Semiparametric regression - Homework 2

Klaudia Weigel

1 Exercise 1

1.1 a)

In this exercise we are going to use `WarsawApts` data, which contains information about apartment prices in Warsaw.

```
library(HRW); data(WarsawApts)
head(WarsawApts)
```

```
##   surface    district n.rooms floor construction.date areaPerMzloty
## 1      20 Srodmiescie       1     7              1970      95.23810
## 2      27        Wola       1     1              1962     117.39845
## 3      28     Mokotow       1     1              1950     114.28571
## 4      28     Mokotow       2     4              1968     114.28571
## 5      30 Srodmiescie       1     1              1952      93.75586
## 6      32     Mokotow       2     3              2007      81.63265
```

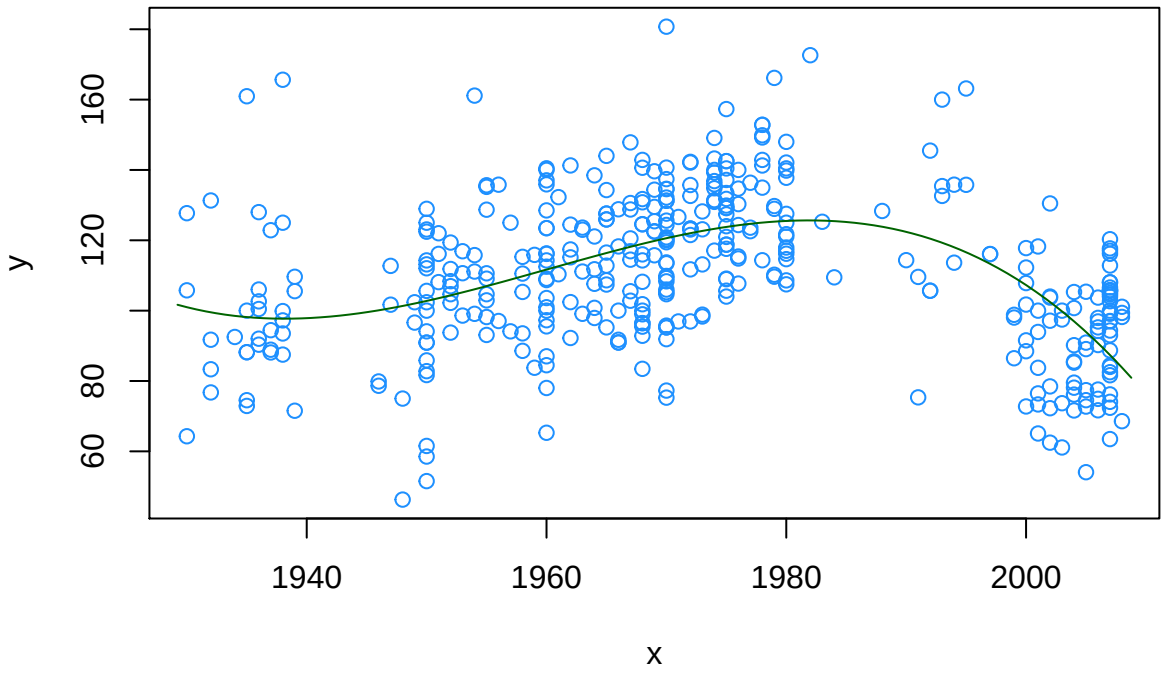
We are going to fit a cubic model in order to predict the price:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i.$$

