Assignment 16; STAT 689

Philip Anderson; panders2@tamu.edu

4/30/2018

```r
# 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 ...
```

```r
# response
y <- sim$bio
# smoothed
x1 <- sim$avgffq
x2 <- sim$avgrecall
# linear
z1 <- sim$age
z2 <- sim$bmi
```

Question 1

Fit a quantile regression with tau=0.5 for the regression of Y on X1 and Z.

```r
mod_one <- quantreg::rqss(y ~ qss(x1, lambda=3.5) + z1 + z2
                          , tau=0.5 # modeling the median
                          )
```

Question 1A

What is statistically significant?

```r
summary(mod_one)
```

```
## Formula:
## y ~ qss(x1, lambda = 3.5) + z1 + z2
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.21308    4.93126   5.316  1.7e-07 ***
## z1          -0.02296    0.06553  -0.350   0.7262
## z2          -0.18223    0.09419  -1.935   0.0537 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of qss terms:
##    EDF Lambda Penalty F value Pr(>F)
## x1   7    3.5    5.98   45.51 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##   Quantile Fidelity at tau = 0.5  is      1196.44
##   Effective Degrees of Freedom = 11        Sample Size = 446
```

The smoothed x1 term, or avgFFQ is the only significant term at the 5% level.

Question 1B

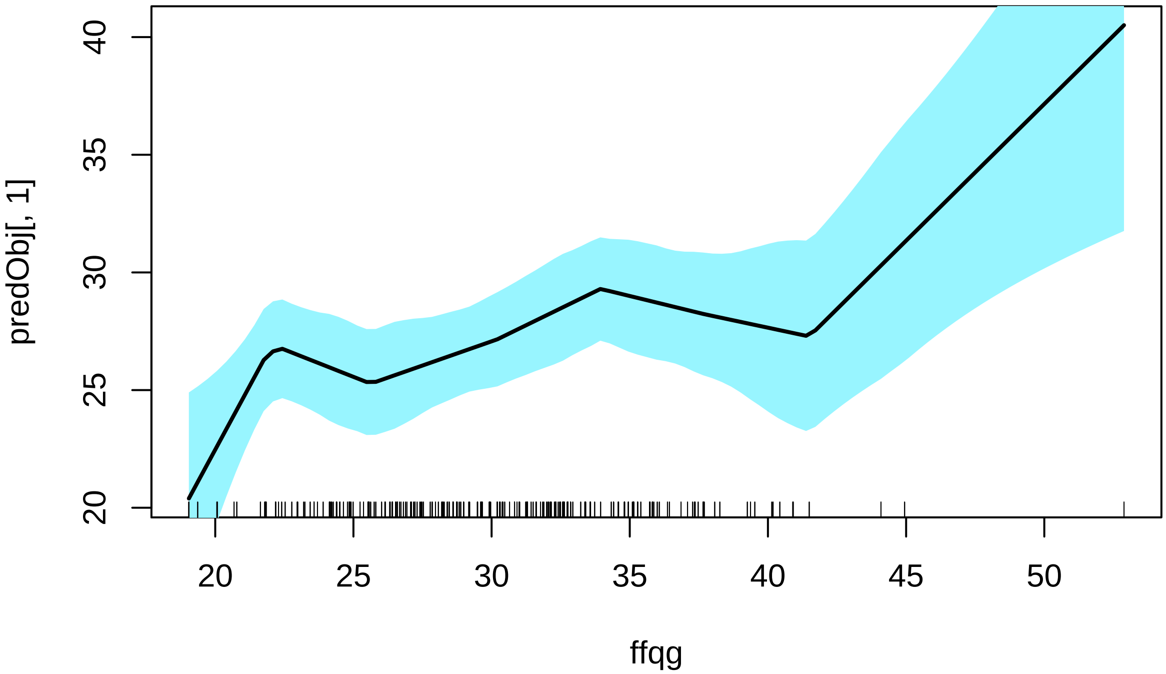
Plot the fitted function when age=55 and BMI=25.

```r
ng <- 101
ffqg <- seq(from=min(x1)
            , to=max(x1)
            , length=ng
            )
newData <- as.data.frame(cbind(ffqg
              , rep(55, ng)
              , rep(25, ng)
                    )
              )
names(newData) <- c("x1", "z1", "z2")

newDataList <- as.list(newData)

predObj <- quantreg::predict.rqss(object=mod_one
                   , newdata=newDataList
                   , interval="confidence"
                   , level=0.95
                   )
```

```r
plot(ffqg, predObj[,1], type="n")
polygon(c(ffqg, rev(ffqg))
        , c(predObj[,2], rev(predObj[,3]))
        , col="cadetblue1"
        , border=F
        )
lines(ffqg, predObj[,1], lwd=2)
rug(x1)
```



Question 2

Fit an ordinary regression, but apply a bivariate spline, te(x1, x2).

```
mod_two <- mgcv::gam(y ~ te(x1, x2) + z1 + z2
                     , method="REML"
                     )
```

Question 2A

What is statistically significant?

```
summary(mod_two)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## y ~ te(x1, x2) + z1 + z2
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.68149    2.80051  10.956  < 2e-16 ***
## z1           0.04080    0.04291   0.951  0.34218
## z2          -0.19682    0.06736  -2.922  0.00366 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df     F  p-value
## te(x1,x2) 3.001  3.003 16.41 3.69e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.108   Deviance explained = 11.8%
## -REML = 1493.9  Scale est. = 48.741    n = 446
```

Of the linear terms, only z2 (BMI) is significant; the bivariate smoothed term of x1 and x2 is also significant.

