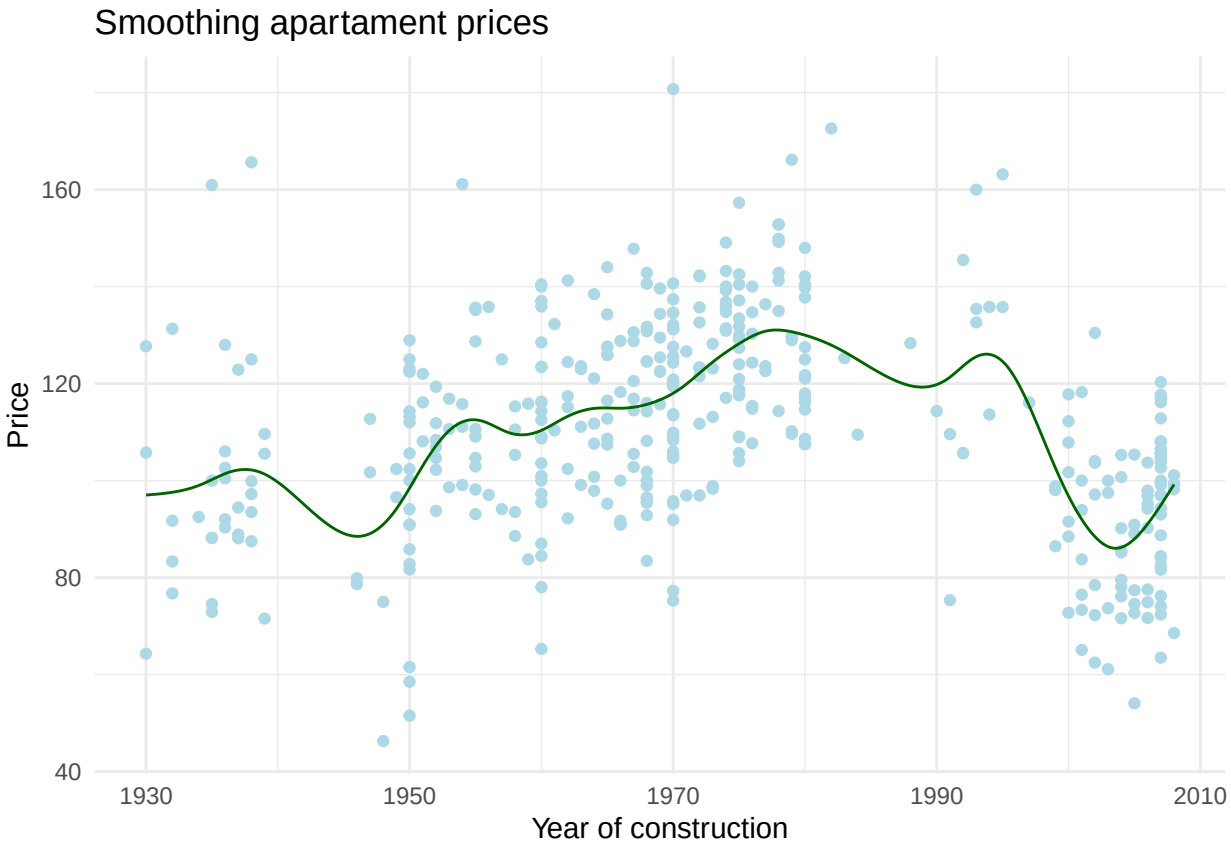
Semiparametic Regression - Assignment 4

Szymon Armata | 341593 | Data Science

08 November 2022

Contents

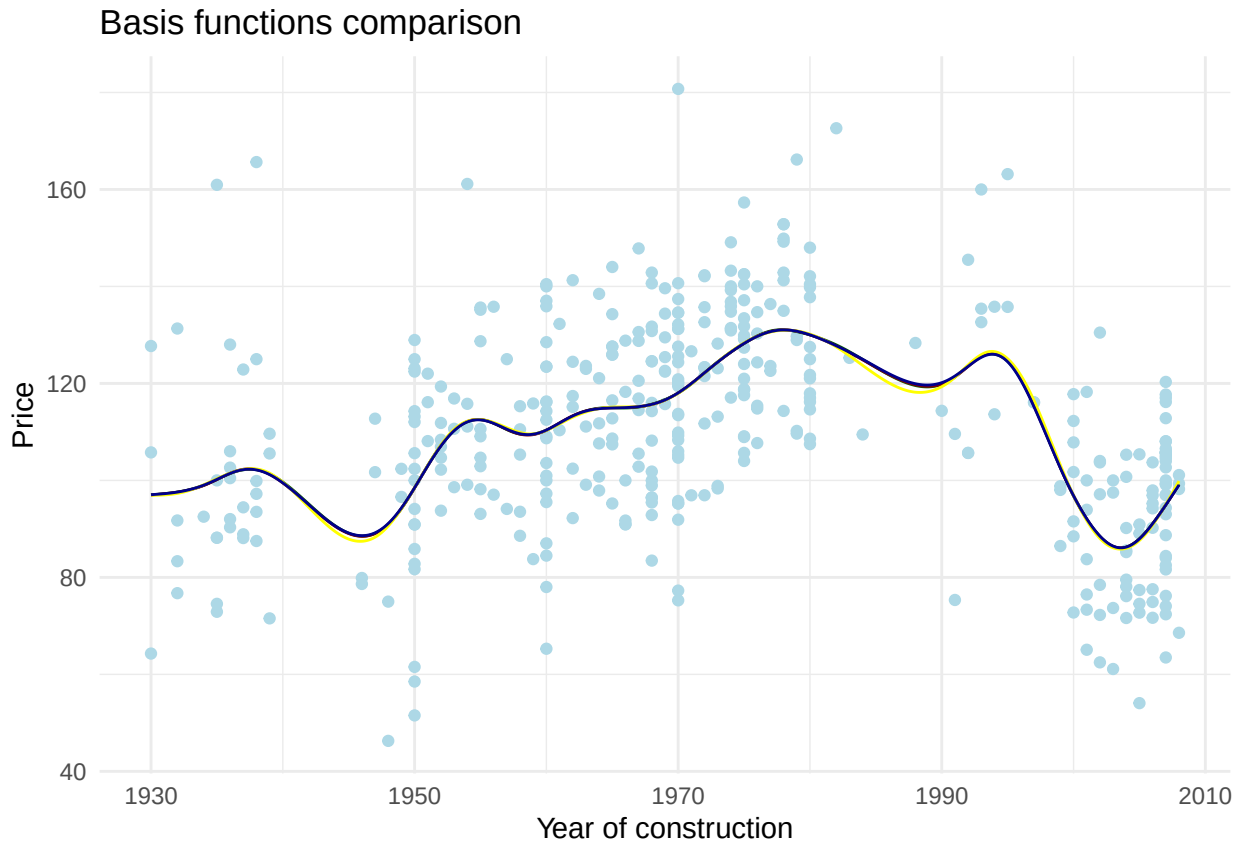
1 Task 1.



Even though there are times when there is a shortage of data (for example, the years 1980–1995) and the line appears to be overfitted, it can be seen that the fitting line represents the trend in apartment prices.

1.1 a

In this task we compare different types of penalized splines: Gaussians process basis functions, P-splines and thin plate regression and show their results on one plot.



