Semiparametic Regression - Assignment 2

Szymon Armata | 341593 | Data Science

23 October 2022

Contents

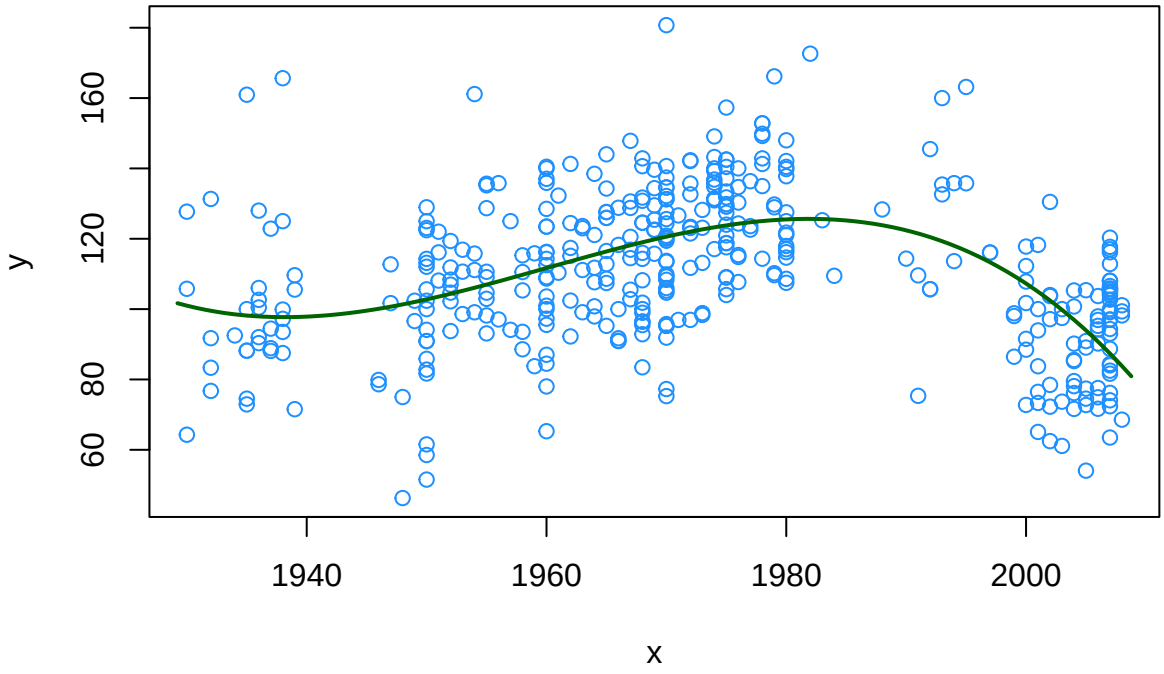
1 Task 1.

