Semiparametric regression - Homework 6

Klaudia Weigel

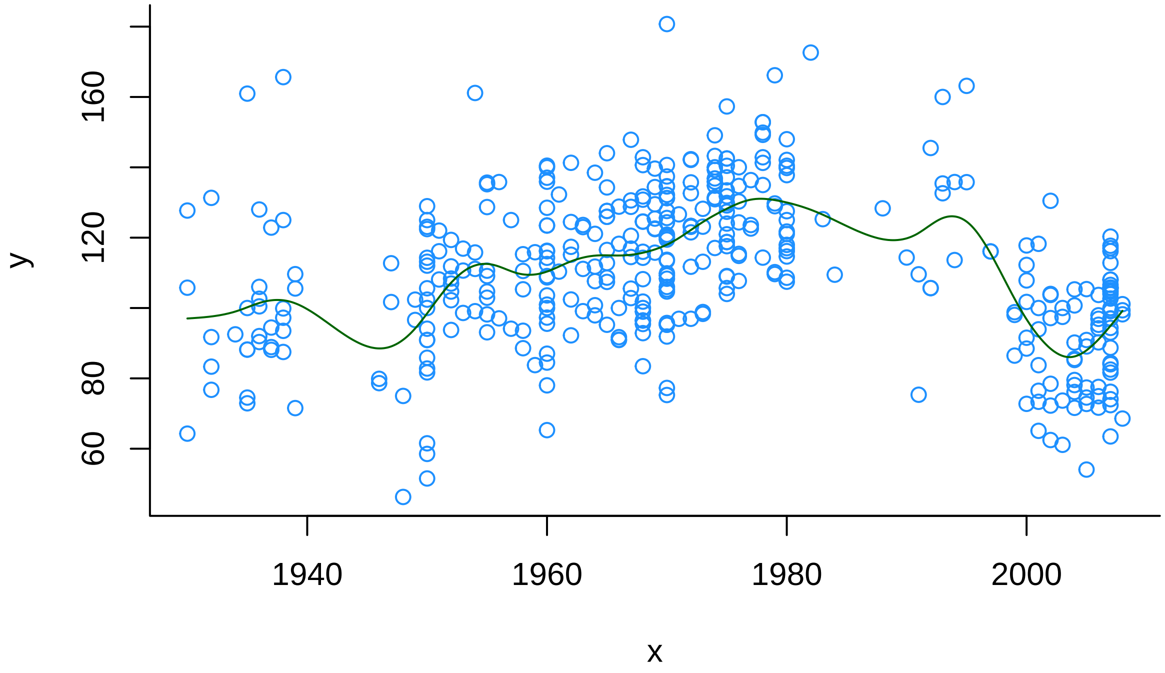
1 Exercise 1

1.1 (a)

We will first fit a penalized spline model to `WarsawApts` data. We will fit the model with 30 cubic spline basis functions and smoothing parameter chosen via generalized cross validation (GCV) method.

```r
library(HRW); library(nlme); library(mgcv)
library(ggplot2); library(tidyr)

data(WarsawApts)
x <- WarsawApts$construction.date
y <- WarsawApts$areaPerMzloty
plot(x, y, bty = 'l', col = 'dodgerblue')
fitGAMcr <- gam(y~s(x, bs = 'cr', k = 30))
xg <- seq(min(x), max(x), length = 1001)
fHatgGAMcr <- predict(fitGAMcr, newdata = data.frame(x = xg))
lines(xg, fHatgGAMcr, col = 'darkgreen')
```