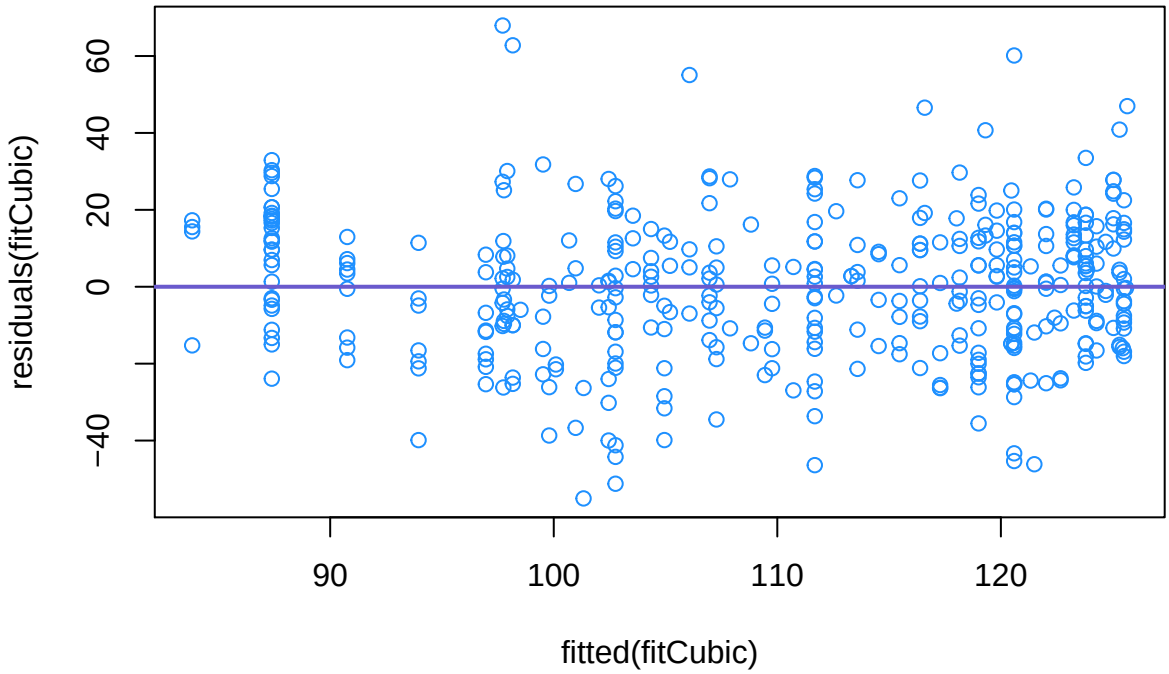
where y_i is the price in złoty and x_i is the construction date.

```
x <-  WarsawApts$construction.date
y <- WarsawApts$areaPerMzloty
fitCubic <- lm(y~poly(x, 3, raw = TRUE))
ng <- 101
xg <- seq(1.01*min(x) - 0.01*max(x), 1.01*max(x) - 0.01*min(x), length = ng)
fHatCubicg <- as.vector(cbind(rep(1,ng), xg, xg^2, xg^3)%*%fitCubic$coef)
```

```
plot(x, y, col = 'dodgerblue')
lines(xg, fHatCubicg, col = 'darkgreen')
```




```
plot(fitted(fitCubic), residuals(fitCubic), col = 'dodgerblue')
abline(0, 0, col = 'slateblue', lwd = 2)
```



From the first plot we can see that the model follows the general trend of the data. We know that in the residuals versus fits plot the points should be randomly scattered around the x axis. Some pattern in the plot might suggest that the variances of the errors are not equal or that the fitted cubic model is inadequate to the data. Here there seems to be no visible pattern in the plot, therefore we have no reason to think that the assumption about the constant variance of the errors is violated. The plot indicates the presence of potential outliers. As we can see, most of the residuals are in the range of -40 and 40, however, there are a few observations whose residuals are around 60.
From both plots we can conclude that the cubic model is a reasonable fit to the data.

1.2 b)

We define the following truncated function

$$(x-\kappa)_+ = \begin{cases} 0 & x \leq \kappa \\ x-\kappa & x \geq \kappa \end{cases}.$$

```
trLin <- function(x, kappa) return((x-kappa)*(x>kappa))
```

1.3 c)

