A Handbook of Statistical Analyses Using R — 2nd Edition

Brian S. Everitt and Torsten Hothorn

CHAPTER 10

Scatterplot Smoothers and Generalised Additive Models: The Men<U+2019>s Olympic 1500m, Air Pollution in the USA, and Risk Factors for Kyphosis

10.1 Introduction

10.2 Scatterplot Smoothers and Generalised Additive Models

10.3 Analysis Using R

10.3.1 Olympic 1500m Times

To begin we will construct a scatterplot of winning time against year the games were held. The R code and the resulting plot are shown in Figure 10.1. There is very clear downward trend in the times over the years, and, in addition there is a very clear outlier namely the winning time for 1896. We shall remove this time from the data set and now concentrate on the remaining times. First we will fit a simple linear regression to the data and plot the fit onto the scatterplot. The code and the resulting plot are shown in Figure 10.2. Clearly the linear regression model captures in general terms the downward trend in the times. Now we can add the fits given by the lowess smoother and by a cubic spline smoother; the resulting graph and the extra R code needed are shown in Figure 10.3.

Both non-parametric fits suggest some distinct departure from linearity, and clearly point to a quadratic model being more sensible than a linear model here. And fitting a parametric model that includes both a linear and a quadratic effect for year gives a prediction curve very similar to the non-parametric curves; see Figure 10.4.

Here use of the non-parametric smoothers has effectively diagnosed our linear model and pointed the way to using a more suitable parametric model; this is often how such non-parametric models can be used most effectively. For these data, of course, it is clear that the simple linear model cannot be suitable if the investigator is interested in predicting future times since even the most basic knowledge of human physiology will tell us that times cannot continue to go down. There must be some lower limit to the time man can run 1500m. But in other situations use of the non-parametric smoothers may point to a parametric model that could not have been identified *a priori*.

It is of some interest to look at the predictions of winning times in future

```
R> plot(time ~ year, data = men1500m)
```

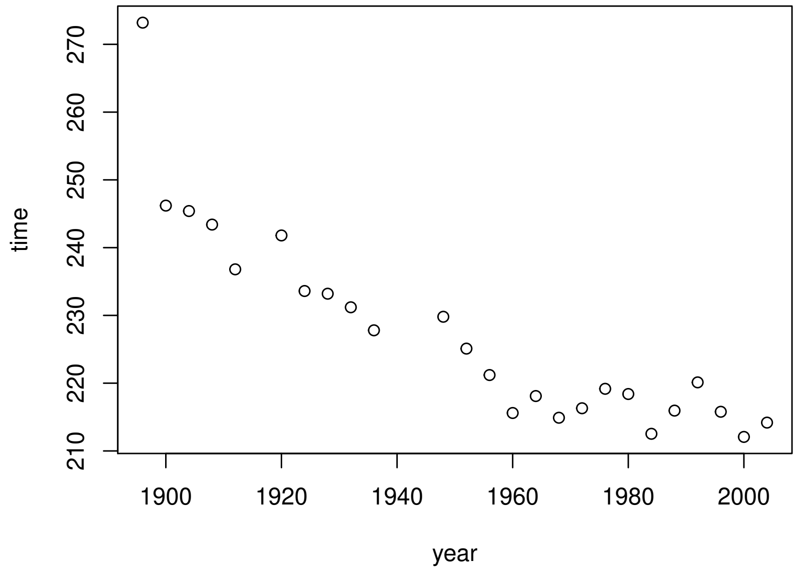

Figure 10.1 Scatterplot of year and winning time.

Olympics from both the linear and quadratic models. For example, for 2008 and 2012 the predicted times and their 95% confidence intervals can be found using the following code

```
R> predict(men1500m_lm,
+          newdata = data.frame(year = c(2008, 2012)),
+          interval = "confidence")
```

```
  fit lwr upr
1 208 205 211
2 207 203 210
```

```
R> predict(men1500m_lm2,
+          newdata = data.frame(year = c(2008, 2012)),
+          interval = "confidence")
```

```
  fit lwr upr
1 214 210 218
2 214 210 219
```

For predictions far into the future both the quadratic and the linear model fail; we leave readers to get some more predictions to see what happens. We can compare the first prediction with the time actually recorded by the winner

```
R> men1500m1900 <- subset(men1500m, year >= 1900)
R> men1500m_lm <- lm(time ~ year, data = men1500m1900)
R> plot(time ~ year, data = men1500m1900)
R> abline(men1500m_lm)
```