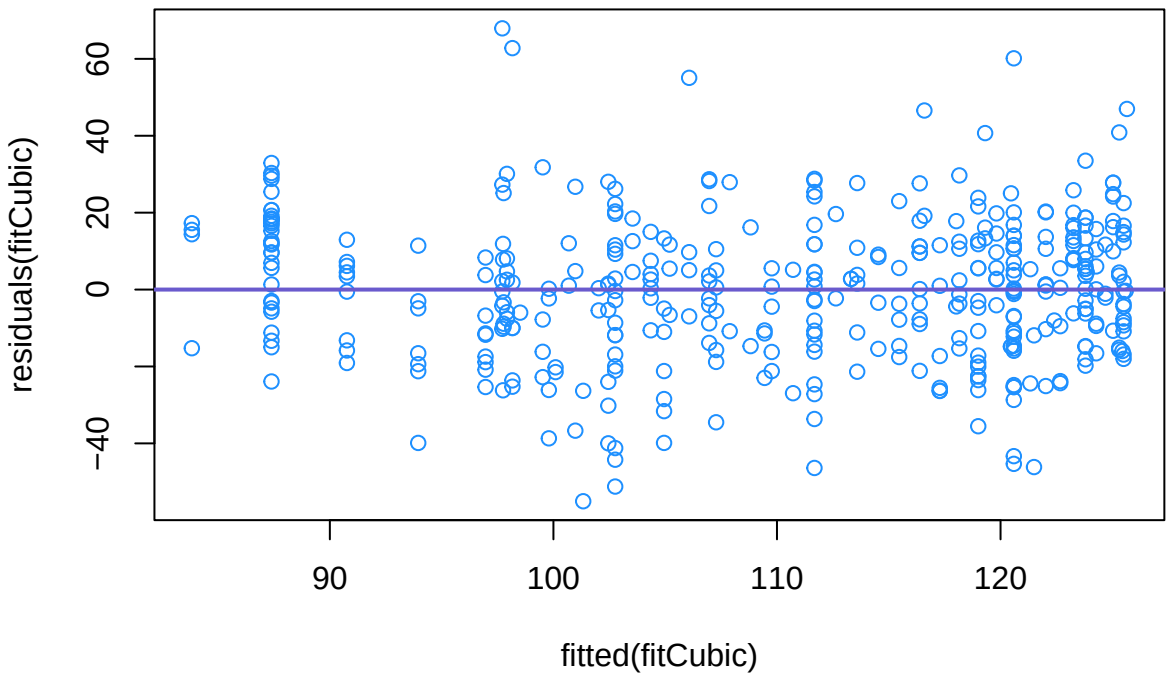
1.1 a

First, we build a cubic regression model using information on Warsaw apartment prices:

$$yi = \beta 0 + \beta 1 + \beta 2xi^2 + \beta 3xi^3 + \epsilon i$$

Next, we plot the fit and examine the residuals:





From the first plot above we can see that a smooth curve provides a sufficient level of data fit.

The second plot shows no dependence between the fitted values and the residuals, which supports homoscedasticity.

1.2 b

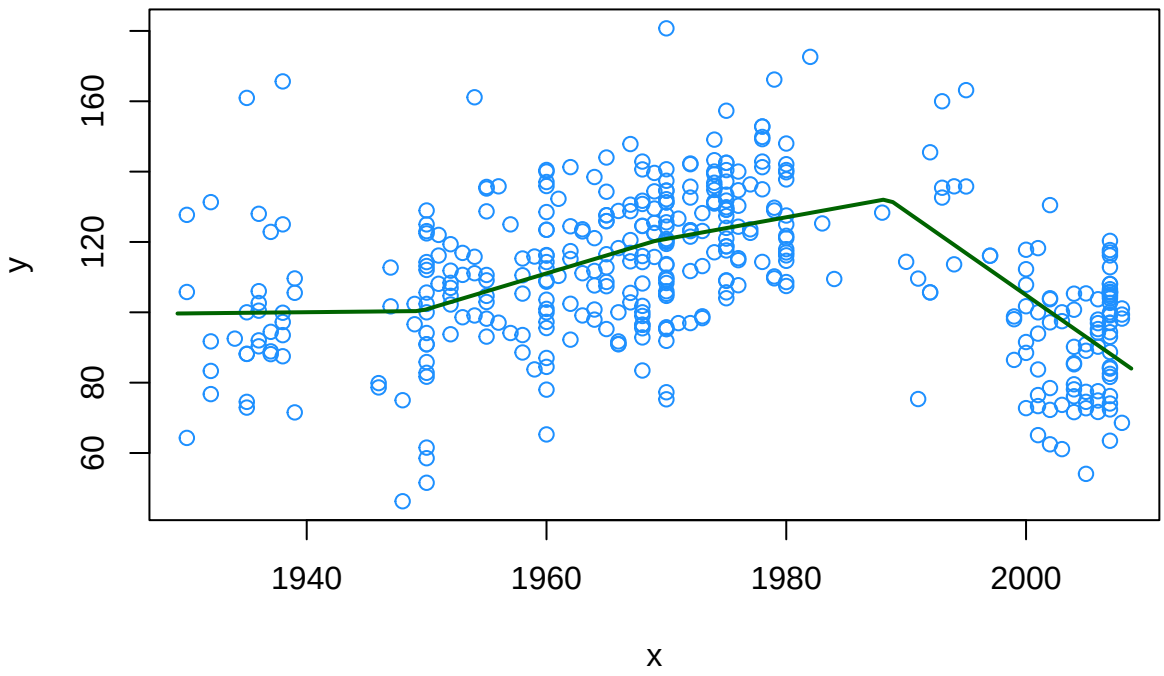
Now we define the truncated line function with a knot at kappa (κ):