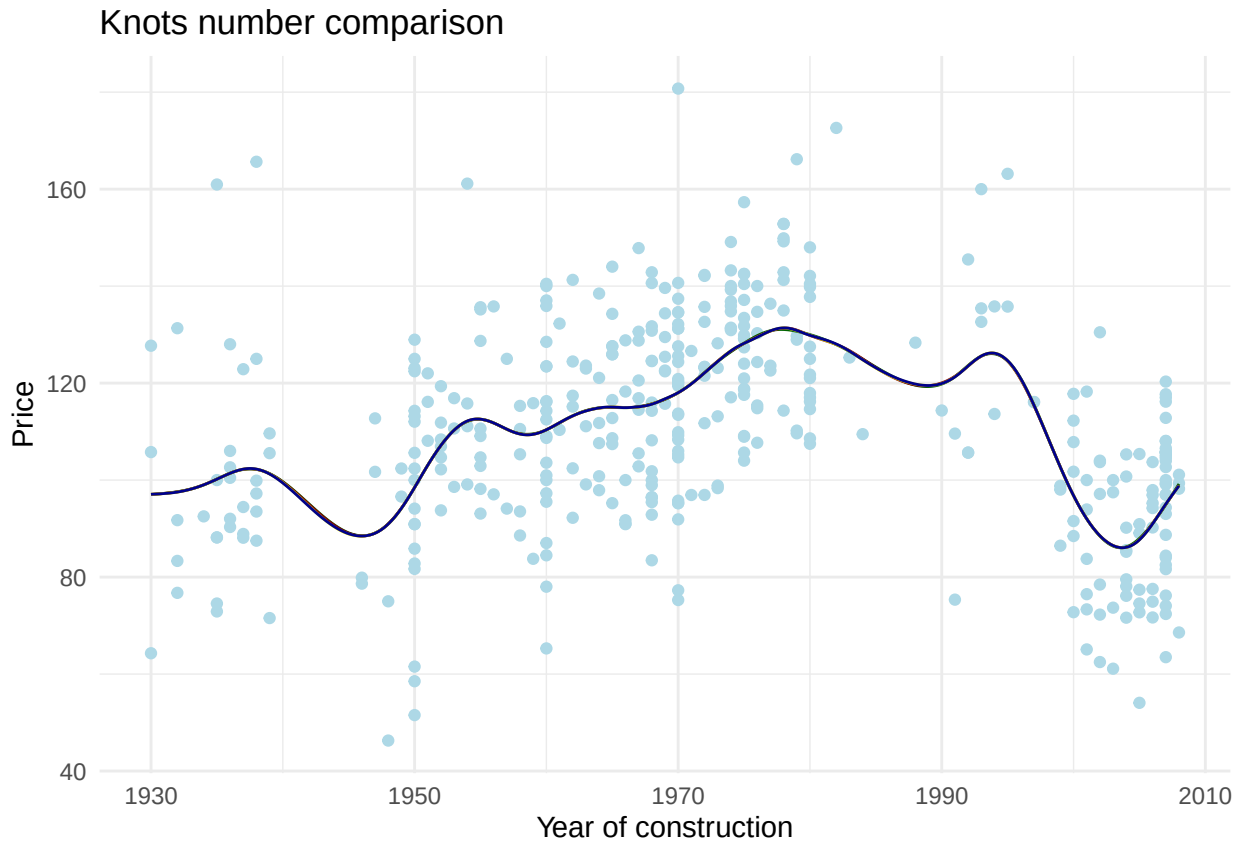
The differences between the various basis functions in this specific example are minimal, as we can see (or rather, as we cannot see).

Only during times when there are few observations are there any discernible differences between them.

We can get the conclusion that the basis functions we choose do not significantly affect line quality (on this particular data).

1.2 b

In this task we compare different numbers of knots (i.e. basis functions) used to fit (30,40,50,60) and show their results on one plot.

