Homework 9

Skylar Liu

2024-03-06

1. Install the package AER. Then get the library(AER), then data(HousePrices) and then library(mgcv)

```
rm(list = ls())
set.seed(382957)
options(repos = list(CRAN="http://cran.rstudio.com/"))

#install and load data and packages
install.packages('AER')
```

```
## Installing package into 'C:/Users/slkoe/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## package 'AER' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\slkoe\AppData\Local\Temp\RtmpAdvbPP\downloaded_packages
```

```
library(AER)
```

```
## Warning: package 'AER' was built under R version 4.3.3

## Loading required package: car

## Loading required package: carData

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival
```

```r
data("HousePrices")
library(mgcv)
```

```
## Loading required package: nlme

## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

2. Run the code >fitGaussAM = gam(price ~ s(lotsize,bs="cr",k=105)

- bedrooms
- factor(bathrooms)
- factor(stories)
- factor(driveway)
- factor(recreation)
- factor(fullbase)
- factor(gasheat)
- factor(aircon)
- garage
- factor(prefer), data = HousePrices,family = gaussian) summary(fitGaussAM)

Is lot size statistically significant at p=0.05?

```r
# run example code
fitGaussAM = gam(price ~ s(lotsize,bs="cr",k=105)
                + bedrooms
                + factor(bathrooms)
                + factor(stories)
                + factor(driveway)
                + factor(recreation)
                + factor(fullbase)
                + factor(gasheat)
                + factor(aircon)
                + garage
                + factor(prefer),
                data = HousePrices,family = gaussian)
summary(fitGaussAM)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## price ~ s(lotsize, bs = "cr", k = 105) + bedrooms + factor(bathrooms) +
##     factor(stories) + factor(driveway) + factor(recreation) +
##     factor(fullbase) + factor(gasheat) + factor(aircon) + garage +
##     factor(prefer)
##
## Parametric coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              36007.2     3512.3  10.252  < 2e-16 ***
## bedrooms                  2220.0     1117.1   1.987 0.047453 *
## factor(bathrooms)2     11893.5     1736.2   6.850 2.21e-11 ***
## factor(bathrooms)3     25249.6     5321.3   4.745 2.74e-06 ***
## factor(bathrooms)4     63728.4    15825.8   4.027 6.55e-05 ***
## factor(stories)2        6030.3     1662.8   3.626 0.000317 ***
## factor(stories)3       13611.6     2916.8   4.667 3.95e-06 ***
## factor(stories)4       19676.3     3094.3   6.359 4.65e-10 ***
## factor(driveway)yes     5313.3     2048.9   2.593 0.009790 **
## factor(recreation)yes   3573.8     1925.8   1.856 0.064094 .
## factor(fullbase)yes     6184.3     1586.8   3.897 0.000111 ***
## factor(gasheat)yes     11560.0     3178.3   3.637 0.000305 ***
## factor(aircon)yes      11420.7     1590.0   7.183 2.55e-12 ***
## garage                  4267.0      852.9   5.003 7.88e-07 ***
## factor(prefer)yes      10612.0     1830.7   5.797 1.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df     F p-value
## s(lotsize) 39.37  47.13 3.806  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.699   Deviance explained = 72.8%
## GCV = 2.3875e+08  Scale est. = 2.1497e+08  n = 546
```

The p-value of lot size is 2^-16, so it is significant at the alpha=0.5 level.

3. Use the anova function to test whether the spline model is better than just modeling lotsize as a linear term.

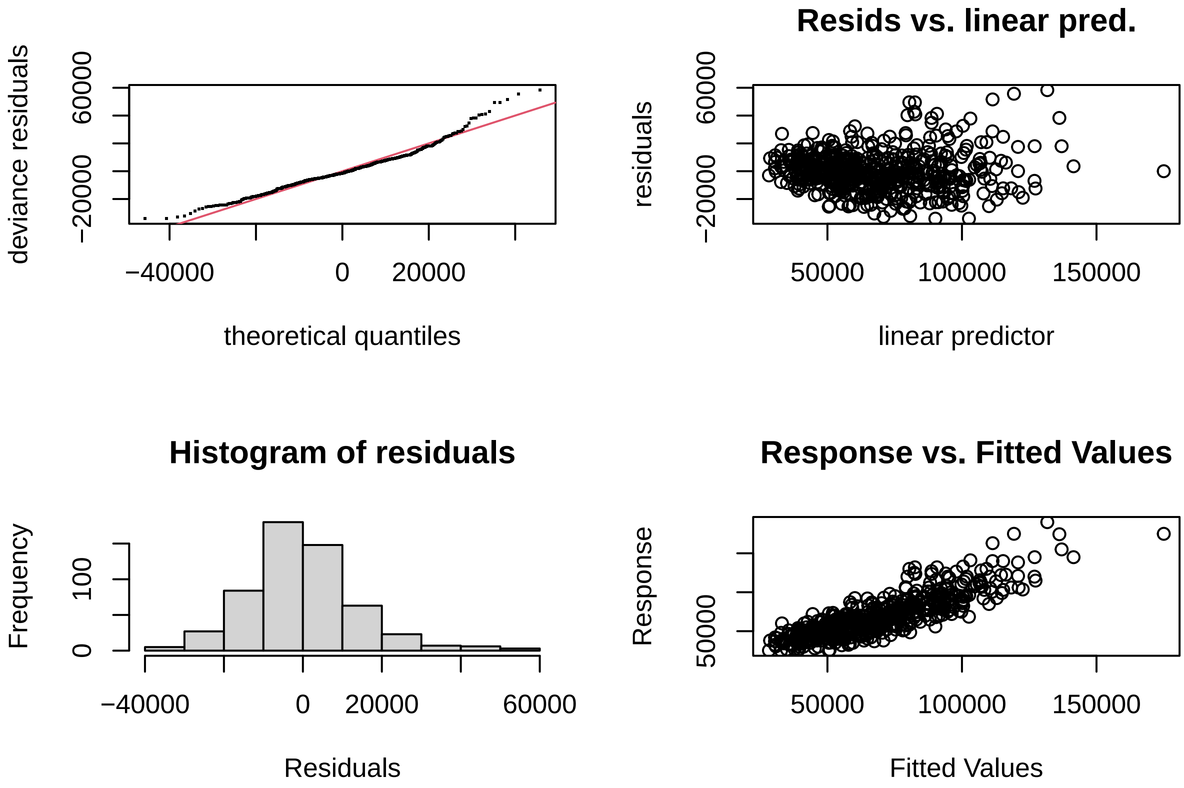
```
# Lotsize as linear term
fitGaussFull = gam(price ~ lotsize
                + bedrooms
                + factor(bathrooms)
                + factor(stories)
                + factor(driveway)
                + factor(recreation)
                + factor(fullbase)
                + factor(gasheat)
                + factor(aircon)
                + garage
                + factor(prefer),
                data = HousePrices,family = gaussian)