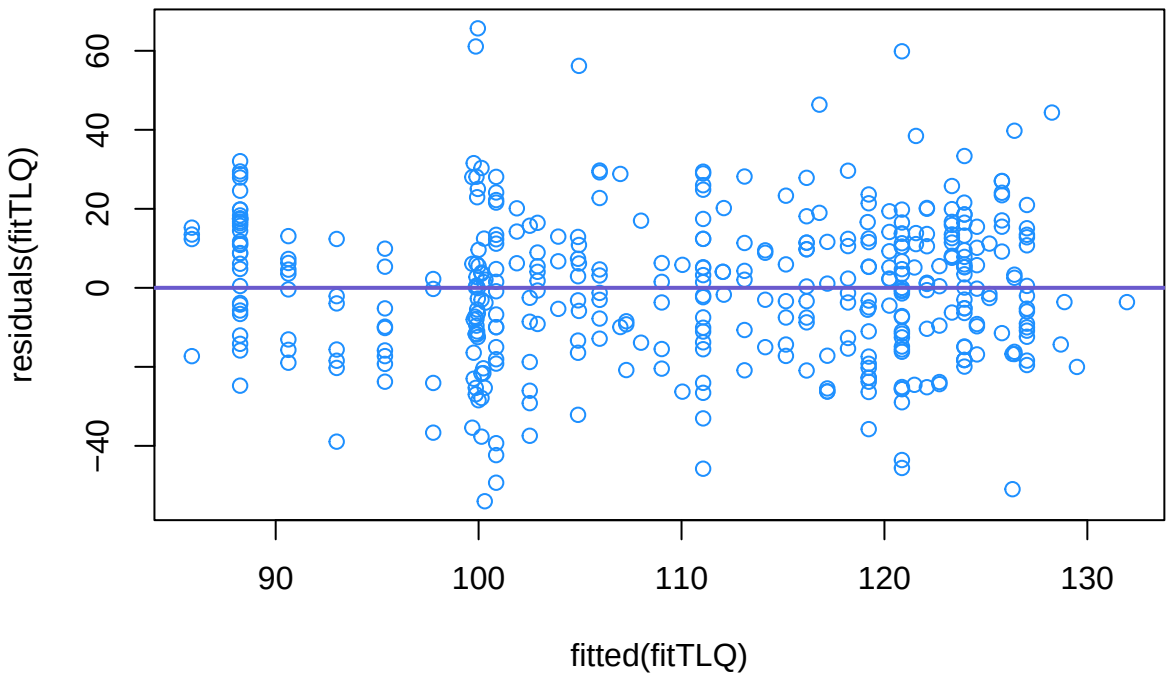
$$(x-\kappa)_+ = (x-\kappa) \cdot 1_{\{x>\kappa\}}$$

1.3 c

Next, we consider the spline regression model to data from **a**:

