Home Work # 8

Q1 Show that for SVM method $f(x) = h(x)^T\beta + \beta_0$ the two optimization problem (1) and (2) are equivalent.

$$\min_{\beta_0,\beta} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i \quad \text{— (1)}$$

Subject to $\xi_i \geq 0$, $y_i(h(x_i)^T\beta + \beta_0) \geq 1 - \xi_i \ \forall i$,

$$\min \sum_{i=1}^{N}[1 - y_i f(x_i)] + \frac{\lambda}{2}||\beta||^2 \quad \text{— (2)}$$

Solution ⇒

$$y_i(h(x_i)^T\beta + \beta_0) \geq 1 - \xi_i$$

$$\xi_i \geq 1 - y_i(h(x_i)^T\beta + \beta_0)$$

$$\xi_i \geq 1 - y_i f(x_i)$$

Let take equation (1)

$$\min_{\beta_0,\beta} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i$$

putting the value of ξ_i in equation (1).

$$\min_{\beta_0,\beta} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}[1 - y_i(h(x_i)^T\beta + \beta_0)]$$

$$\{ f(x_i) = h(x_i)^T\beta + \beta_0 \}$$

$$\min_{\beta_0, \beta} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{N} [1 - y_i f(x_i)] \quad — (4)$$

we know that $\quad C = 1/\lambda$

So we can replace C with $1/\lambda$ in equation (4)

$$\min_{\beta_0, \beta} \frac{1}{2} \|\beta\|^2 + \frac{1}{\lambda} \sum_{i=1}^{N} [1 - y_i f(x_i)]$$

cross multiplication of λ.

$$\min_{\beta_0, \beta} \frac{\lambda}{2} \|\beta\|^2 + \sum_{i=1}^{N} [1 - y_i f(x_i)] \quad — (5)$$

Now we can see the both equation (1) and (2) are equivalent to each other with condition

$$\boxed{C = 1/\lambda}$$

Now we can apply Lagrangian :→

$$L(x, \{\lambda_i\}) = f(x) + \sum \lambda_i g_i(x)$$

$$\lambda_i \geq 0$$

Solution $\frac{\partial L}{\partial x} = 0$ or $\frac{\partial L}{\partial \beta}, \frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \xi}$

$$L = \left[\frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{n} \xi_i\right] + \sum_{j=1}^{N} a_j \left[1 - \xi_i - y_i [h(x_i)^T \beta + \beta_0]\right]$$
$$+ \sum \lambda_i (\xi_i)$$

$$\frac{\partial L}{\partial \beta} = \beta - \sum a_i y_i h^T(x_i) = 0$$

$$\beta = \sum a_i y_i h^T(x_i)$$

$$\frac{\partial L}{\partial \beta_0} = \sum a_i y_i = 0 \qquad \sum a_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = C - a_i - \lambda_i = 0$$

again with ~~we~~ new values.

$$L = \left[\frac{1}{2}\left(\sum a_j y_j h^T(x_j)\right)^T\left(\sum a_i y_i h^T(x_i)\right) + C\sum \xi_i\right.$$

$$+ \sum_{j=1}^{N}\left[a_j\left\{1 - \xi_j - y_j\left[\left(\sum a_i y_i h^T(x_i)\right)^T h^T(x_j) + \beta_0\right)\right\}\right]$$

$$+ \sum \lambda_j (-\xi_j)$$

$$= \frac{1}{2}\sum a_j a_i y_j y_i h(x_i) h^T(x_j) + C\sum_i \xi_i$$

$$+ \sum_{i=1}^{N} a_i - \sum a_i \xi_i + \sum \lambda_i (-\xi_i)$$

$$= \sum a_i + \frac{1}{2}\sum a_i a_j y_i y_j h(x_i) h^T(x_j)$$

$$+ \sum \xi_i \underbrace{(C - a_i - \lambda_i)}_{\to 0} =$$

So the Dual form of SVM

$$\max_{a_i} L(\{\lambda_i\}, \{a_i\}) = \sum a_i + \frac{1}{2} \sum a_i a_j y_i y_j h(x_i) h(x_j)^T$$

#

Home work 8

Arinjay Jain

December 4, 2020

```r
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```r
require(gam)
```

```
## Loading required package: gam
```

```
## Loading required package: splines
```

```
## Loading required package: foreach
```

```
## Loaded gam 1.16.1
```

```r
SAheard <- read.table("http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data",sep=",",
summary(SAheard)
```

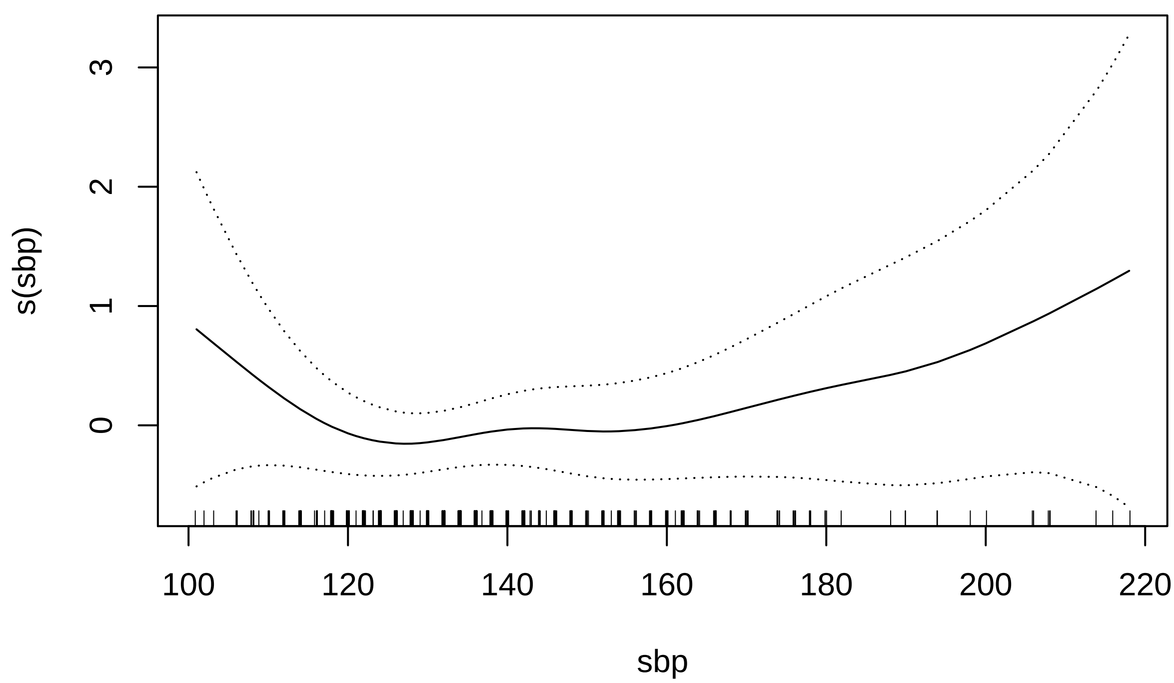
```
##       sbp           tobacco             ldl           adiposity
##  Min.   :101.0   Min.   : 0.0000   Min.   : 0.980   Min.   : 6.74
##  1st Qu.:124.0   1st Qu.: 0.0525   1st Qu.: 3.283   1st Qu.:19.77
##  Median :134.0   Median : 2.0000   Median : 4.340   Median :26.11
##  Mean   :138.3   Mean   : 3.6356   Mean   : 4.740   Mean   :25.41
##  3rd Qu.:148.0   3rd Qu.: 5.5000   3rd Qu.: 5.790   3rd Qu.:31.23
##  Max.   :218.0   Max.   :31.2000   Max.   :15.330   Max.   :42.49
##     famhist        typea         obesity         alcohol             age
##  Absent :270   Min.   :13.0   Min.   :14.70   Min.   :  0.00   Min.   :15.00
##  Present:192   1st Qu.:47.0   1st Qu.:22.98   1st Qu.:  0.51   1st Qu.:31.00
##                Median :53.0   Median :25.80   Median :  7.51   Median :45.00
##                Mean   :53.1   Mean   :26.04   Mean   : 17.04   Mean   :42.82
##                3rd Qu.:60.0   3rd Qu.:28.50   3rd Qu.: 23.89   3rd Qu.:55.00
##                Max.   :78.0   Max.   :46.58   Max.   :147.19   Max.   :64.00
##       chd
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3463
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

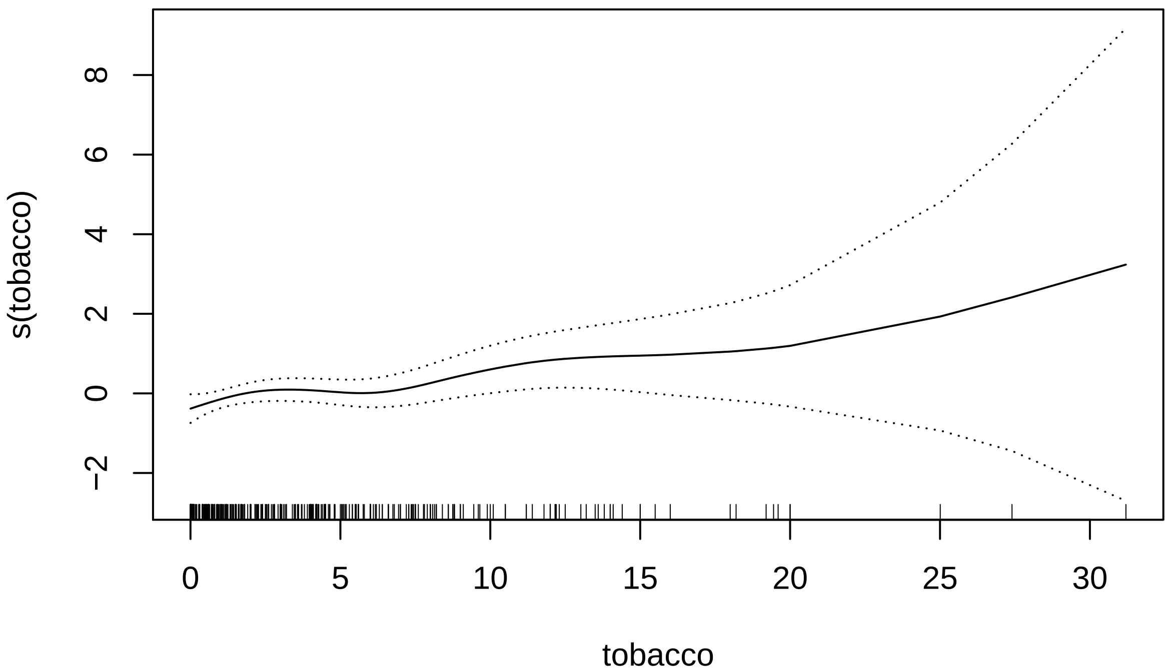
```r
names(SAheard)
```