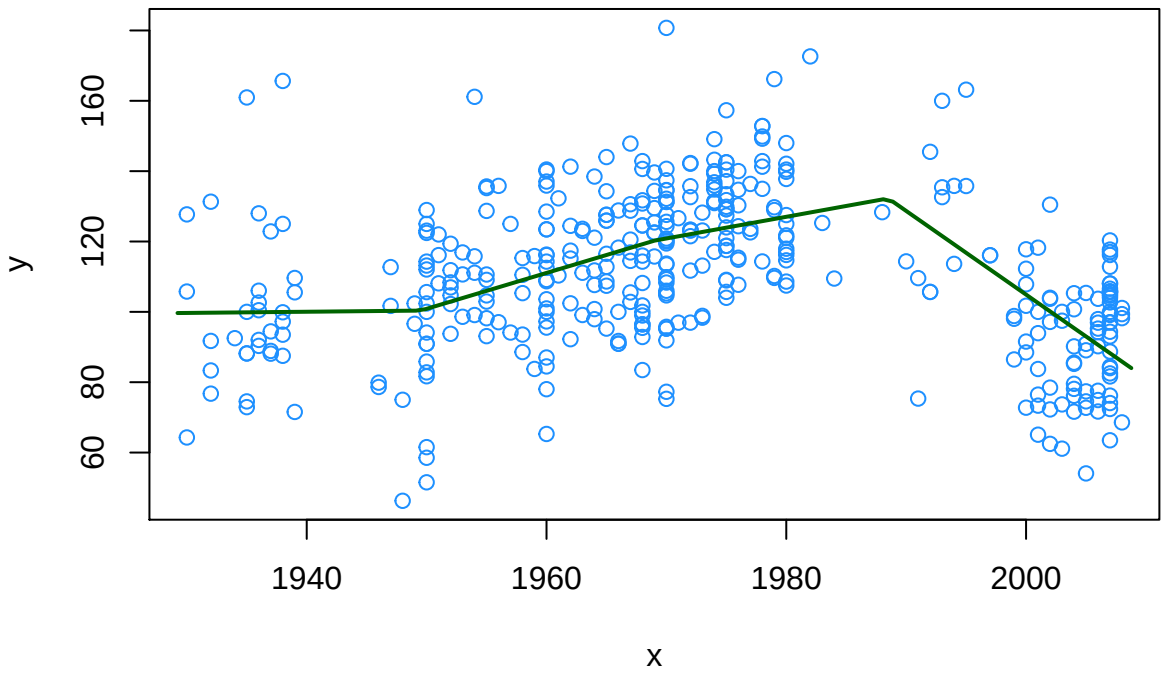
We now consider a spline regression model

$$y_i = \beta_0 + \beta_1 x_i + u_1(x_i-\kappa_1)_+ + u_2(x_i-\kappa_2)_+ + u_3(x_i-\kappa_3)_+ + \epsilon_i,$$

where $\kappa_1, \kappa_2, \kappa_3$ are equally spaced knots over the range of x_is.

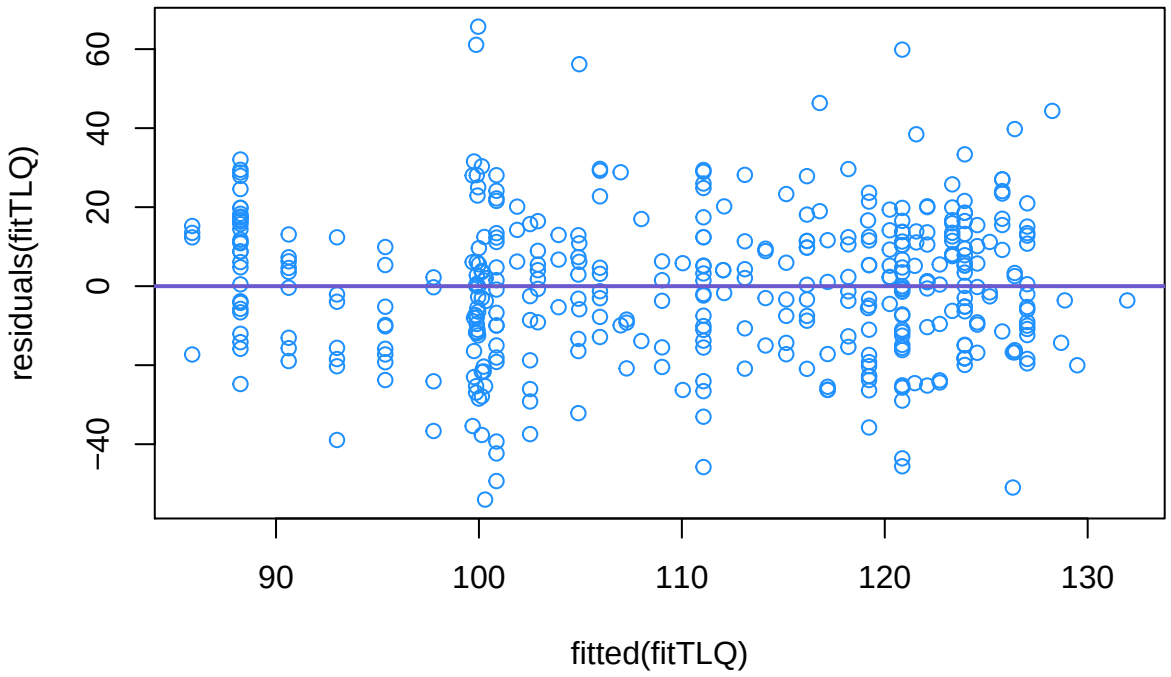
```
knots <- seq(min(x), max(x), length = 5)[-c(1,5)]
X <- cbind(1,x)
for(k in 1:3) X <- cbind(X, trLin(x, knots[k]))
fitTLQ <- lm(y ~ -1 + X)
Xg <- cbind(1, xg)
for (k in 1:3) Xg <- cbind(Xg, trLin(xg, knots[k]))
fHatTLQg <- as.vector(Xg%*%fitTLQ$coef)
```

```
plot(x, y, col = 'dodgerblue')
lines(xg, fHatTLQg, col = 'darkgreen', lwd = 2)
```



