portfolio_6

2023-03-08

Generalized Additive Models

For this task we will be using the wesdr dataset from the gss package. The dataset records wether or not diabetic patients developed retinopathy (*ret* column) (0=no, 1=yes) along with three predictor variables: duration of disease in tears (dur), glycosylated hemoglobin - the percentage of hemoglobin bound to glucuse in the blood (gly) and body mass index (bmi):

```
library(gss)
data(wesdr)
head(wesdr)
```

```
##    dur  gly  bmi ret
## 1 10.3 13.7 23.8   0
## 2  9.9 13.5 23.5   0
## 3 15.6 13.8 24.8   0
## 4 26.0 13.0 21.6   1
## 5 13.8 11.1 24.6   1
## 6 31.1 11.3 24.6   1
```

Let's now split our data into a training and testing set:

```
n.test <- round(0.15 * nrow(wesdr))
test_ind <- sample(seq_len(nrow(wesdr)), size = n.test)

train <- wesdr[-test_ind, ]
test <- wesdr[test_ind, ]
```

Let's now fit a GAM model, where we will choose the penalty parameters, $\{\lambda_j\}_{j=1}^p$, using Generalized Cross Validation (GCV) which is the method gam uses by default:

```
library(mgcv)
```

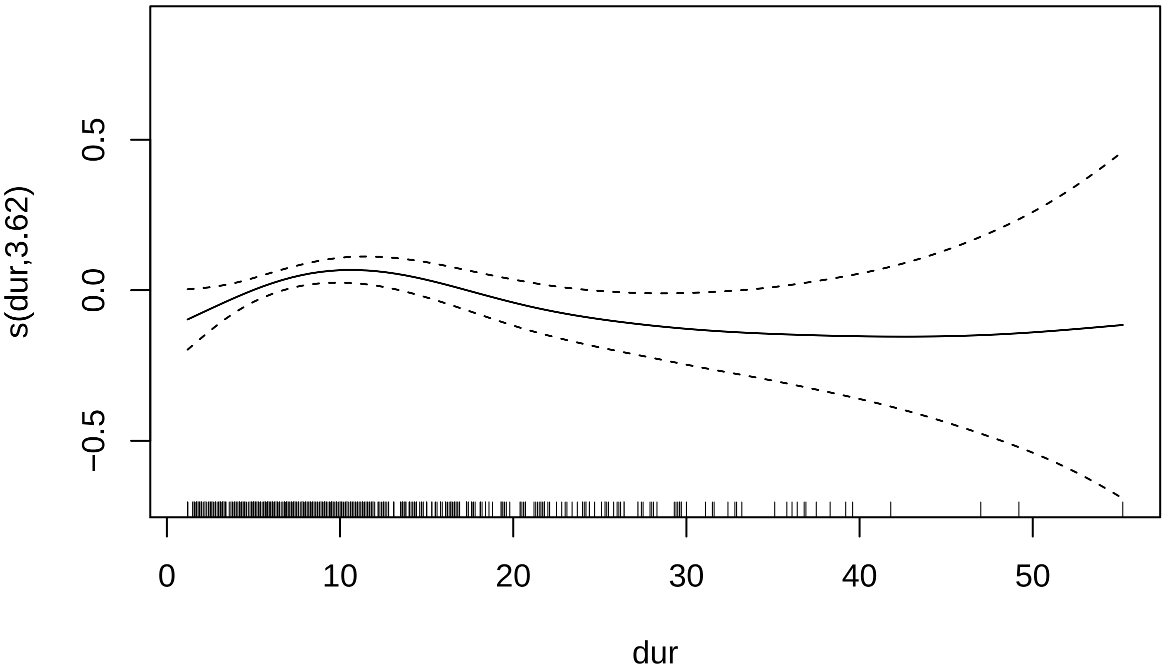
```
## Loading required package: nlme
```

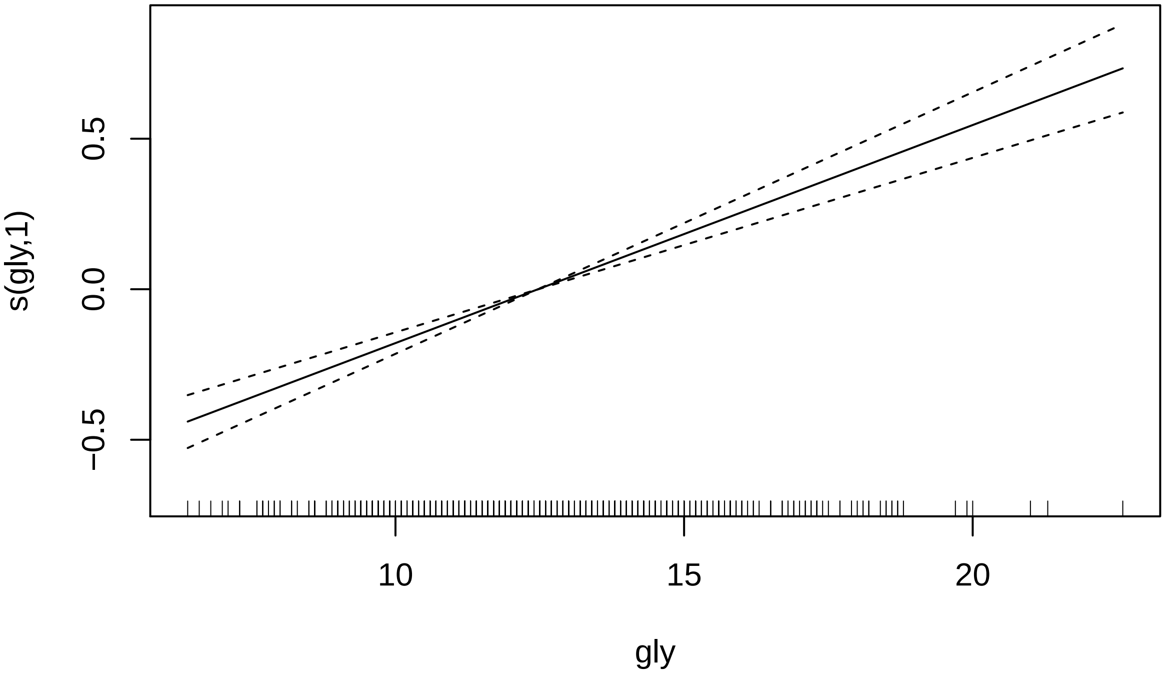
```
## This is mgcv 1.8-42. For overview type 'help("mgcv-package")'.
```