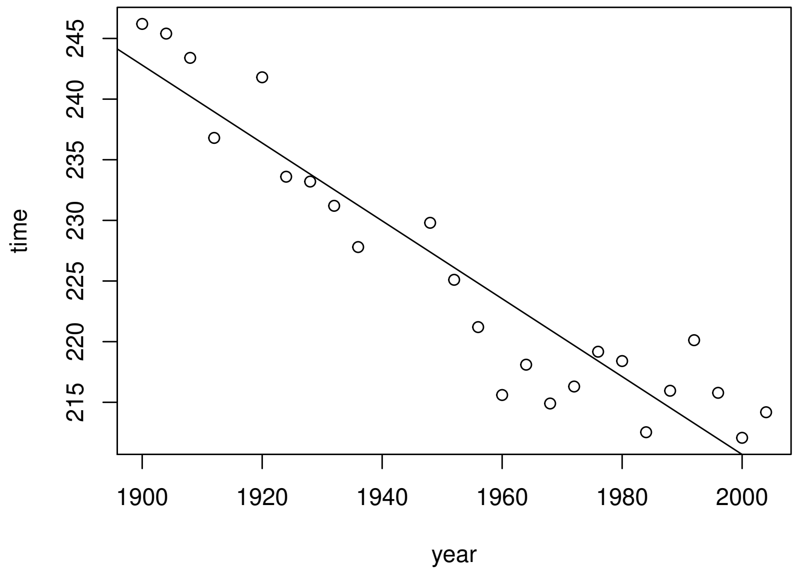

Figure 10.2 Scatterplot of year and winning time with fitted values from a simple linear model.

of the men's 1500m in Beijing 2008, Rashid Ramzi from Brunei, who won the event in 212.94 seconds. The confidence interval obtained from the simple linear model does not include this value but the confidence interval for the prediction derived from the quadratic model does.

10.3.2 Air Pollution in US Cities

Unfortunately, we cannot fit an additive model for describing the SO_2 concentration based on all six covariates because this leads to more parameters than cities, i.e., more parameters than observations when using the default parameterisation of **mgcv**. Thus, before we can apply the `gam` function from package **mgcv**, we have to decide which covariates should enter the model and which subset of these covariates should be allowed to deviate from a linear regression relationship.

```
R> x <- men1500m1900$year
R> y <- men1500m1900$time
R> men1500m_lowess <- lowess(x, y)
R> plot(time ~ year, data = men1500m1900)
R> lines(men1500m_lowess, lty = 2)
R> men1500m_cubic <- gam(y ~ s(x, bs = "cr"))
R> lines(x, predict(men1500m_cubic), lty = 3)
```

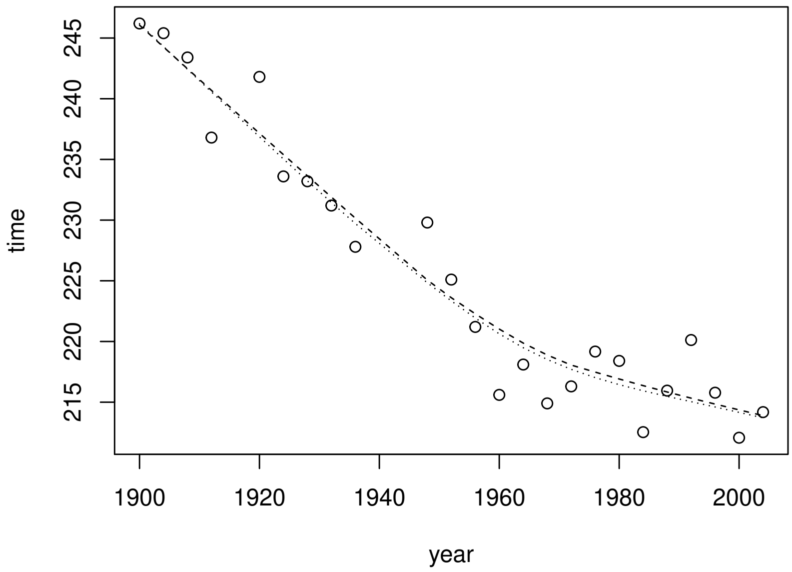

Figure 10.3 Scatterplot of year and winning time with fitted values from a smooth non-parametric model.

