6.4 Case Study: Air Pollution in Chicago

Air Pollution in Chicago

The relationship between air pollution and health is a controversial topic and there is a great deal of epidemiological work attempting to elucidate the links.

In this exercise, the response of interest is the daily death rate in Chicago **`death`**, over a number of years. Possible explanatory variables are level of ozone **`o3median`**, levels of sulphur dioxide **`so2median`**, mean daily temperature, **`tmpd`**, and levels of particular matter, **`pm10median`** (as generated from diesel exhaust, for example). In addition, the death rate tends to vary with **`time`**.

```
library(gamair)
data(chicago)
```

Propose various GAM models for modelling the relationship between air pollution and health. Note that a conventional approach to modelling these data would be to assume that the observed numbers of deaths are Poisson random variables. Use the UBRE score for model selection.

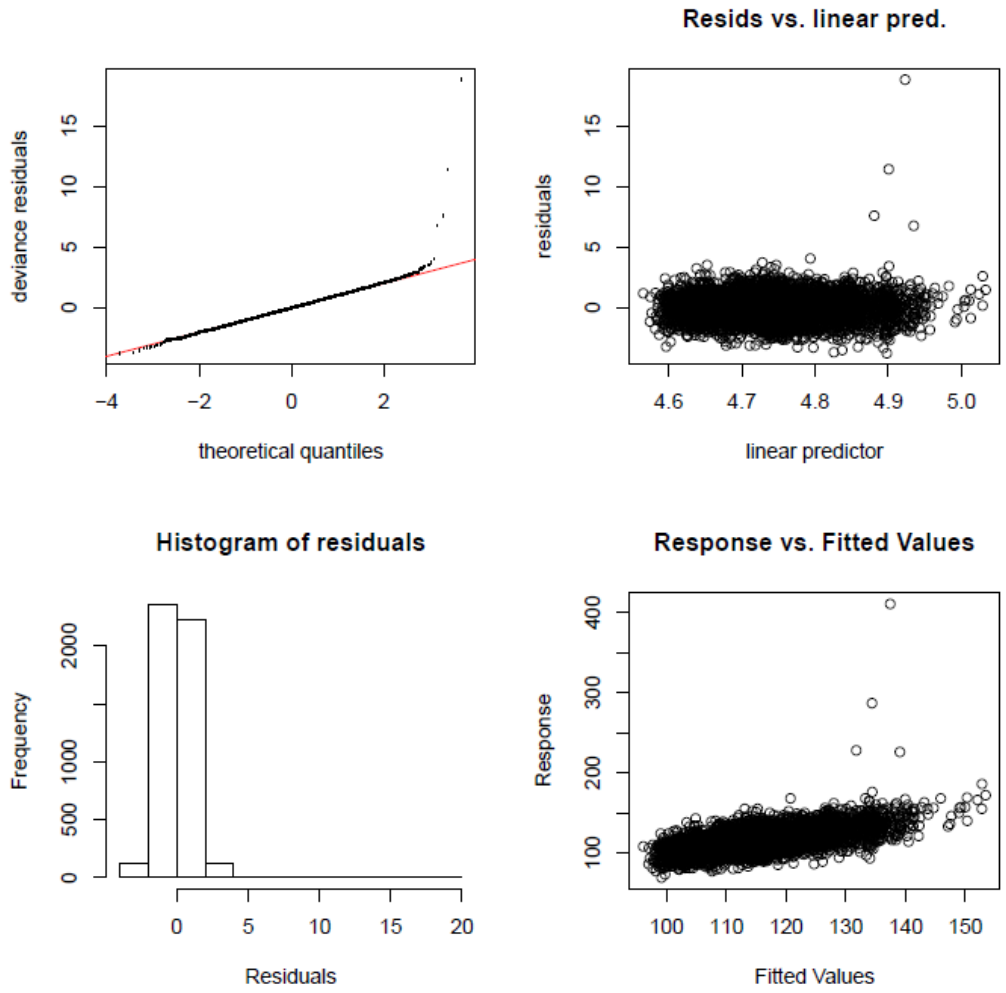
Model 1

Given here is a possible model.