```
##  [1] "sbp"       "tobacco"   "ldl"       "adiposity" "famhist"   "typea"
##  [7] "obesity"   "alcohol"   "age"       "chd"
```

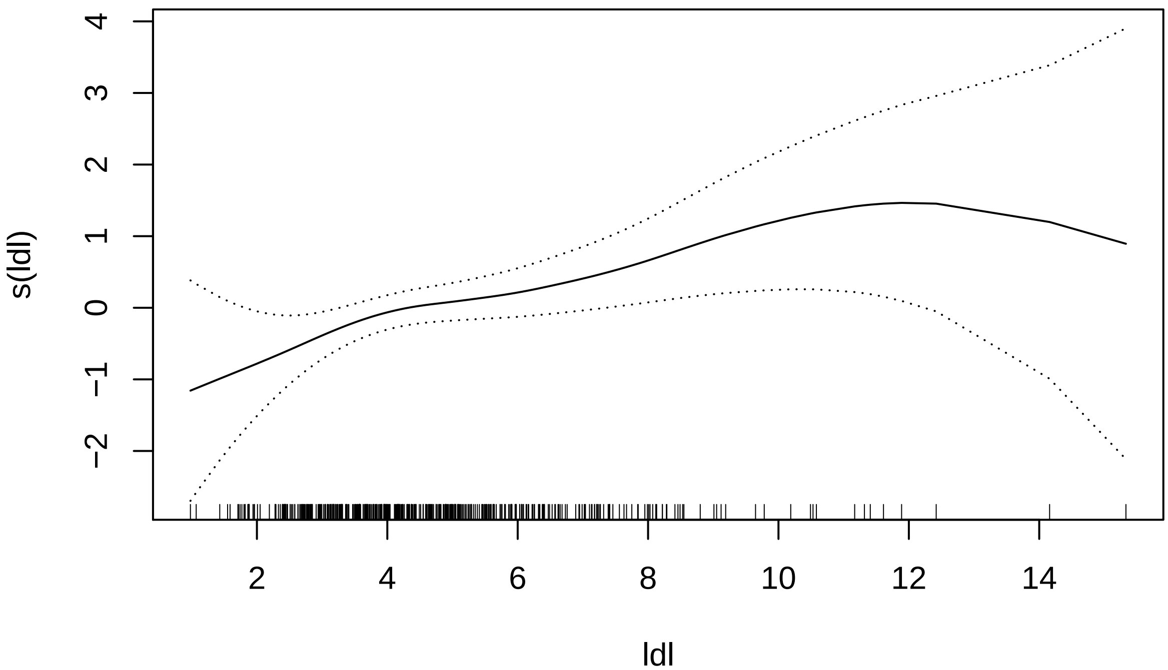
part A