As briefly discussed in Section **??**, we can fit an additive model using the iterative boosting algorithm as described by Bühlmann and Hothorn (2007). The complexity of the model is determined by an AIC criterion, which can also be used to determine an appropriate number of boosting iterations to choose. The methodology is available from package **mboost** (Hothorn et al., 2012). We start with a small number of boosting iterations (100 by default) and compute the AIC of the corresponding 100 models:

```
R> library("mboost")
R> USair_boost <- gamboost(SO2 ~ ., data = USairpollution)
R> USair_aic <- AIC(USair_boost)
R> USair_aic
```

```
R> men1500m_lm2 <- lm(time ~ year + I(year^2),
+                     data = men1500m1900)
R> plot(time ~ year, data = men1500m1900)
R> lines(men1500m1900$year, predict(men1500m_lm2))
```

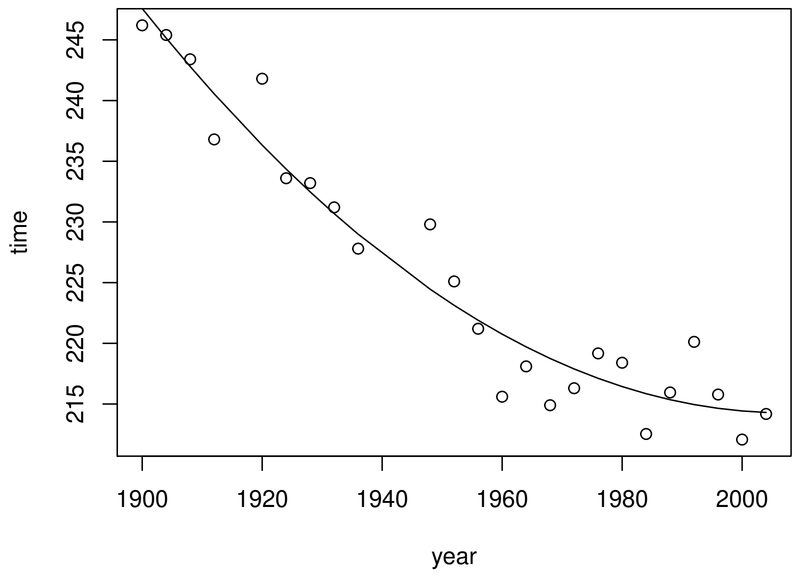

Figure 10.4 Scatterplot of year and winning time with fitted values from a quadratic model.

```
[1] 6.77
Optimal number of boosting iterations: 47
Degrees of freedom (for mstop = 47): 8.31
```

The AIC suggests that the boosting algorithm should be stopped after 47 iterations. The partial contributions of each covariate to the predicted SO_2 concentration are given in Figure 10.5. The plot indicates that all six covariates enter the model and the selection of a subset of covariates for modelling isn't appropriate in this case. However, the number of manufacturing enterprises seems to add linearly to the SO_2 concentration, which simplifies the model. Moreover, the average annual precipitation contribution seems to deviate from zero only for some extreme observations and one might refrain from using the covariate at all.

As always, an inspection of the model fit via a residual plot is worth the effort. Here, we plot the fitted values against the residuals and label the

```
R> USair_gam <- USair_boost[mstop(USair_aic)]
R> layout(matrix(1:6, ncol = 3))
R> plot(USair_gam, ask = FALSE)
```

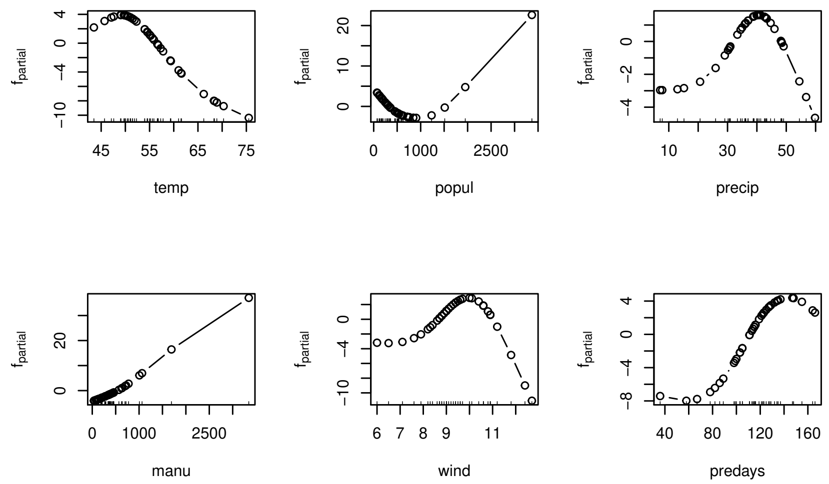

Figure 10.5 Partial contributions of six exploratory covariates to the predicted SO_2 concentration.

points with the name of the corresponding city using the `textplot` function from package **wordcloud**. Figure 10.6 shows at least two extreme observations. Chicago has a very large observed and fitted SO_2 concentration, which is due to the huge number of inhabitants and manufacturing plants (see Figure 10.5 also). One smaller city, Providence, is associated with a rather large positive residual indicating that the actual SO_2 concentration is underestimated by the model. In fact, this small town has a rather high SO_2 concentration which is hardly explained by our model. Overall, the model doesn't fit the data very well, so we should avoid overinterpreting the model structure too much. In addition, since each of the six covariates contributes to the model, we aren't able to select a smaller subset of the covariates for modelling and thus fitting a model using `gam` is still complicated (and will not add much knowledge anyway).