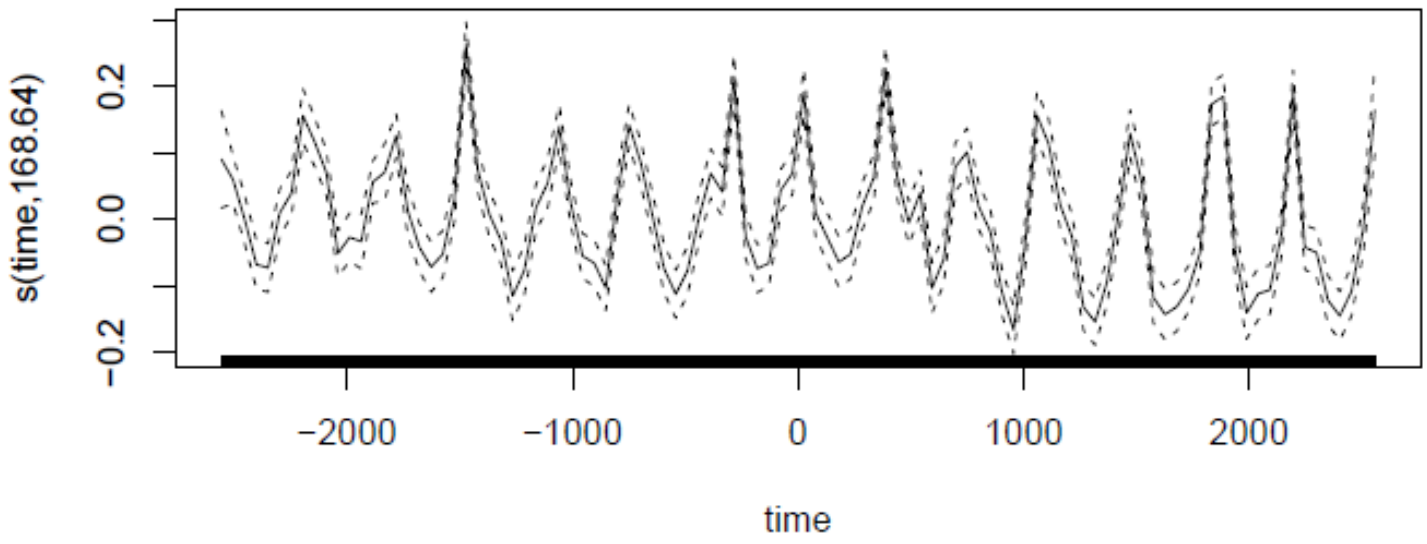
Run the code in the workspace and examine the outputs.

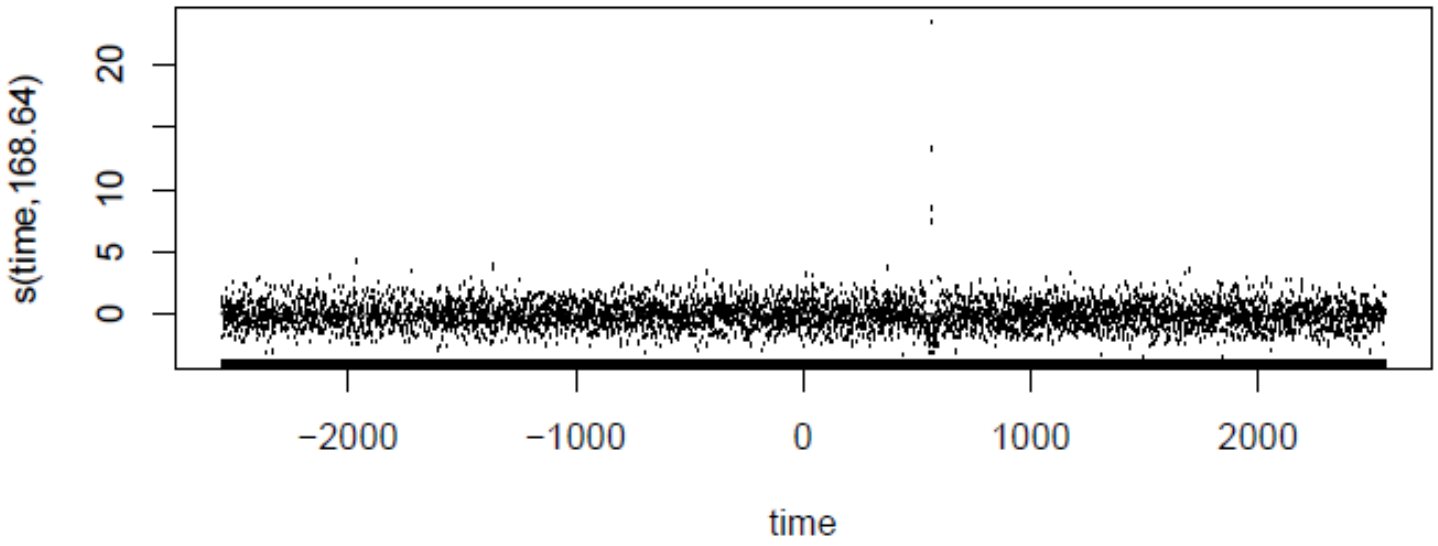
```
Method: UBRE Optimizer: outer newton
ull convergence after 3 iterations.
Gradient range [3.514595e-08,3.514595e-08]
(score 0.2546689 & scale 1).
Hessian positive definite, eigenvalue range [0.004247567,0.004247567].
Model rank = 204 / 204

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(time) 199 169    0.92  <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Family: poisson
Link function: log

Formula:
death ~ s(time, bs = "cr", k = 200) + pm10median + so2median +
    o3median + tmpd

Estimated degrees of freedom:
169 total = 173.64
UBRE score: 0.2546689
```