```r
SAheard_Gam <- gam(chd ~ s(sbp) + s(tobacco) + s(ldl) + s(adiposity) + s(typea) +
    s(obesity) + s(alcohol) + s(age) + famhist,data=SAheard,family=binomial)
plot(SAheard_Gam,se=TRUE)
```

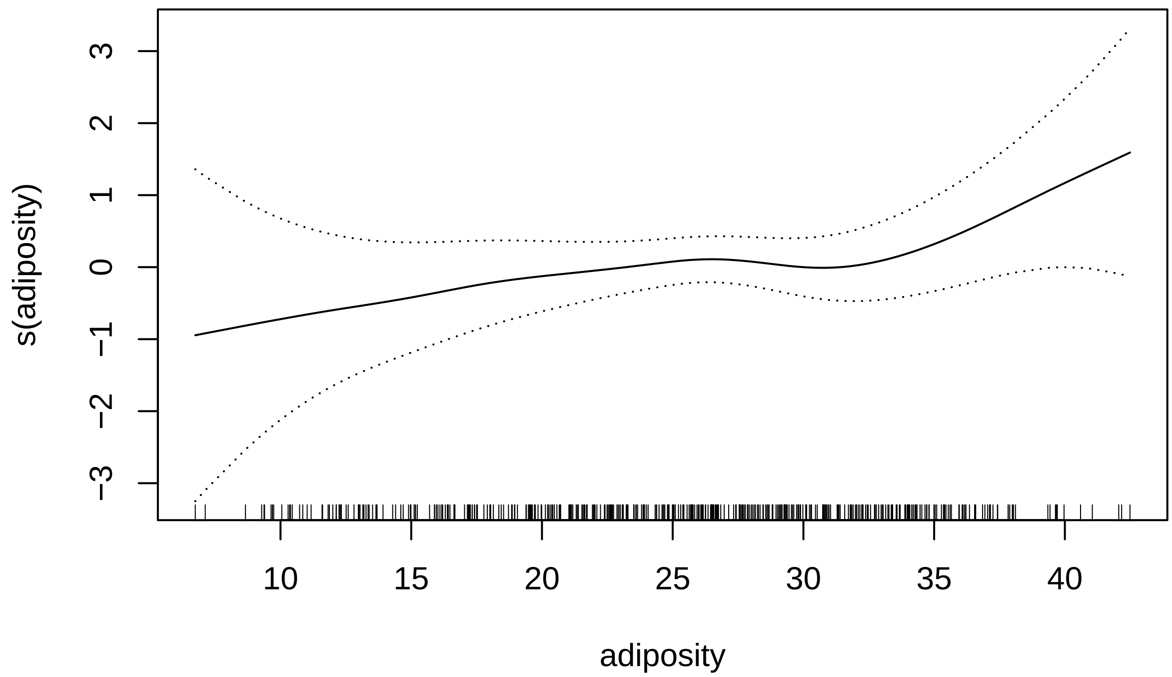




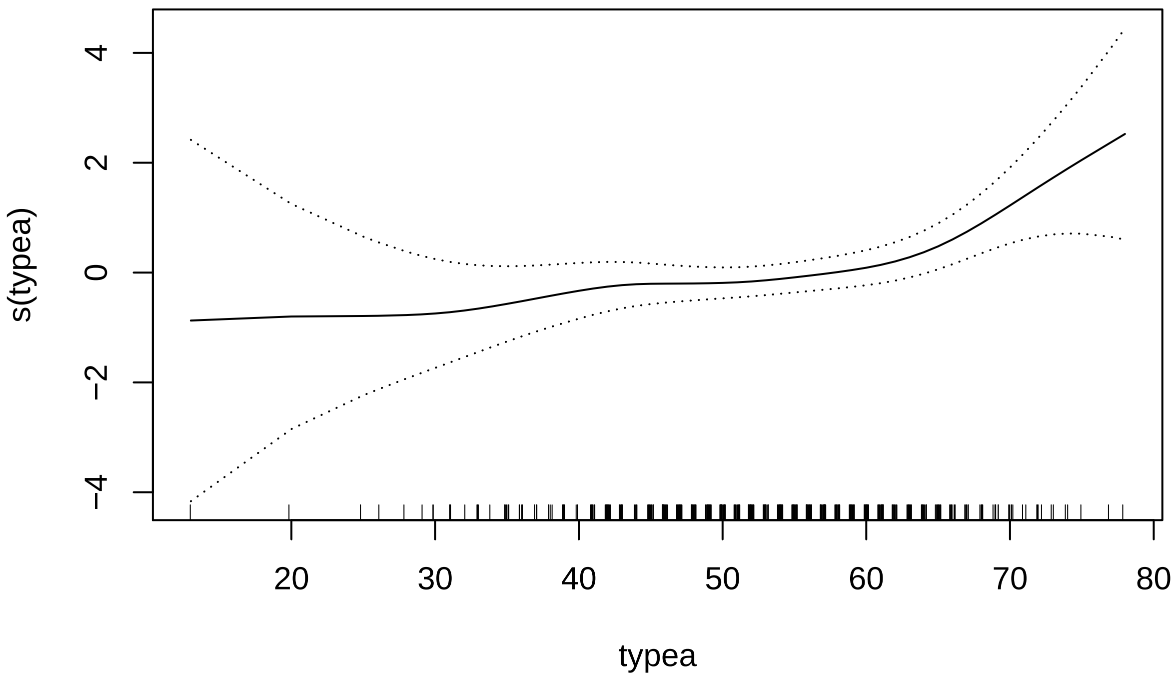
s(tobacco)
−2
0
2
4
6
8
0
5
10
15
20
25
30
tobacco



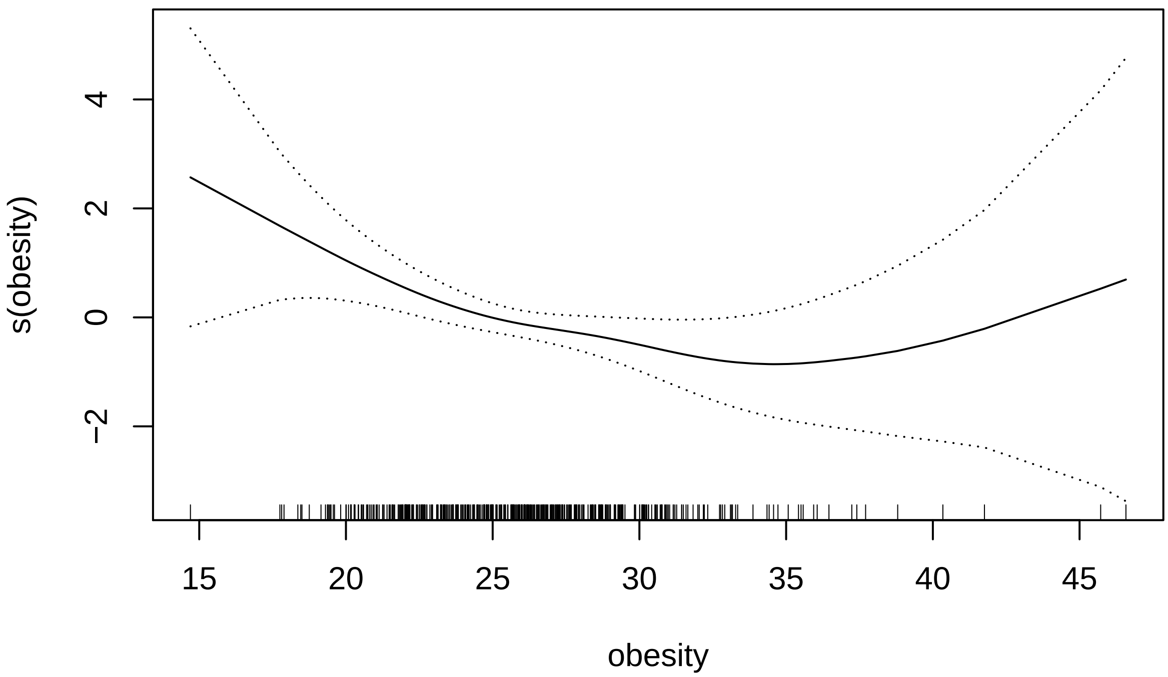
s(ldl)
4
3
2
1
0
−1
−2
2
4
6
8
10
12
14
ldl

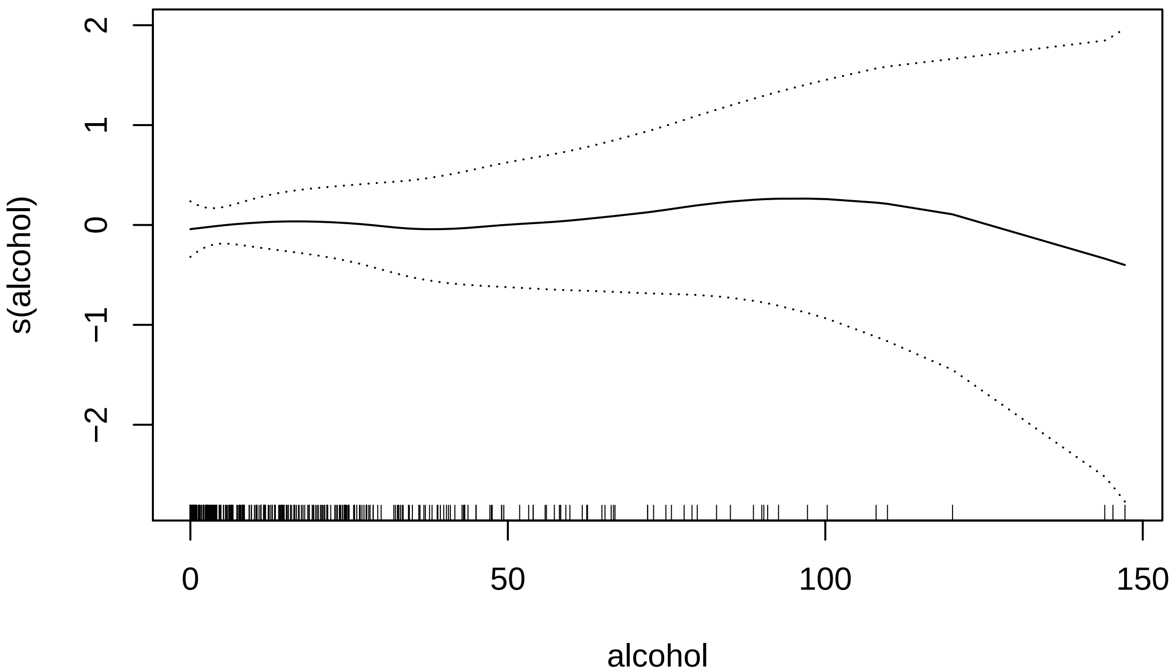