# Test lotsize spline vs no spline
anova(fitGaussAM,fitGaussFull,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: price ~ s(lotsize, bs = "cr", k = 105) + bedrooms + factor(bathrooms) +
##     factor(stories) + factor(driveway) + factor(recreation) +
##     factor(fullbase) + factor(gasheat) + factor(aircon) + garage +
##     factor(prefer)
## Model 2: price ~ lotsize + bedrooms + factor(bathrooms) + factor(stories) +
##     factor(driveway) + factor(recreation) + factor(fullbase) +
##     factor(gasheat) + factor(aircon) + garage + factor(prefer)
##   Resid. Df Resid. Dev      Df    Deviance  Pr(>Chi)
## 1    483.87 1.0569e+11
## 2    530.00 1.2617e+11 -46.129 -2.0483e+10 2.836e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of the anova is 2.84^-5 which is statistically significatn at the 0.05 level, indicating that the spline model is better than no spline.

4. Use gam.check(fitGaussAM) to see if you have enough basis functions. Do you?

```
# Check basis functions
gam.check(fitGaussAM)
```



```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 6 iterations.
## The RMS GCV score gradient at convergence was 1.477487 .
## The Hessian was positive definite.
## Model rank =  119 / 119
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                k'   edf k-index p-value
## s(lotsize) 104.0  39.4    1.08    0.97
```

The p-value is large at 0.97, indicating that there are enough basis functions present.