The plots show clear problems as a result of some substantial outliers. Plotting the estimated smooth without and with partial residuals emphasises the size of the outliers:





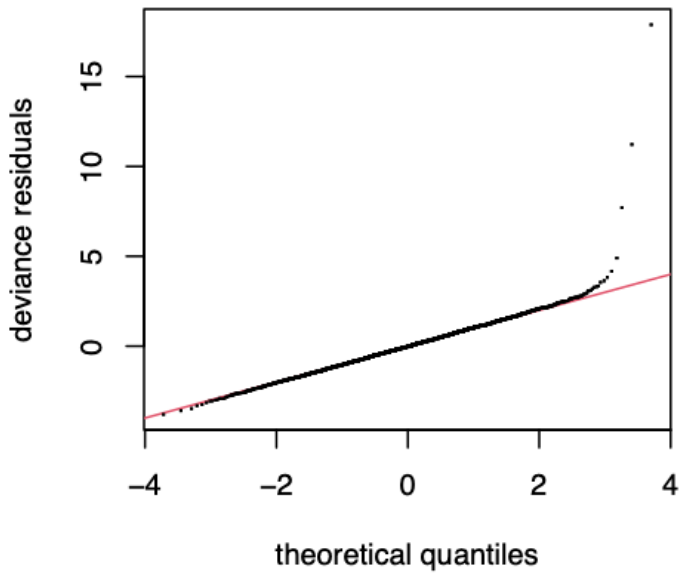
Examination of the data indicates that the outliers are the four highest daily death rates occurring in the data and that they occur on consecutive days. This peak is associated with a period of very high temperatures and high ozone.