10.3.3 Risk Factors for Kyphosis

Before modelling the relationship between kyphosis and the three exploratory variables age, starting vertebral level of the surgery and number of vertebrae involved, we investigate the partial associations by so-called *spinograms*, as

```
R> SO2hat <- predict(USair_gam)
R> SO2 <- USairpollution$SO2
R> plot(SO2hat, SO2 - SO2hat, type = "n",
+       xlim = c(-30, 110), ylim = c(-30, 60))
R> textplot(SO2hat, SO2 - SO2hat, rownames(USairpollution),
+           show.lines = FALSE, new = FALSE)
R> abline(h = 0, lty = 2, col = "grey")
```

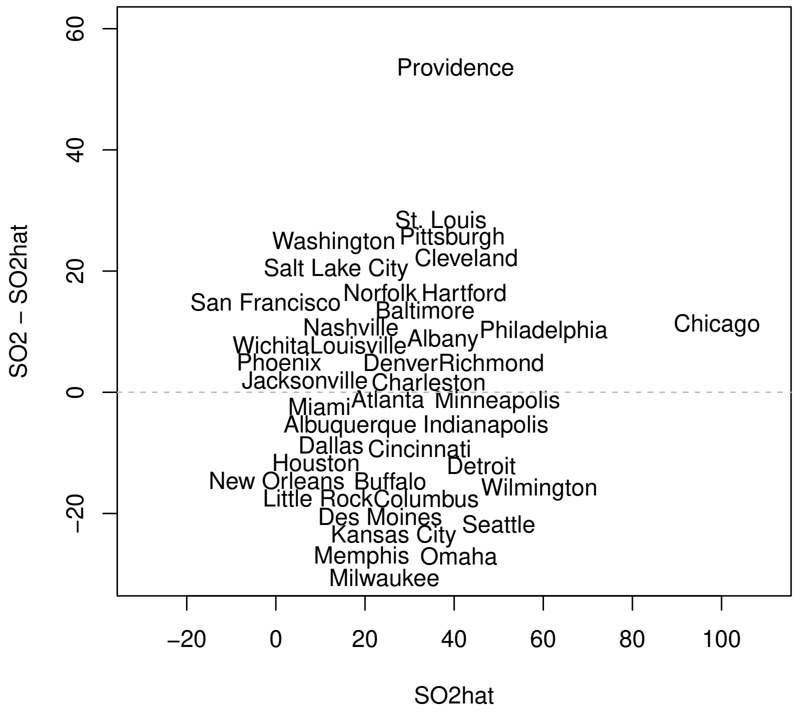

Figure 10.6 Residual plot of SO_2 concentration.

```
R> layout(matrix(1:3, nrow = 1))
R> spineplot(Kyphosis ~ Age, data = kyphosis,
+            ylevels = c("present", "absent"))
R> spineplot(Kyphosis ~ Number, data = kyphosis,
+            ylevels = c("present", "absent"))
R> spineplot(Kyphosis ~ Start, data = kyphosis,
+           ylevels = c("present", "absent"))
```

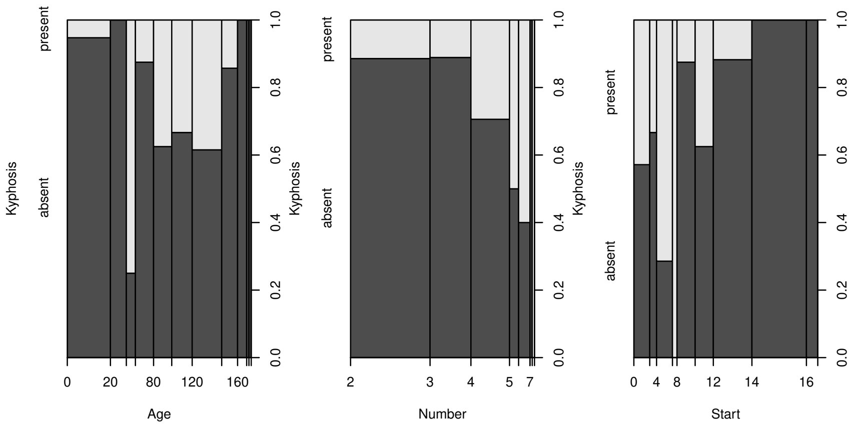

Figure 10.7 Spinograms of the three exploratory variables and response variable `kyphosis`.

introduced in Chapter 2. The numeric exploratory covariates are discretised and their empirical relative frequencies are plotted against the conditional frequency of kyphosis in the corresponding group. Figure 10.7 shows that kyphosis is absent in very young or very old children, children with a small starting vertebral level and high number of vertebrae involved.

