Assignment 15; STAT 689

Philip Anderson; panders2@tamu.edu

4/21/2018

```
rm(list=ls())
```

```
# bring in data
sim <- read.csv("/Users/panders2/Documents/schools/tamu/stat_689/homework/semiparametric-regression/mis
names(sim) <- tolower(names(sim))
str(sim)
```

```
## 'data.frame':    446 obs. of  11 variables:
##  $ id       : int  1 1 2 2 3 3 4 4 5 5 ...
##  $ meas     : int  1 2 1 2 1 2 1 2 1 2 ...
##  $ age      : int  49 49 62 62 46 46 51 51 69 69 ...
##  $ bmi      : num  31.3 31.3 21 21 19.1 ...
##  $ truth    : num  27.9 27.9 23.1 23.1 26.5 ...
##  $ ffq      : num  36.5 46.5 26.5 20.9 23.5 ...
##  $ recall   : num  30.5 38.8 25.2 16.2 23.3 ...
##  $ bio      : num  26.3 20.5 21.9 17.6 29.6 ...
##  $ avgffq   : num  41.5 41.5 23.7 23.7 25.6 ...
##  $ avgrecall: num  34.7 34.7 20.7 20.7 23.2 ...
##  $ avgbio   : num  23.4 23.4 19.8 19.8 31.3 ...
```

Question 1

Run a random-intercept logistic spline regression with Y=indicator that Bio < 27.5, X=FFQ (spline), and Z=(Age, BMI) (linear).

```
# generate a binary class for Biomarkers
sim$bio_bin <- ifelse(sim$bio < 27.5, 1, 0)
# check
summary(sim[sim$bio_bin==1, ]$bio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.44   20.25   22.84   22.49   25.75   27.34
```

```
summary(sim[sim$bio_bin==0, ]$bio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.61   29.62   32.08   33.79   36.45   52.90
```

```
mod_one <- mgcv::gamm(bio_bin ~ s(ffq) + age + bmi
                      , random=list(id = ~ 1)
                      , family=binomial
                      , data=sim
                      )
```

```
##
##  Maximum number of PQL iterations:  20
```

```
## iteration 1
```

```
## iteration 2
## iteration 3

## iteration 4

## iteration 5
```

Question 2

Which among X and Z are statistically significant predictors?

```
summary(mod_one$gam)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## bio_bin ~ s(ffq) + age + bmi
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.0080984  0.9210452  -1.095    0.274
## age         -0.0003272  0.0140430  -0.023    0.981
## bmi          0.0409440  0.0221343   1.850    0.065 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##        edf Ref.df    F p-value
## s(ffq)   1      1 9.68 0.00198 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0317
##   Scale est. = 1          n = 446
```

Of the included predictor variables, only X, or the smoothed FFQ term is statistically significant.

Question 3

Graph the fitted probabilities for people who are 55 years old and whose BMI = 25.

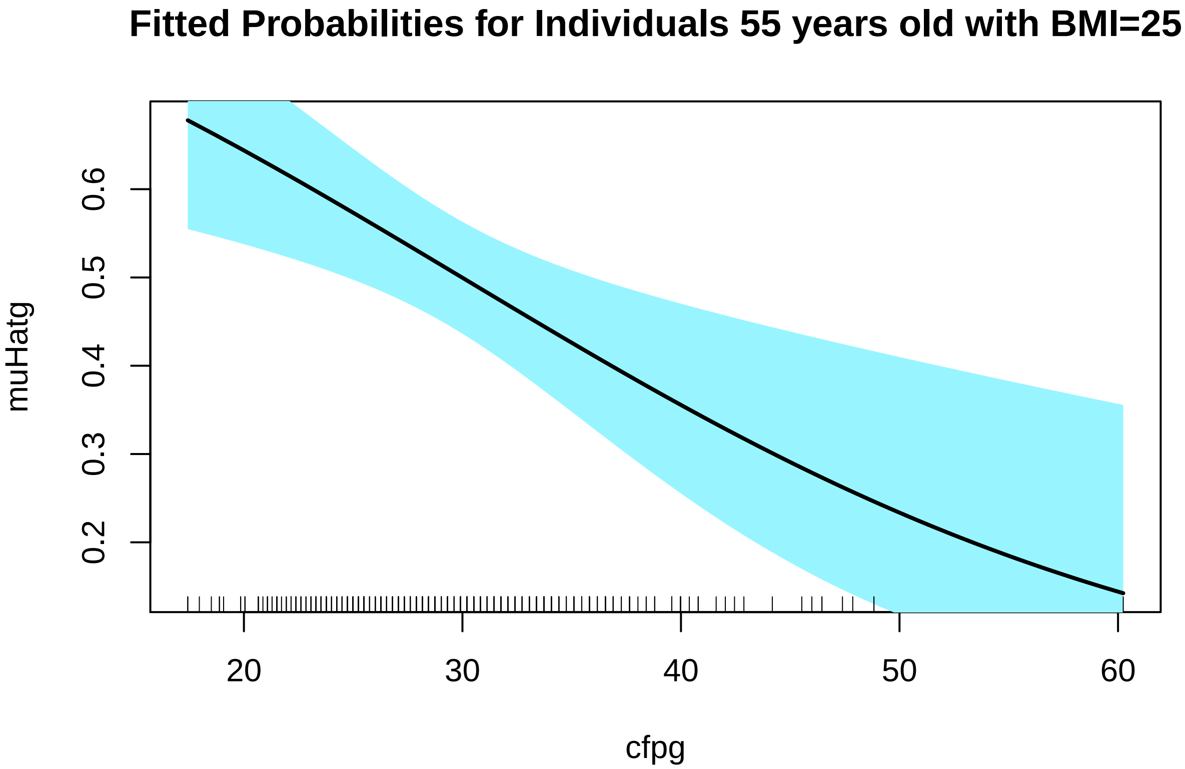
```
ng <- 101
cfpg <- seq(from=min(sim$ffq)
            , to=max(sim$ffq)
            , length=ng
              )
newData <- as.data.frame(cbind(cfpg
              , rep(55, ng)
              , rep(25, ng)
                               )
                        )
names(newData) <- c("ffq", "age", "bmi")
```

```
newDataList <- as.list(newData)

predObj <- predict(mod_one$gam, newdata=newDataList, se=T)

muHatg <- 1/(1+exp(-predObj$fit))
aa      <- predObj$fit + 2*predObj$se
bb      <- predObj$fit - 2*predObj$se
lowergg <- 1 / (1 + exp(-bb))
uppergg <- 1 / (1 + exp(-aa))
```

```
plot(cfpg, muHatg, type="n", main="Fitted Probabilities for Individuals 55 years old with BMI=25")
polygon(c(cfpg, rev(cfpg))
        , c(lowergg, rev(uppergg))
        , col="cadetblue1"
        , border=F
        )
lines(cfpg, muHatg, lwd=2)
rug(sim$ffq)
```


Fitted Probabilities for Individuals 55 years old with BMI=25
muHatg
cfpg