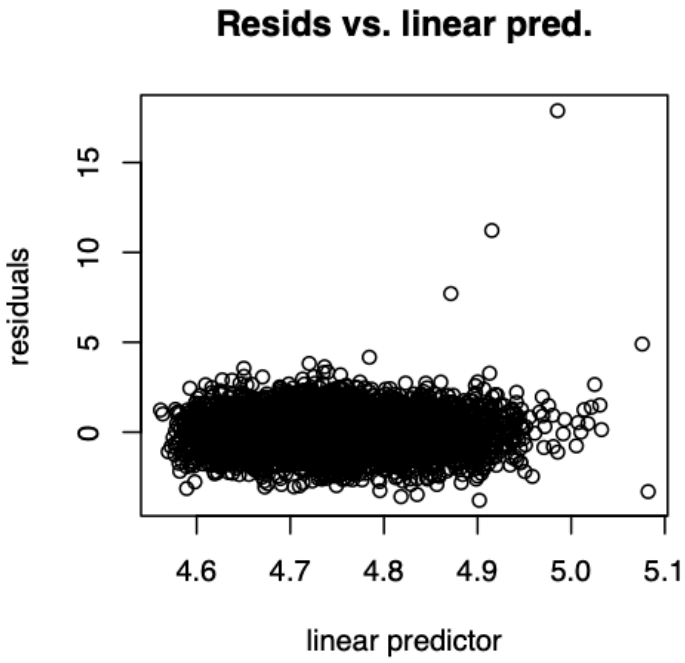
Model 2

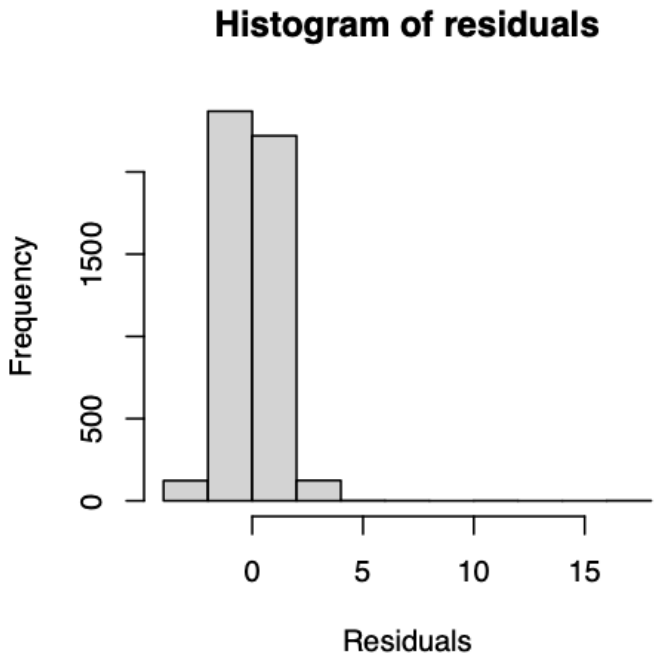
One immediate possibility for this issue is that model 1 is not flexible enough and that some non-linear response of death rate to temperature and ozone is required. We therefore consider a second model. This is given in the workspace.

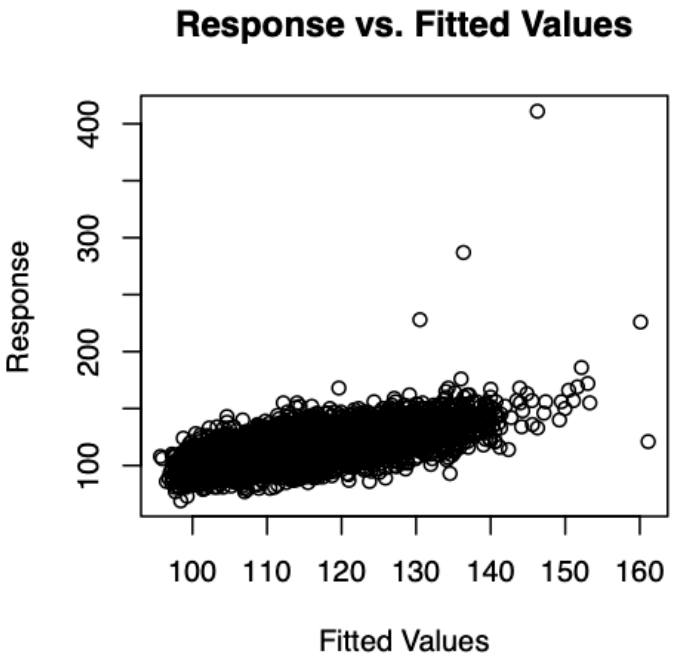
```
Method: UBRE   Optimizer: outer newton
full convergence after 8 iterations.
Gradient range [-6.661706e-07,4.670663e-08]
(score 0.2410737 & scale 1).
Hessian positive definite, eigenvalue range [5.217842e-05,0.004273638].
Model rank =  236 / 236

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(time)       199.00 167.93    0.94  <2e-16 ***
s(pm10median)   9.00   6.86    1.02    0.90
s(so2median)    9.00   7.38    0.99    0.20
s(o3median)     9.00   1.58    1.00    0.36
s(tmpd)         9.00   8.27    1.02    0.95
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Family: poisson
Link function: log

Formula:
death ~ s(time, bs = "cr", k = 200) + s(pm10median, bs = "cr") +
    s(so2median, bs = "cr") + s(o3median, bs = "cr") + s(tmpd,
    bs = "cr")

Estimated degrees of freedom:
167.93   6.86   7.38   1.58   8.27  total = 193.03
UBRE score: 0.2410737
```









This model is also not appropriate. The diagnostic plots are similar as in the case of model 1.