The logistic additive model needed to describe the conditional probability of kyphosis given the exploratory variables can be fitted using function `gam`. Here, the dimension of the basis (k) has to be modified for `Number` and `Start` since these variables are heavily tied. As for generalised linear models, the `family` argument determines the type of model to be fitted, a logistic model in our case:

```
R> kyphosis_gam <- gam(Kyphosis ~ s(Age, bs = "cr") +
+       s(Number, bs = "cr", k = 3) + s(Start, bs = "cr", k = 3),
+       family = binomial, data = kyphosis)
R> kyphosis_gam
```

```
Family: binomial
Link function: logit

Formula:
```

```
R> trans <- function(x)
+       binomial()$linkinv(x)
R> layout(matrix(1:3, nrow = 1))
R> plot(kyphosis_gam, select = 1, shade = TRUE, trans = trans)
R> plot(kyphosis_gam, select = 2, shade = TRUE, trans = trans)
R> plot(kyphosis_gam, select = 3, shade = TRUE, trans = trans)
```

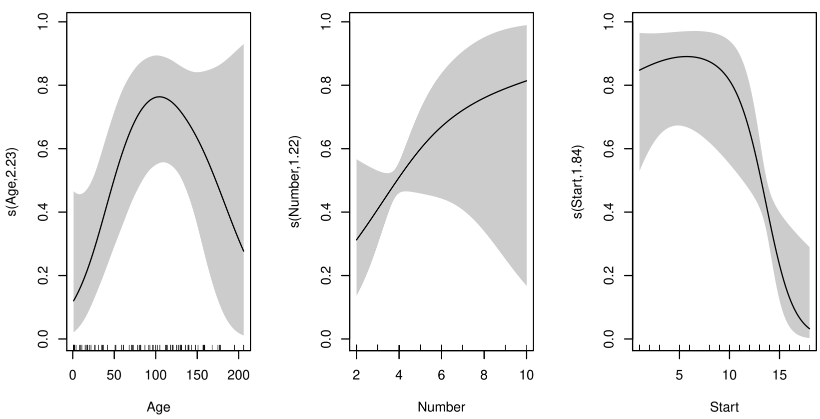

Figure 10.8 Partial contributions of three exploratory variables with confidence bands.

```
Kyphosis ~ s(Age, bs = "cr") + s(Number, bs = "cr", k = 3) +
    s(Start, bs = "cr", k = 3)

Estimated degrees of freedom:
2.23 1.22 1.84  total = 6.29

UBRE score: -0.234
```

The partial contributions of each covariate to the conditional probability of kyphosis with confidence bands are shown in Figure 10.8. In essence, the same conclusions as drawn from Figure 10.7 can be stated here. The risk of kyphosis being present decreases with higher starting vertebral level and lower number of vertebrae involved. Children about 100 months old are under higher risk compared to younger or older children.

Summary

Additive models offer flexible modelling tools for regression problems. They stand between generalised linear models, where the regression relationship is assumed to be linear, and more complex models like random forests (see Chapter 9) where the regression relationship remains unspecified. Smooth functions describing the influence of covariates on the response can be easily interpreted.

Variable selection is a technically difficult problem in this class of models; boosting methods are one possibility to deal with this problem.

Exercises

Ex. 10.1 Consider the body fat data introduced in Chapter 9, Table ??. First fit a generalised additive model assuming normal errors using function `gam`. Are all potential covariates informative? Check the results against a generalised additive model that underwent AIC-based variable selection (fitted using function `gamboost`).

Ex. 10.2 Try to fit a logistic additive model to the glaucoma data discussed in Chapter 9. Which covariates should enter the model and how is their influence on the probability of suffering from glaucoma?

Bibliography

Bühlmann, P. and Hothorn, T. (2007), "Boosting algorithms: Regularization, prediction and model fitting," *Statistical Science*, 22, 477–505.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2012), ***mboost**: Model-Based Boosting*, URL `http://CRAN.R-project.org/package=mboost`, R package version 2.1-2.