5. Consider a house with a lot size of 5000 square feet, three bedrooms, two bathrooms, two stories, a driveway, no recreation room, a finished basement, hot water heating, no air conditioning, two garage places, and located outside of the preferred neighborhood of Windsor, Canada. Predict the mean price of all houses under those constraints.

```
# Set up prediction parameters
example <- data.frame(lotsize = 5000,
                      bedrooms = 3,
                      bathrooms = 2,
                      stories = 2,
                      driveway = "yes",
                      recreation = "no",
                      fullbase = "yes",
                      gasheat = "yes",
                      aircon = "no",
                      garage = 2,
                      prefer = "no")

# Predict mean price of all houses with give constraints
predicted_price <- predict(fitGaussAM, newdata = example, type = "response", se.fit = TRUE)
mean = predicted_price$fit
se   = predicted_price$se.fit
mean
```

```
##        1
## 91617.43
```

The predicted mean housing price is $91,617.43.

6. Find an approximate 95% confidence interval for the mean price above.

```
# 95% CI for the mean housing price
n = length(HousePrices$price)
t = qt(0.975, df = n-2)
upperCI = mean + t * se
lowerCI = mean - t * se
c(lowerCI, upperCI)
```

```
##         1         1
##  81561.41 101673.45
```

The 95% CI for housing prices is ($81,561.41, $101,673.45)

7. Now restart R and and install the packages aplore3 and gam. Make sure you detach mgcv before doing this.

a. Add in library(aplore3); library(gam); data(icu); help(icu).

```
rm(list = ls())

detach("package:mgcv", unload=TRUE)
library(aplore3)
```

```
## Warning: package 'aplore3' was built under R version 4.3.3
```

```
library(gam)
```

```
## Warning: package 'gam' was built under R version 4.3.3
```

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

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

```
## Loaded gam 1.22-3
```

```
data(icu)
help(icu)
```

```
## starting httpd help server ...
```

```
##  done
```

8. Let Y be the response that the patient dies. The help file will help you figure out what this variable is. Run stepwise regression (step.Gam) with possible splines for age (age), heart rate (hra) and systolic blood pressure (sys), with 3 EDF. The binary predictor (no spline) is gender.

```
# Make Y numeric
icu$sta = as.numeric(icu$sta == "Died")
Y = icu$sta

# Run a baseline linear regression
baseGam = gam::gam(Y ~ as.factor(gender)
                   + age
                   + hra
                   + sys,
                   family = binomial,
                   data=icu)

# Run the stepwise regression
stepGam = step.Gam(baseGam,
                  scope =
                    list("gender" = ~1 + as.factor(gender),
                         "age" = ~1 + age + s(age,3),
                         "hra" = ~1 + hra + s(hra,3),
                         "sys" = ~1 + sys + s(sys,3)),
                  family = binomial,data = icu)
```

```
## Start:  Y ~ as.factor(gender) + age + hra + sys; AIC= 193.226
## Step:1 Y ~ as.factor(gender) + age + hra + s(sys, 3) ; AIC= 188.5008
## Step:2 Y ~ age + hra + s(sys, 3) ; AIC= 186.5004
## Step:3 Y ~ age + s(sys, 3) ; AIC= 184.5167
```

9. What is the indicated model?

   The indicated model (Step:3) includes age and systolic blood pressure as a spline [s(sys, 3)].

10. Detach the gam package and rerun using mgcv::gam with the indicated stepwise model: use k = 10 and bs=”cr”. What terms are statistically significant at the 0.05 level?

```
# Detach the gam package and load mgcv
detach("package:gam", unload=TRUE)
library(mgcv)
```

```
## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```
# Run stepwise using mgcv
mgcvStep = gam(Y ~ age
               + s(sys,k = 10,bs="cr"),
               family = binomial,
               data=icu)
summary(mgcvStep)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## Y ~ age + s(sys, k = 10, bs = "cr")
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.19528    0.74954  -4.263 2.02e-05 ***
## age          0.02806    0.01139   2.464   0.0137 *
## ---
## 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(sys) 4.893  5.942  18.11 0.00667 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.157   Deviance explained = 15.8%
## UBRE = -0.088085  Scale est. = 1         n = 200
```

Both age and systolic blood pressure are significant at the 0.05 level.