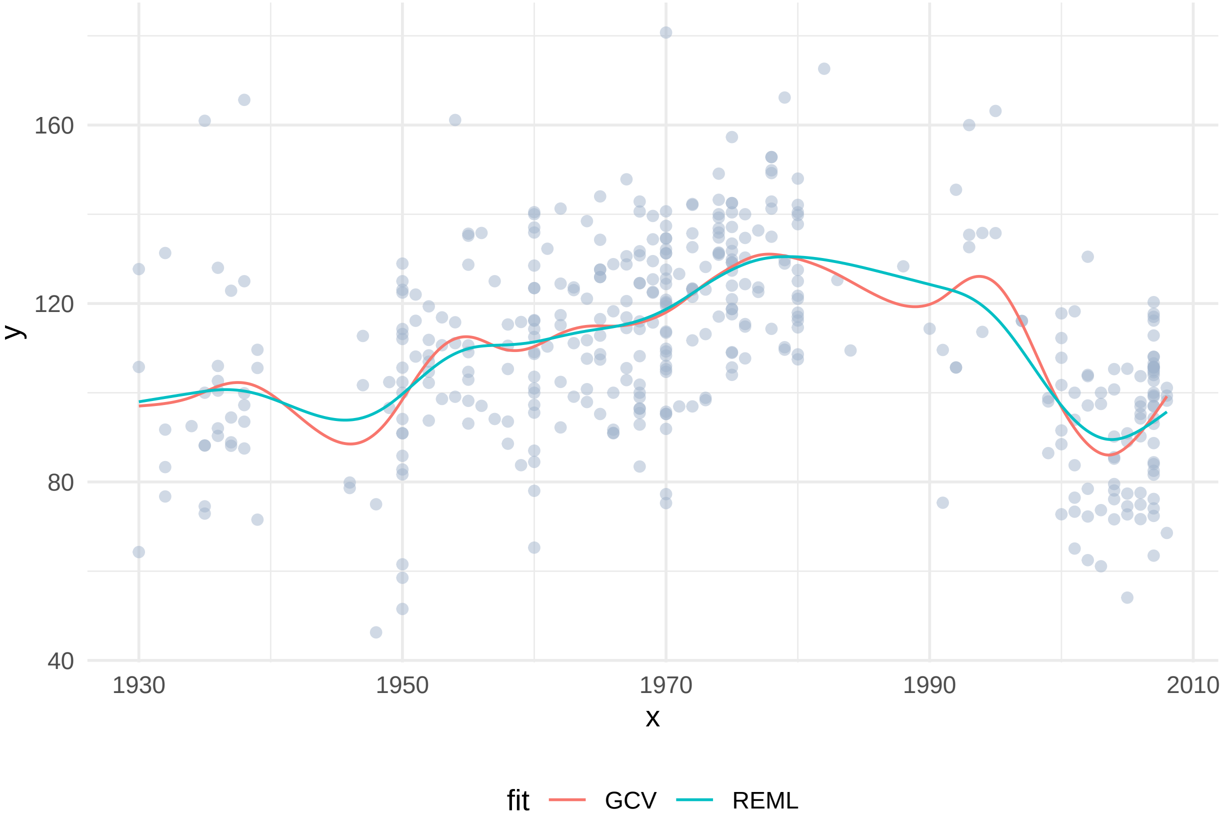
1.2 (b)

Alternatively restricted maximum likelihood (REML) may be used for smoothness selection.

```
fitGAMcrREML <- gam(y ~ s(x, bs = "cr", k = 30), method = "REML")
fHatgGAMcrREML <- predict(fitGAMcrREML, newdata = data.frame(x = xg))
```

Let us plot both fit.

```
data_plot <- data.frame(cbind(xg, fHatgGAMcr, fHatgGAMcrREML))
data_plot %>%
  gather('fit','value', -xg) %>%
  ggplot() +
  geom_point(data = data.frame(cbind(x,y)), aes(x = x, y = y),
             color = 'lightsteelblue3', alpha=0.5) +
  geom_line(aes(x = xg, y = value, color = fit), size = 0.5) +
  scale_color_hue(labels = c("GCV", "REML")) +
  theme_minimal() +
  theme(legend.position = 'bottom')
```



1.3 (c)

We see in the plot above that the REML-based fit is more smooth and has less wiggle than the GCV-based fit. It seems that the REML fit is less prone to overfitting and undersmoothing, whereas the GCV fit is quite sensitive to potential outliers. Overall we can conclude that the choice of the smoothing parameter has a significant impact on the fit.