q4, we have moderate evidence that the odds of having heart disease is about 1.53 times those of a patient whose level of PFOA is Q1.
- The probability of having heart disease seems to be highest when a patient's level of PFOA is Q3.