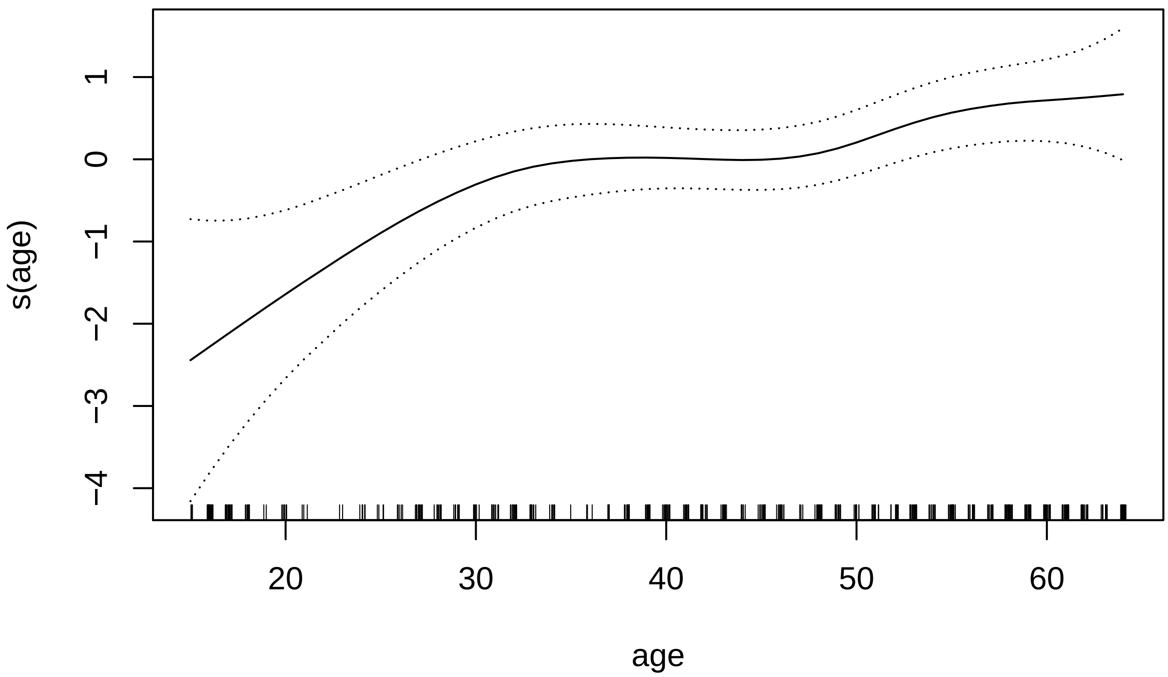
s(adiposity)
3
2
1
0
−1
−2
−3
10
15
20
25
30
35
40
adiposity

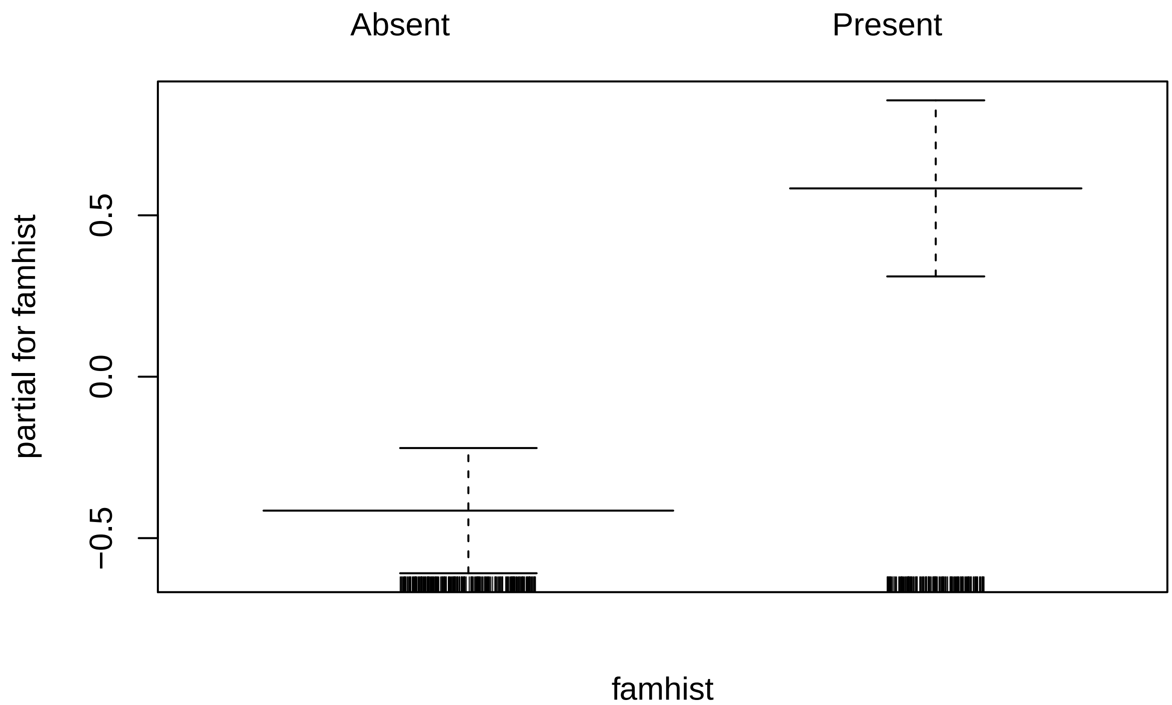


s(typea)
4
2
0
−2
−4
20
30
40
50
60
70
80
typea


s(obesity)
4
2
0
−2
15
20
25
30
35
40
45
obesity

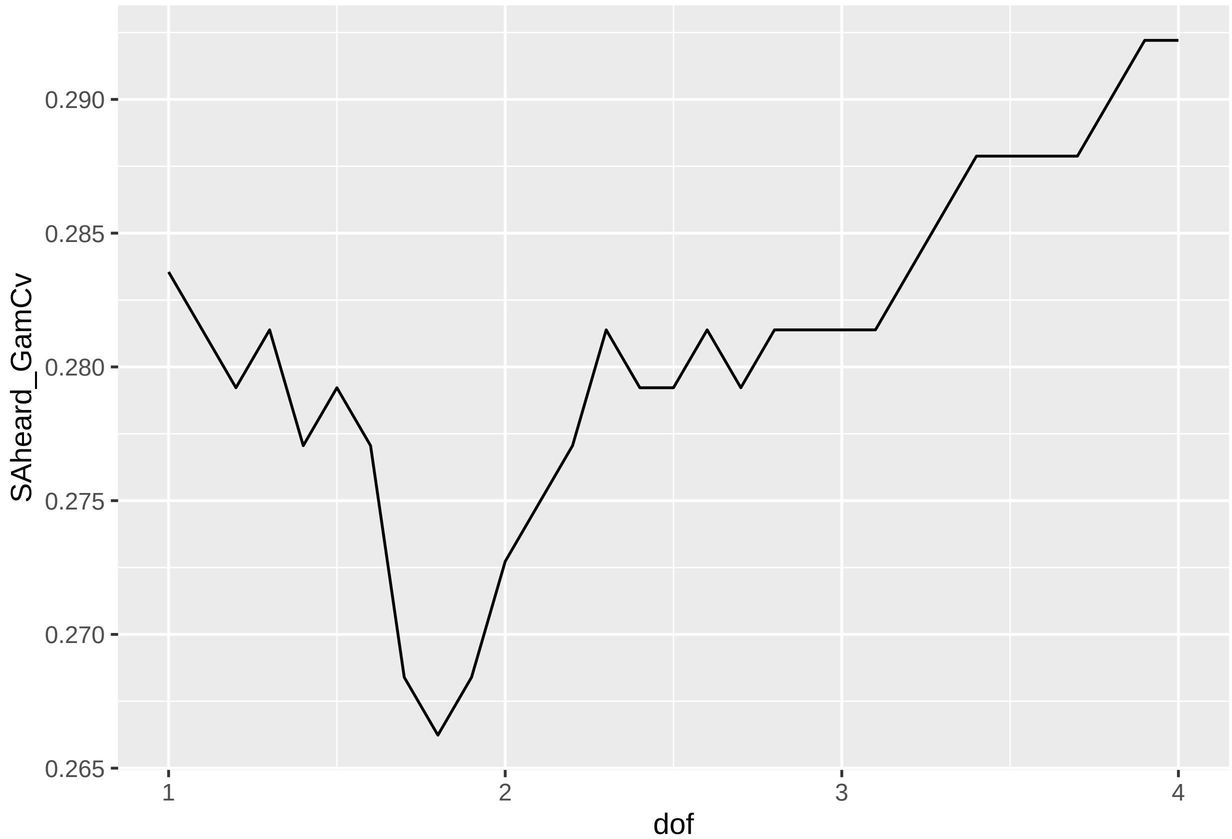

2
1
0
−1
−2
s(alcohol)
0
50
100
150
alcohol


s(age)
1
0
−1
−2
−3
−4
20
30
40
50
60
age

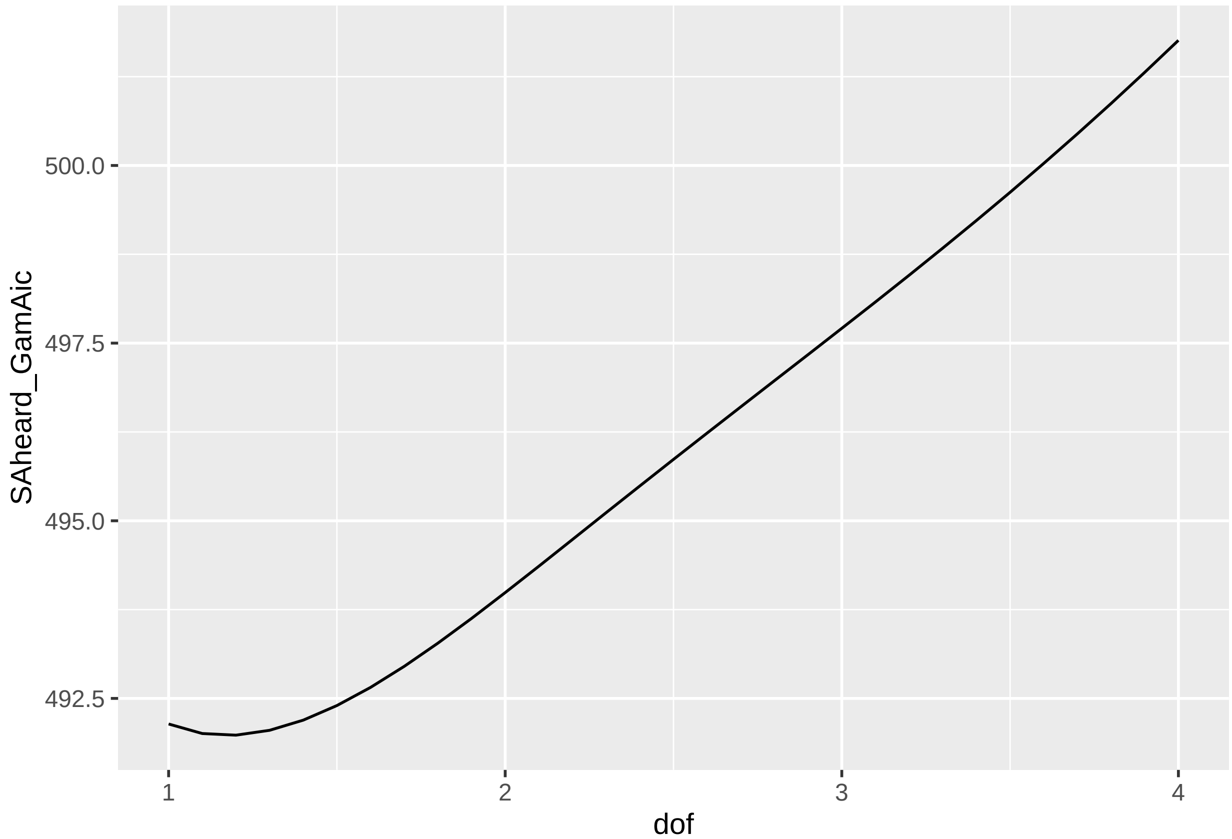


Part B