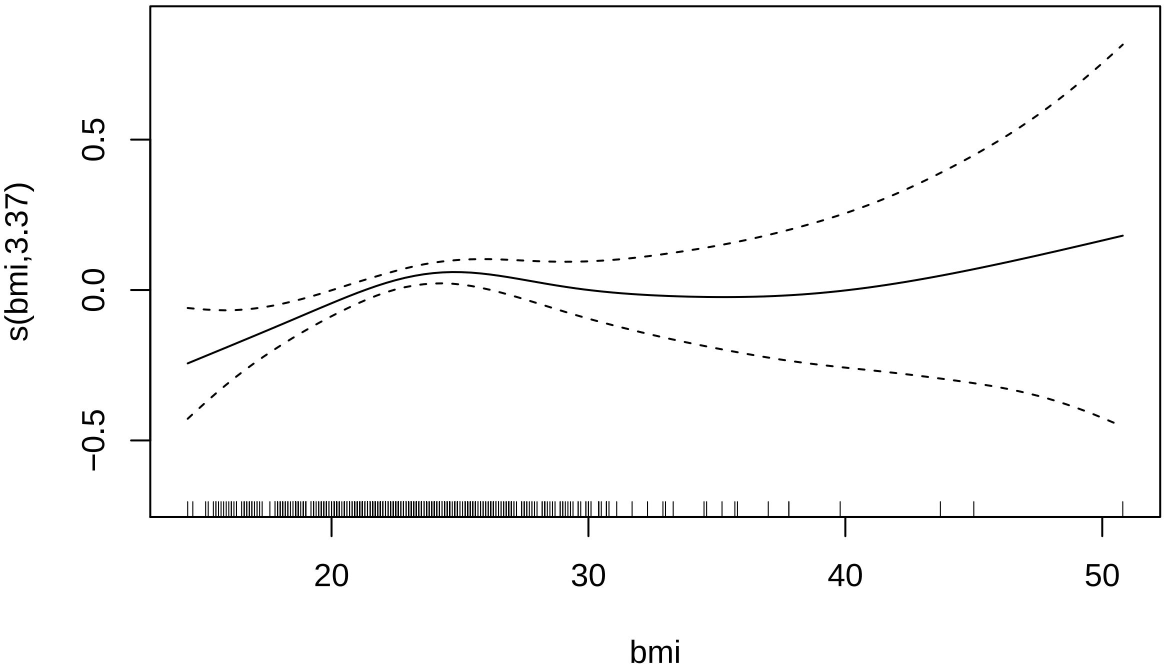
```
fit <- gam(ret~s(dur)+s(gly)+s(bmi), data = train)
```

Let's now plot our estimated functions:

```
plot(fit)
```


s(dur,3.62)
0.5
0.0
−0.5
0
10
20
30
40
50
dur


s(gly,1)
0.5
0.0
−0.5
10
15
20
gly



From the above we see that the model for gly is linear, however, the models for the two other features are non-linear. Especially the model for the feature dur, it could be argued that perhaps a linear model would be suitable for the bmi. So for this dataset it appears that using a GAM is a reasonable choice (as opposed to picking a GLM for example).

Let's now evaluate the performance of our fitted GAM by evaluating its performance on the test set:

```
preds <- predict(fit, newdata = test)
diff <- as.vector(test[,"ret"]) - preds
mse <- mean((diff)^2)
mse
```

```
## [1] 0.176479
```

Let's now fit a GLM and compare its performance on the test set versus the GAM:

```
fit2 <- gam(ret~dur+gly+bmi, data = train)

preds2 <- predict(fit2, newdata = test)
diff2 <- as.vector(test[,"ret"]) - preds2
mse2 <- mean((diff2)^2)
mse2
```

```
## [1] 0.1876876
```

Comparing the mse's we see that the mse achieved by fitting a GLM is smaller than that obtained by fitting a GAM. This suggests that modelling all the features using linear models is perhaps a better option and that the bmi and dur features are better modeled linearly.