As in the previous point the shape of the fitted line follows the general shape of our data.

```
plot(fitted(fitTLQ), residuals(fitTLQ), col = 'dodgerblue')
abline(0, 0, col = 'slateblue', lwd = 2)
```



This plot is also similar to the one obtained in point b). Points seem to be scattered randomly around the x axis, and there is no visible pattern. There is therefore no reason to think that the assumption of homoscedasticity is broken or that the model is inadequate.

1.4 d)

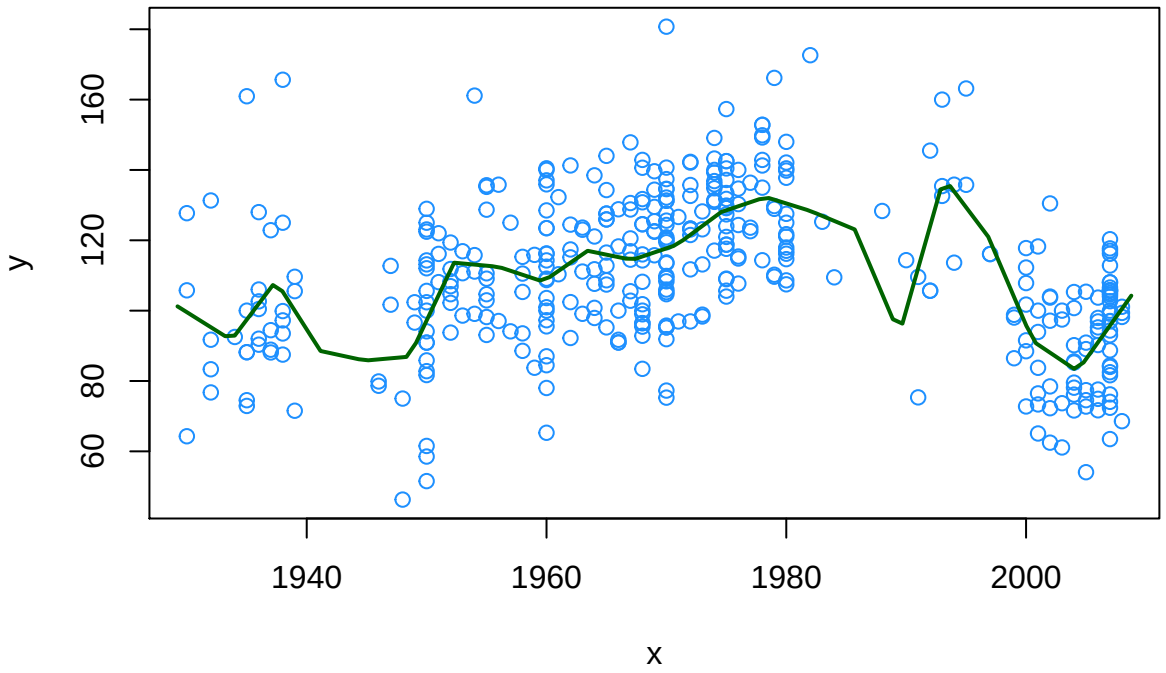
We will now fit the following spline regression model

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{20} u_k (x_i - \kappa_k)_+ + \epsilon_i,$$

where $\kappa_1, \ldots, \kappa_{20}$ are equally spaced knots over the range of the data.