```
library(boot)
Loss <- function(r, pi=0) mean(abs(r-pi)>0.5)
likelihood <- function(r,pi=0) -mean(log(r*pi+(1-r)*(1-pi)))
dof <- seq(1,4,by=0.1)

#Using cross-validation with 7 folds

SAheard_GamCv <- numeric(length(dof))

for(i in seq(along=dof)){
formGam <- as.formula(paste("chd~famhist+",paste("s(",names(SAheard[1,1:9])[-5], ",df=", dof[i], ")",se
SAheard_Gam_CV <- gam(formGam,family=binomial,data=SAheard)
tmp  <- cv.glm(SAheard,SAheard_Gam_CV,Loss,7)
set.seed(tmp$seed)
SAheard_GamCv[i] <- tmp$delta[1]
}

qplot(dof,SAheard_GamCv,geom="line")
```

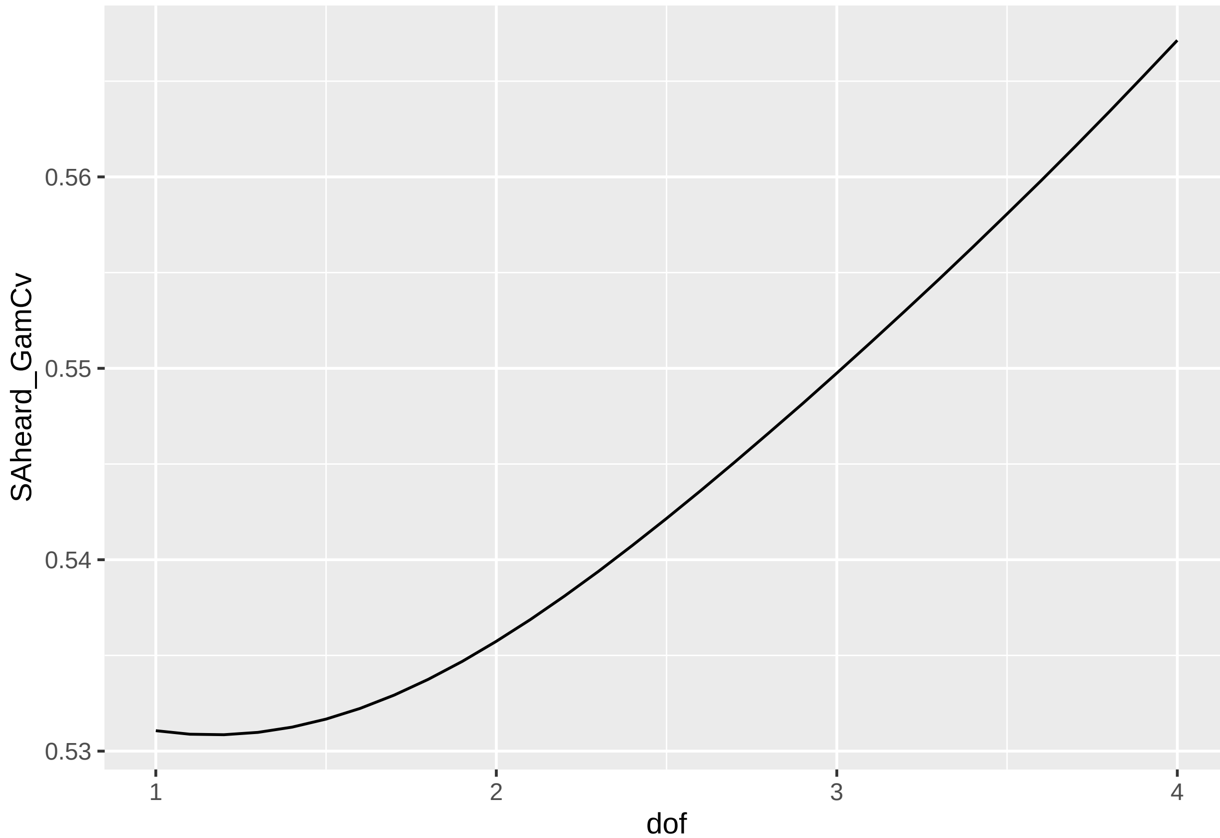