Examination of the data

More detail examination of the data shows that the highest temperatures were recorded in the few days **preceding** the high mortalities, when there were also high ozone levels recorded. We, therefore, need to produce lagged variables in the following way:

$$\texttt{o3}_\texttt{i} = \sum_{j=i-3}^{i} \texttt{o3median}_\texttt{j}$$

with similar definitions for the other predictor variables.

Model 3

Given the suggestion from the data that a combination of high ozone and high temperature might lead to very high death rates a third model was tried.

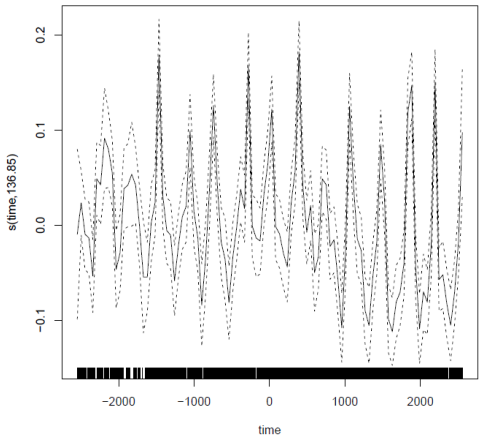
```
Family: poisson
Link function: log

Formula:
death ~ s(time, bs = "cr", k = 200) +
        te(o3, tmpd, k = 8) +
        te(pm10, tmpd, k = 6)

Estimated degrees of freedom:
136.37  36.55   8.18  total = 182.1

UBRE score: 0.1585013
```

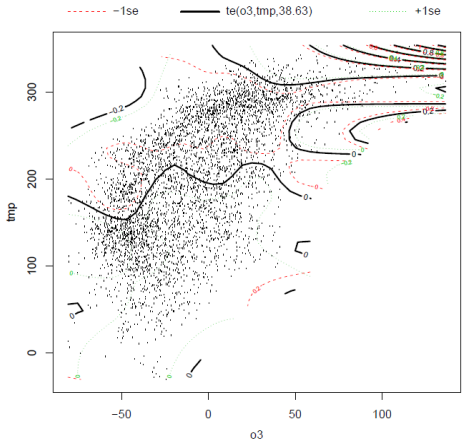
This model produces much better results. Having reached the stage of having a model that is not obviously wrong, it is worth proceeding to see if it can be simplified.

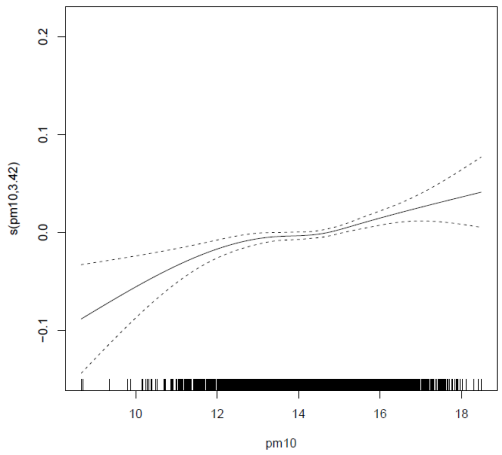
Model 4

Following the results from model 3, consider this simplified model.

```
Family: poisson
Link function: log

Formula:
death ~ s(time, bs = "cr", k = 200) +
        te(o3, tmpd, k = 8) +
        s(pm10, bs = "cr", k = 6)

Estimated degrees of freedom:
136.85  38.63   3.42  total = 179.91

UBRE score: 0.1587634
```







Based on the UBRE score the most favourable model is `m3`. Further experimentation with other models worsens the fit. As it can be visible from the final figure the notion that several days of high ozone and temperature can cause elevated death rates can explain these data, but given the way that the data have been used to develop a model, we would really need to see if the same model works well in other locations or time periods, before giving it too much credence.