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

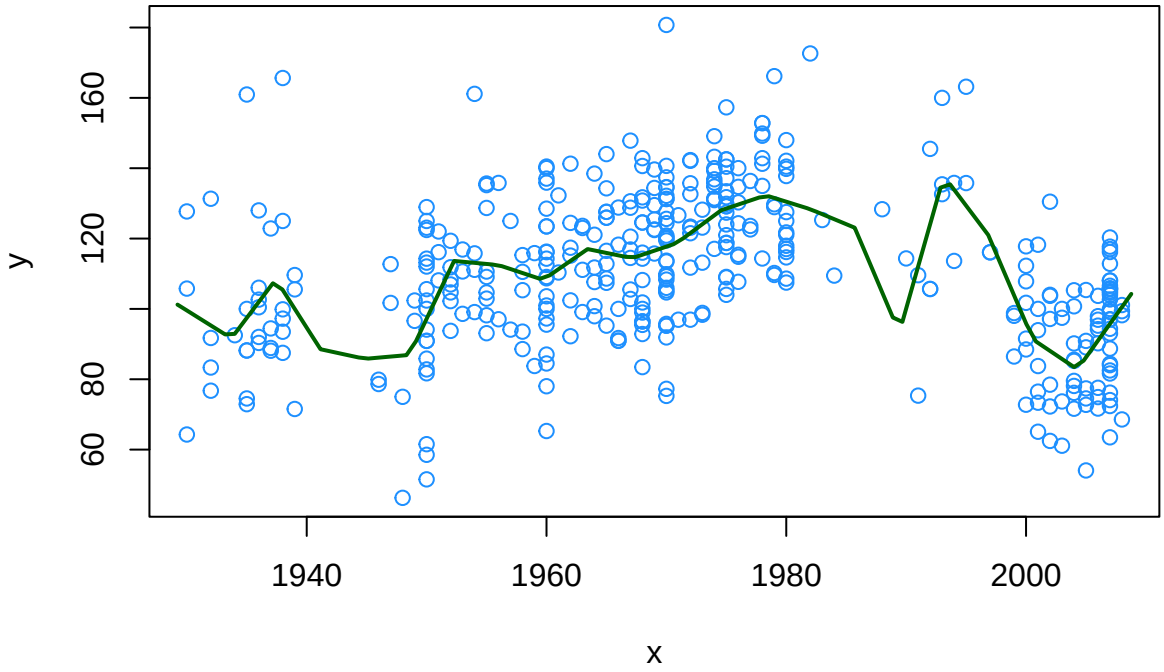

y
160
120
80
60
1940
1960
1980
2000
x

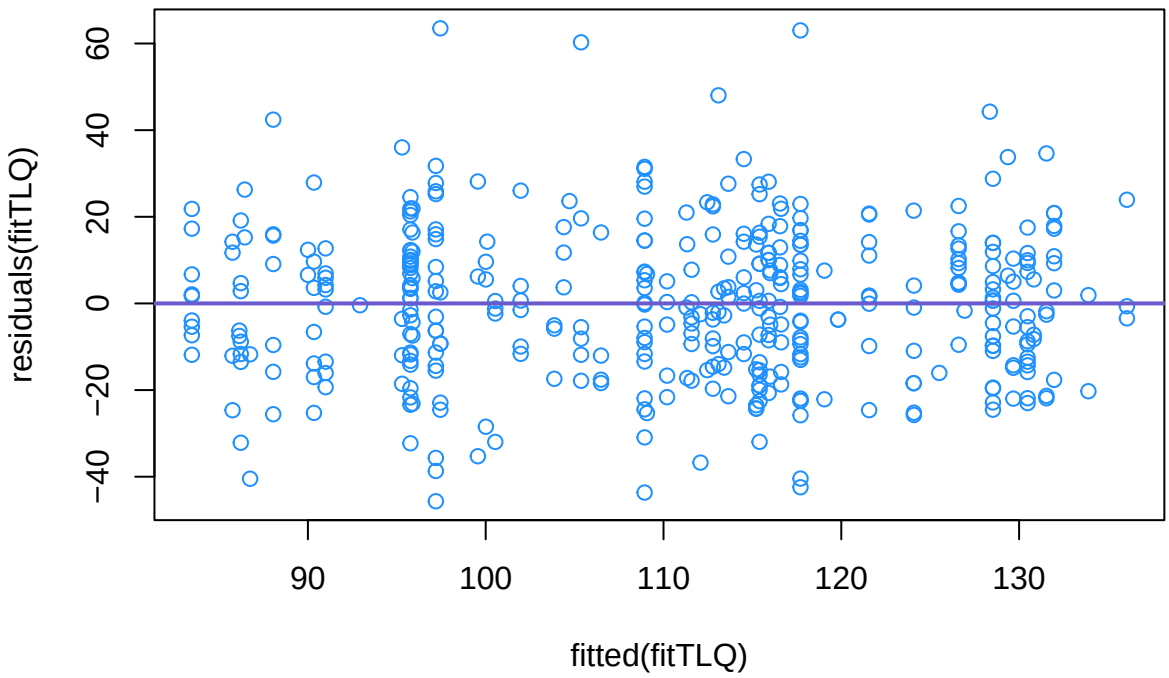


On the first plot we obtain curve which is quite good fitted to the data, but not that smooth like on the plot from **a**.

The second plot is similar to the one from point **a**. No dependence is seen between the residuals and the fitted values. This means good properties of the model.

1.4 d





On the first plot, we can see that there are too many knotes and the curve is overfitted. There are too many fluctuations in the fitted line which do not necessarily show dependencies in the actual data.

For example - in years 1940-1950 and 1980-1997, there is an extreme lack of data. In the second period we have about 14 observations to model, but the fitted line shows many dependencies (it has 4 extrema).

On the other hand, the second plot does not show any dependence between the residuals and the fitted values.

The model satisfied one of the assumptions, but the visual diagnosis of the fitted line suggests that it is not well constructed. We could use fewer knots to repair this model.

We could also do some better data mining to find some observations that we could subtract from our sample (i.e. which residuals don't appear to be normal by default).