```
###use AIC criteria. using the effective degrees of freedom

SAheard_GamAic <- numeric(length(dof))
for(i in seq(along=dof)){
formGam <- as.formula(paste("chd~famhist+",paste("s(",names(SAheard[1,1:9])[-5], ",df=", dof[i], ")",sep
SAGam_Aic <- gam(formGam,family=binomial,data=SAheard)
SAheard_GamAic[i]  <- SAGam_Aic$aic
}

qplot(dof,SAheard_GamAic,geom="line")
```



Part C

```
for(i in seq(along=dof)){
formGam <- as.formula(paste("chd~famhist+",paste("s(",names(SAheard[1,1:9])[-5], ",df=", dof[i], ")",sep
SAGam <- gam(formGam,family=binomial,data=SAheard)
tmp  <- cv.glm(SAheard,SAGam,likelihood,7)
set.seed(tmp$seed)
SAheard_GamCv[i] <- tmp$delta[1]
}

qplot(dof,SAheard_GamCv,geom="line")
```


SAheard_GamCv
0.56
0.55
0.54
0.53
1
2
3
4
dof

Part D Using MGCV

Arinjay Jain

December 4, 2020

```r
library(mgcv)
```

```
## Warning: package 'mgcv' was built under R version 3.6.3
```

```
## Loading required package: nlme
```

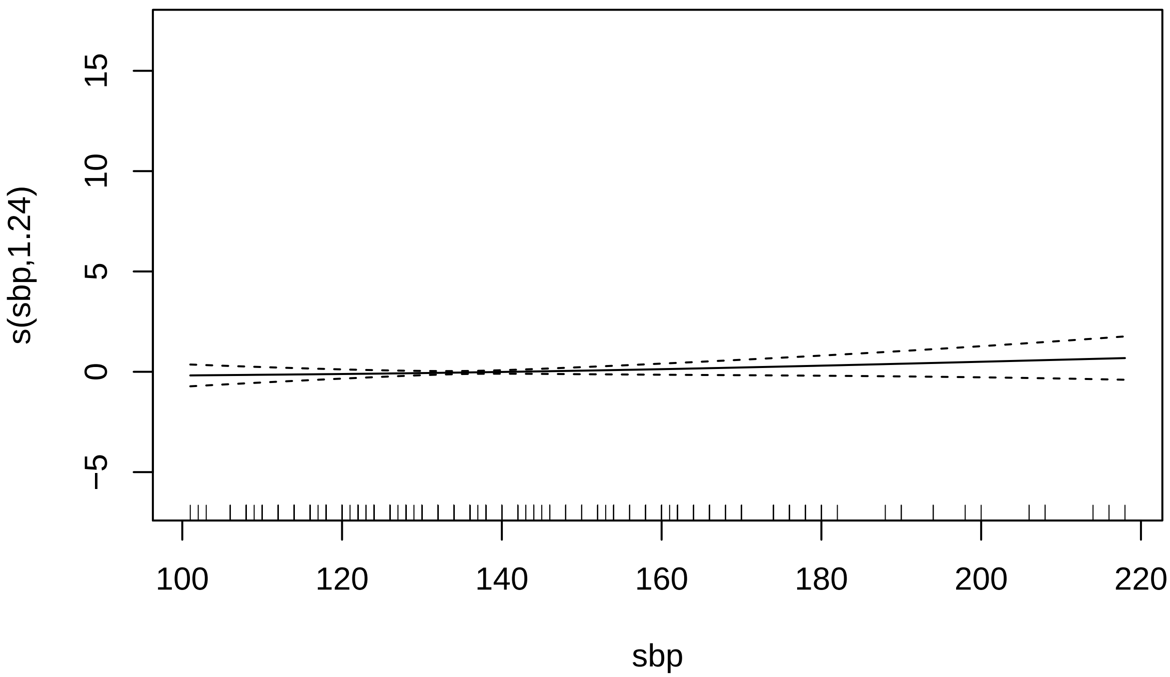
```
## This is mgcv 1.8-33. For overview type 'help("mgcv-package")'.
```

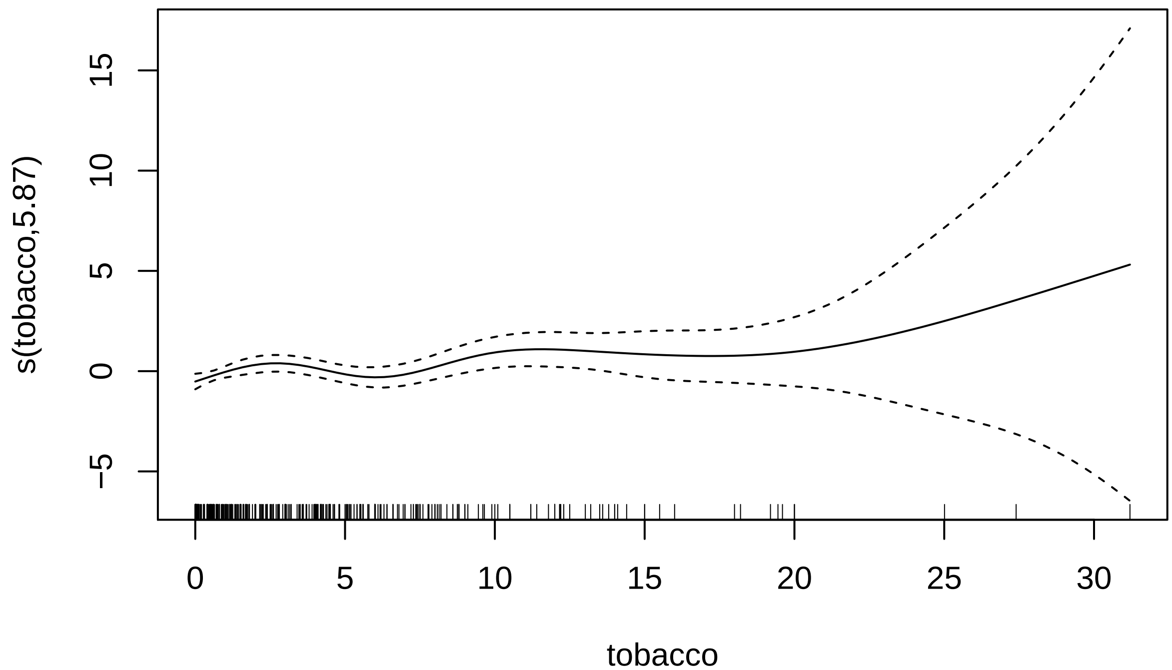
```r
SA_mgcv <- read.table("http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data",sep=",",]