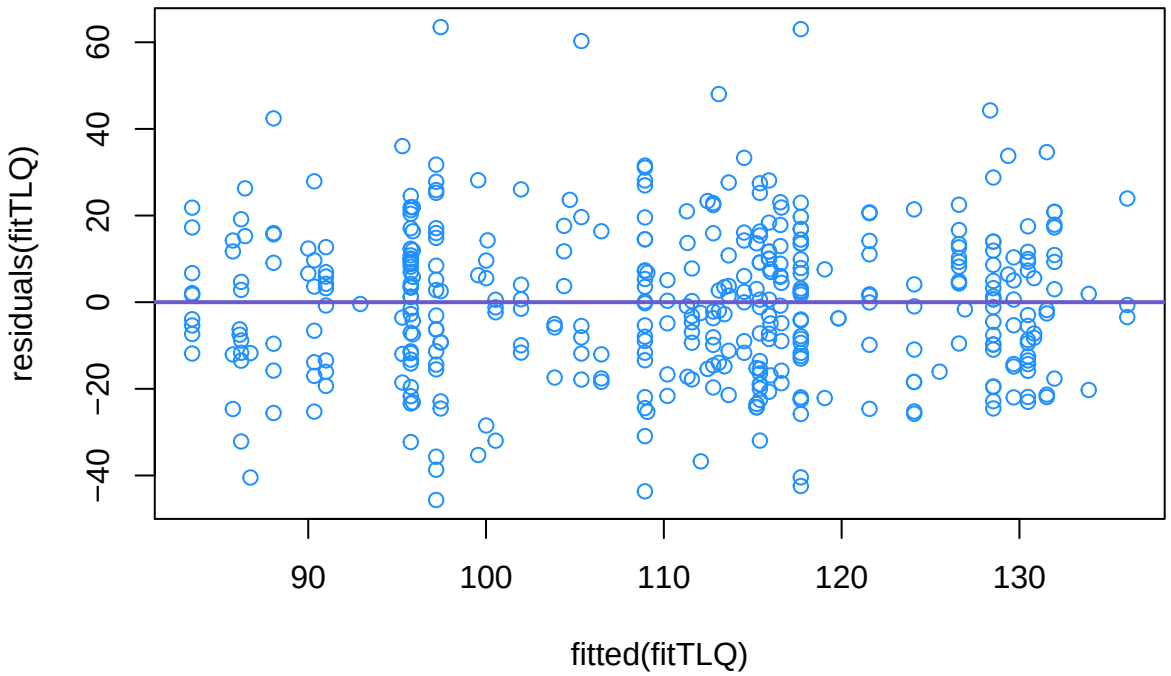
```
knots <- seq(min(x), max(x), length = 22)[-c(1,22)]
X <- cbind(1,x)
for(k in 1:20) X <- cbind(X, trLin(x, knots[k]))
fitTLQ <- lm(y ~ -1 + X)
Xg <- cbind(1, xg)
for (k in 1:20) Xg <- cbind(Xg, trLin(xg, knots[k]))
fHatTLQg <- as.vector(Xg%*%fitTLQ$coef)
```

```
plot(x, y, col = 'dodgerblue')
lines(xg, fHatTLQg, col = 'darkgreen', lwd = 2)
```



The model clearly overfits to the training data. In some time periods there are very few observations (for example in 1940-1949 there are 8 observations) and the model overfits to those observations instead of following the general shape.

```
plot(fitted(fitTLQ), residuals(fitTLQ), col = 'dodgerblue')
abline(0, 0, col = 'slateblue', lwd = 2)
```



The residual versus fits plot suggests that the model is adequate and the assumption of homoscedasticity is not broken. It is a natural result as the model overfits to the training data, therefore it will perform well on these particular observations. However it might not do very well when provided with some new information. Overall this model is not a good fit. We should probably opt for less knots or use more data to fit our model.