Question 2B

Plot the functions when age=55 and BMI=25.

X1 materials first.

```
ng <- 101
x1g <- seq(from=min(x1), to=max(x1), length=ng)

newDataDF <- as.data.frame(
          cbind(
          x1g
          , rep(mean(x2), ng)
          , rep(55, ng)
          , rep(25, ng)
          )
  )
names(newDataDF) <- c("x1", "x2", "z1", "z2")
newDataList <- as.list(newDataDF)

predObj_one <- predict(mod_two, newDataList, se=T)

muHatg_one <- predObj_one$fit
aa_one     <- predObj_one$fit + 2*predObj_one$se
bb_one     <- predObj_one$fit - 2*predObj_one$se
lowergg_one <- 1 / (1 + exp(-bb_one))
uppergg_one <- 1 / (1 + exp(-aa_one))
```

X2 materials second.

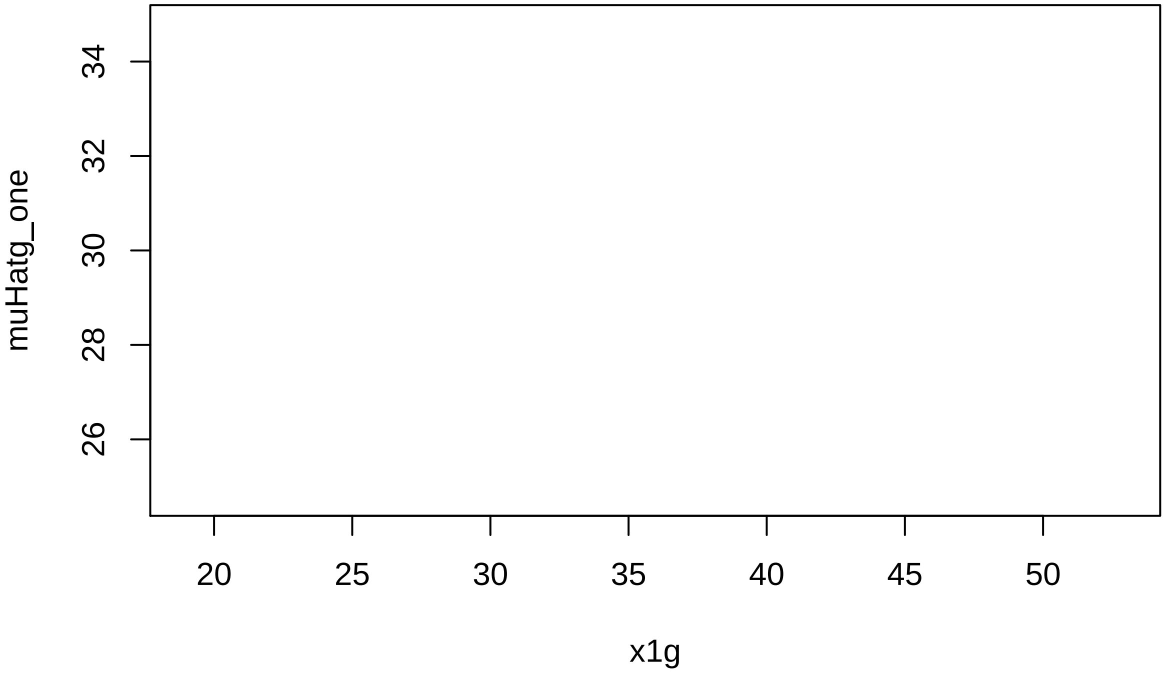
```
ng <- 101
x2g <- seq(from=min(x2), to=max(x2), length=ng)

newDataDF <- as.data.frame(
          cbind(
          rep(mean(x1), ng)
          , x2g
          , rep(55, ng)
          , rep(25, ng)
          )
  )
names(newDataDF) <- c("x1", "x2", "z1", "z2")
newDataList <- as.list(newDataDF)

predObj_two <- predict(mod_two, newDataList, se=T)
```

```
muHatg_two <- predObj_two$fit
aa_two     <- predObj_two$fit + 2*predObj_two$se
bb_two     <- predObj_two$fit - 2*predObj_two$se
lowergg_two <- 1 / (1 + exp(-bb_two))
uppergg_two <- 1 / (1 + exp(-aa_two))
```

```
plot(x1g, muHatg_one, type="n")
polygon(c(x1g, rev(x1g))
       , c(lowergg_one, rev(uppergg_one))
       , col="cadetblue1"
        )
```


muHatg_one
x1g