formGam_mgcv <- as.formula(paste("chd~famhist+",paste("s(",names(SA_mgcv[1,1:9])[-5],")",sep="",collaps
SAGam_mgcv <- gam(formGam_mgcv,family=binomial,data=SA_mgcv)
summary(SAGam_mgcv)
```

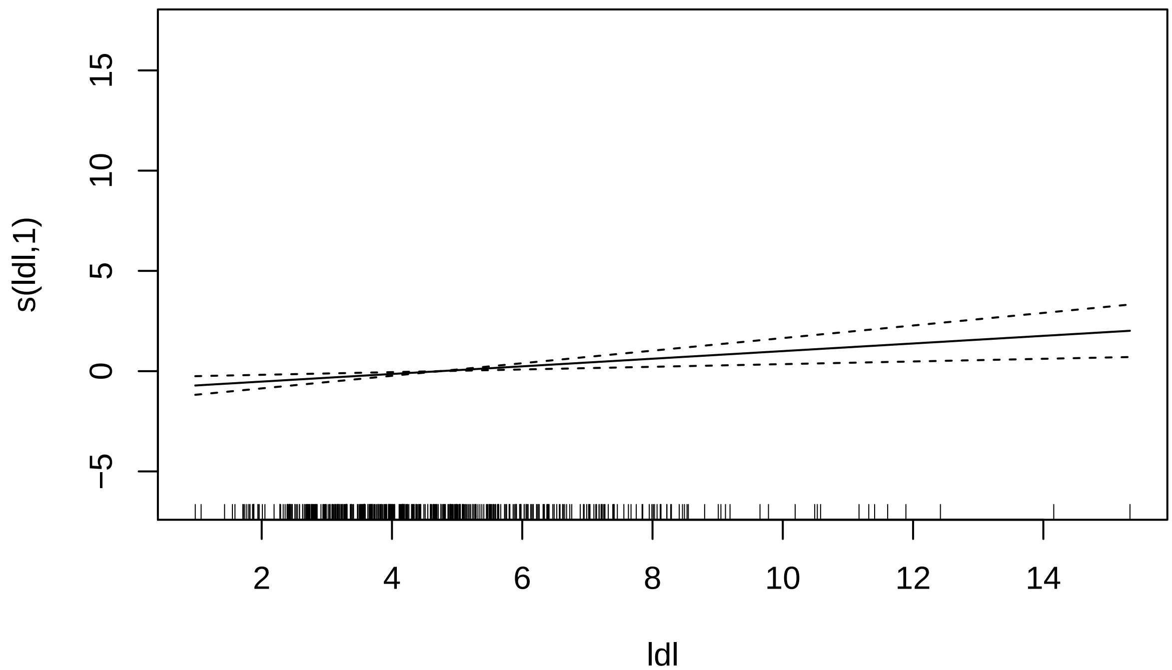
```
##
## Family: binomial
## Link function: logit
##
## Formula:
## chd ~ famhist + s(sbp) + s(tobacco) + s(ldl) + s(adiposity) +
##     s(typea) + s(obesity) + s(alcohol) + s(age)
##
## Parametric coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.3331     0.1796  -7.421 1.16e-13 ***
## famhistPresent   0.9443     0.2347   4.023 5.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df Chi.sq p-value
## s(sbp)       1.235  1.434  1.979 0.33042
## s(tobacco)   5.865  6.989 16.500 0.02033 *
## s(ldl)       1.000  1.000  9.466 0.00209 **
## s(adiposity) 1.000  1.000  1.186 0.27624
## s(typea)     3.329  4.203 13.386 0.01161 *
## s(obesity)   2.204  2.840  5.530 0.11458
## s(alcohol)   1.000  1.000  0.013 0.90773
## s(age)       3.394  4.228 12.687 0.01349 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.267   Deviance explained = 25.8%
## UBRE = 0.048901  Scale est. = 1         n = 462
```

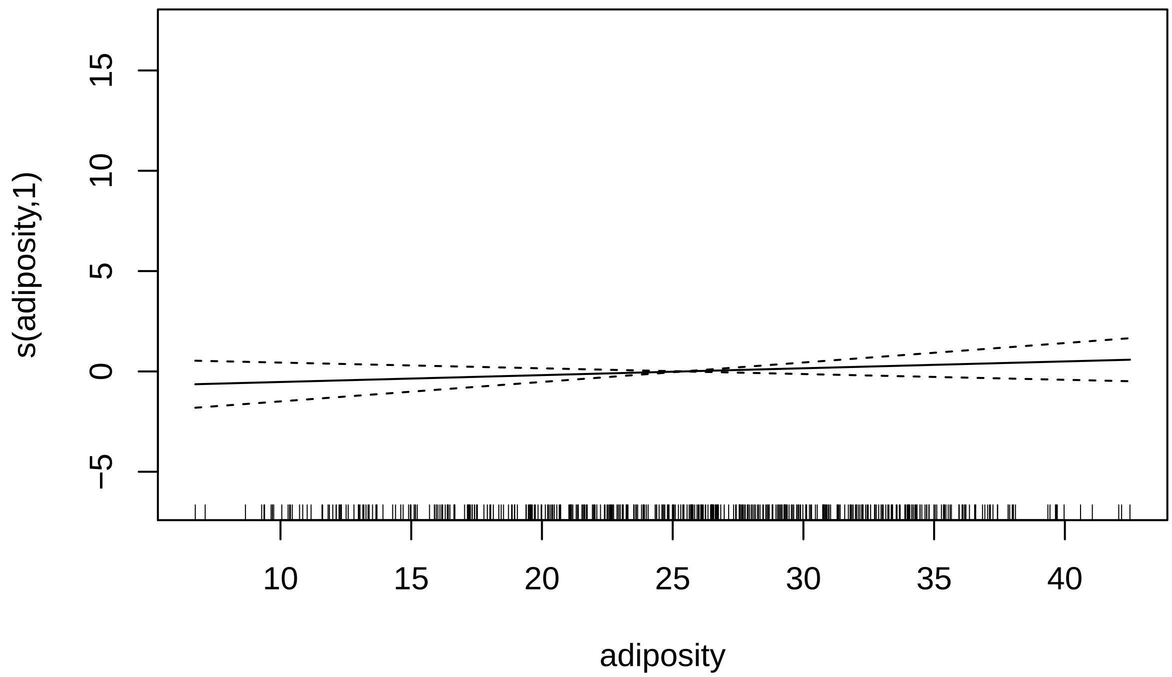
```r
plot(SAGam_mgcv,se=TRUE)
```



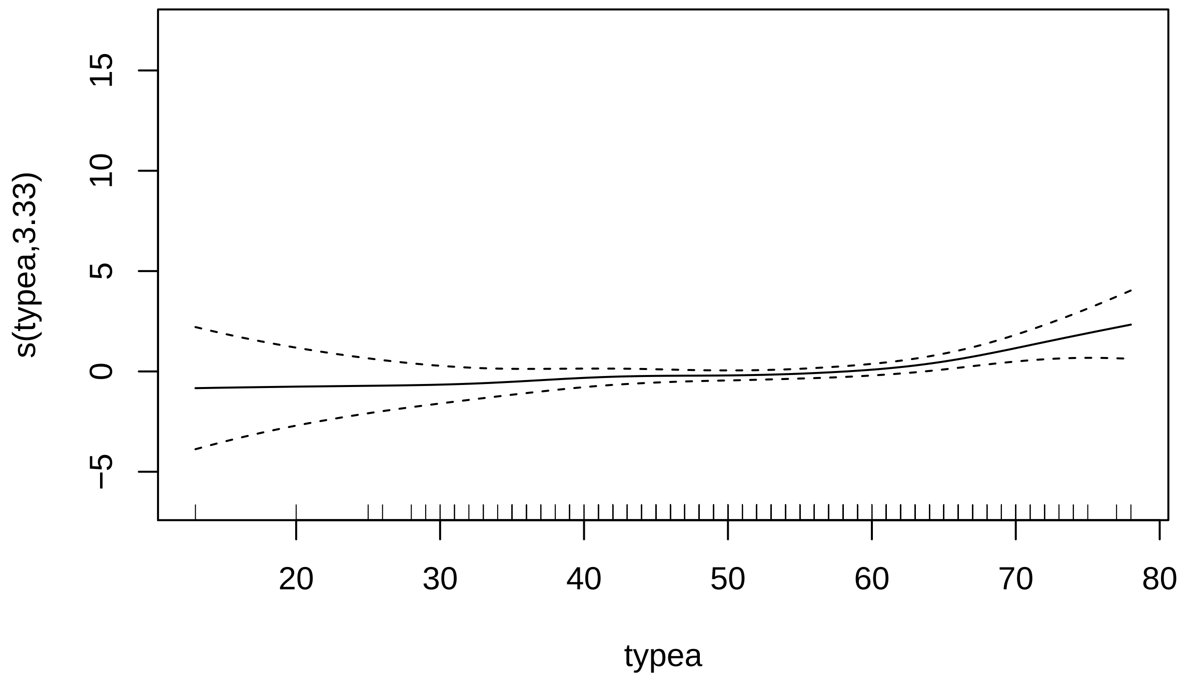

s(tobacco,5.87)
15
10
5
0
−5
0
5
10
15
20
25
30
tobacco

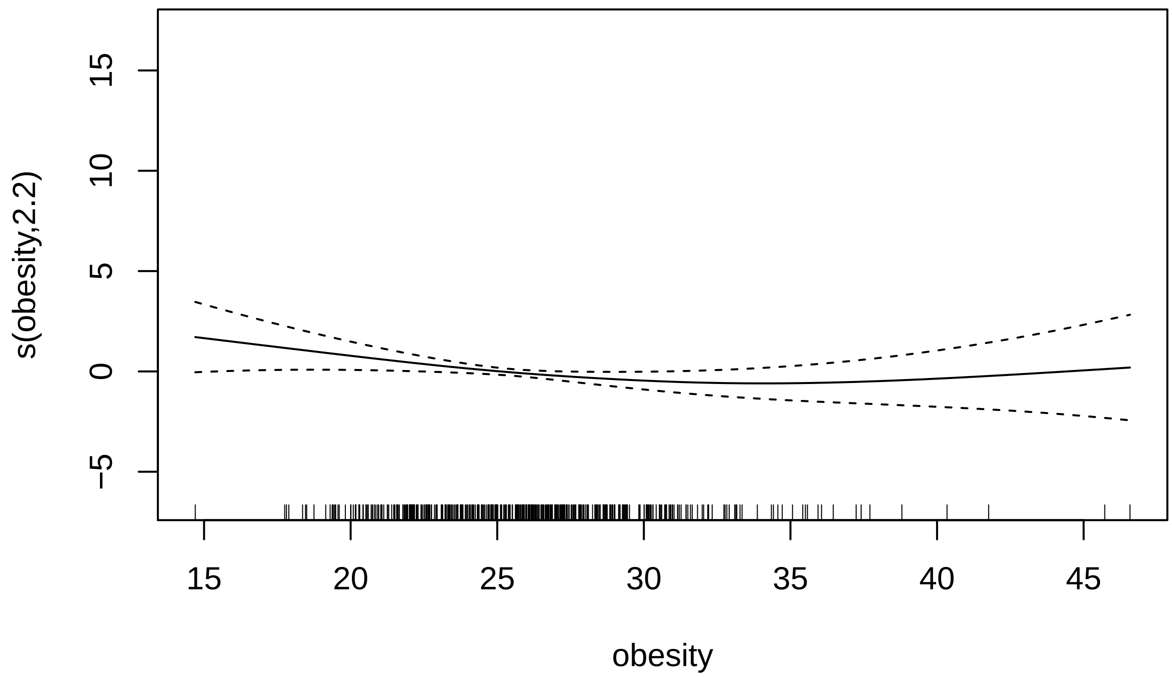


s(ldl,1)
15
10
5
0
−5
2
4
6
8
10
12
14
ldl

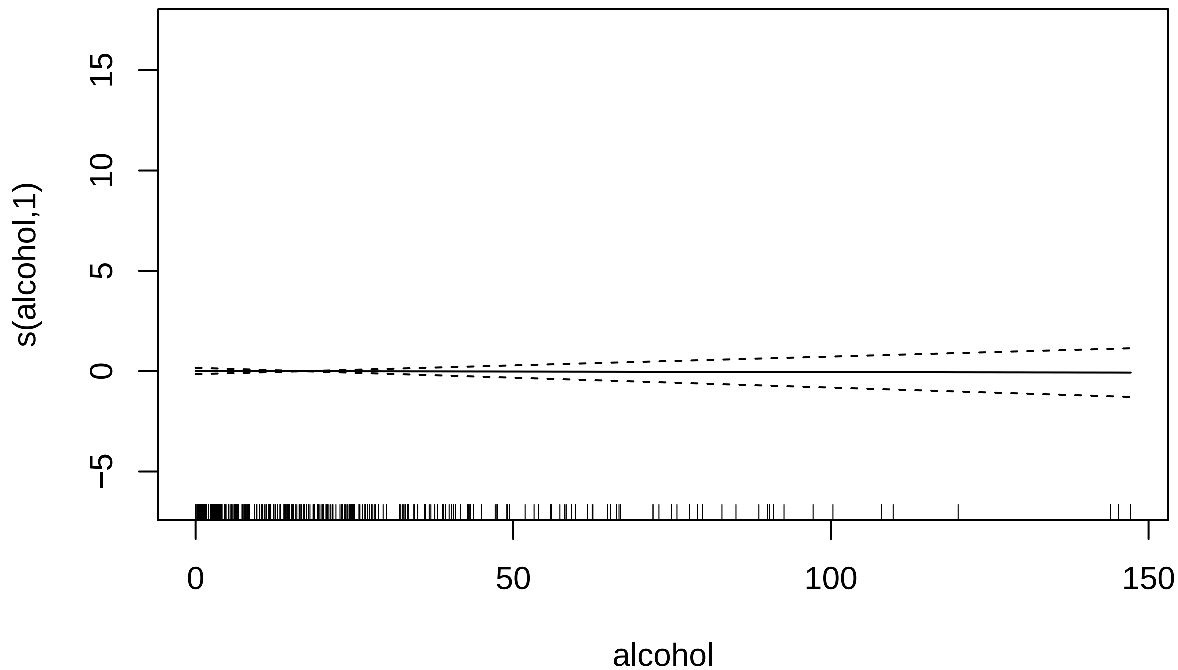


s(adiposity,1)
15
10
5
0
−5
10
15
20
25
30
35
40
adiposity

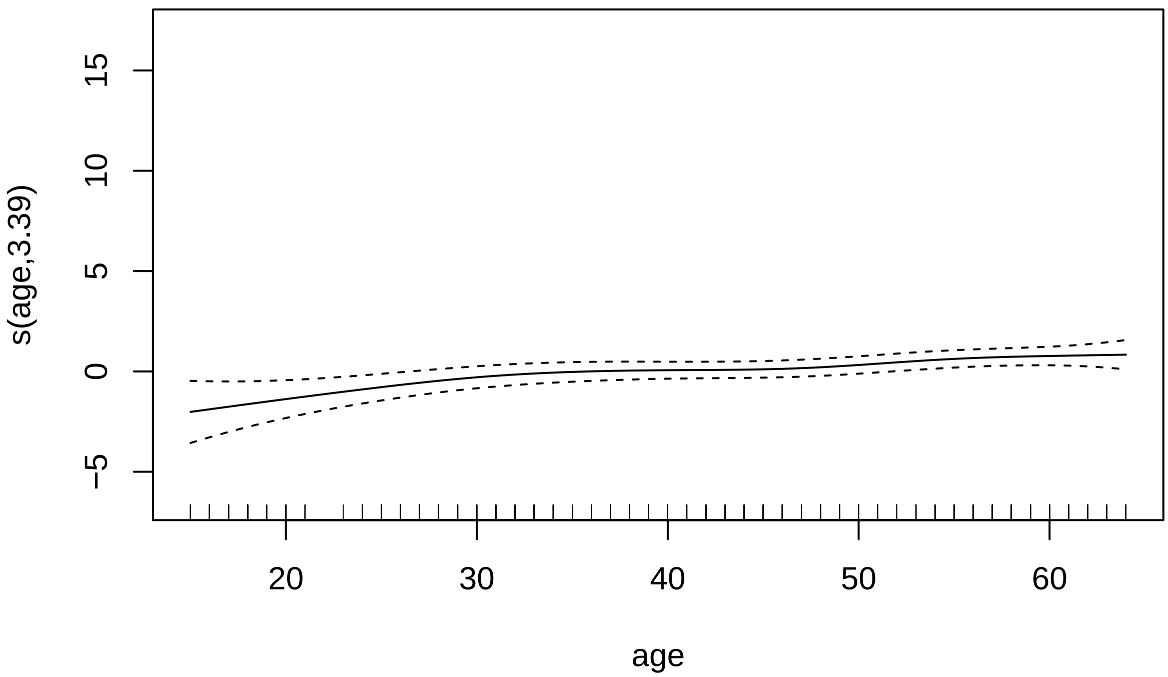

s(typea,3.33)
15
10
5
0
−5
20
30
40
50
60
70
80
typea



s(obesity,2.2)
15
10
5
0
−5
15
20
25
30
35
40
45
obesity



s(alcohol,1)
15
10
5
0
−5
0
50
100
150
alcohol



15
10
5
0
−5
s(age,3.39)
20
30
40
50
60
age