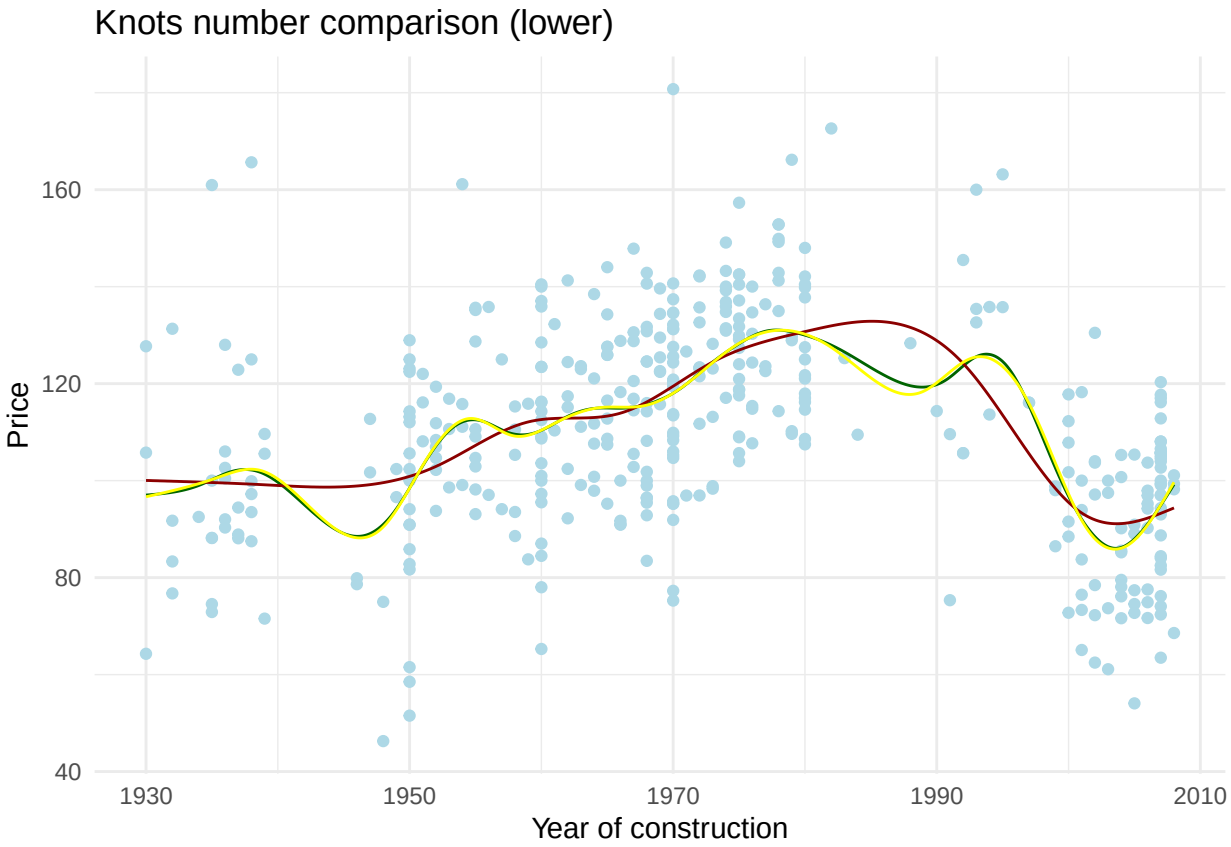


There is no significant difference between plots of all four fits.

Models with more than 30 nodes are not very different from the baseline one. It can therefore be assumed that the 30-node model represents all relevant trends in the data as well as the extremes.

The relation between number of knots and fit is logarithmic, that means we have to choose hundreds of knots to see the difference, but on provided data that number of knots does not make any sense.

Now, we do the same for lower numbers (10,20,30).



In the case of this plot, the situation is a little different than before.

Even a 20-knot model appears to be pretty close to versions with 30 knots or more. It depicts every trend in the data and every extreme that can be seen in higher-order models as well as the 20-knot model.

10-knot model is much different, because it fits the data less closely, but in my opinion, it is not overfitted. It captures the general trend in the data, but is less sensitive for the outliners.

1.3 c

Based on the graphs in subsections a and b, two main conclusions can be reached in the context of the relative influence of the type of basis and number of basis functions.

Type of basis did not have much impact on fit. Both Gaussian process basis functions, P-splines and thin plate regression splines gave an almost identical plot shape, making it difficult to speak of much significance. I would assess their influence as very weak.

The situation is slightly different in terms of the impact of number of basis function on fit. Up to a certain level, differences in the shapes of the plots could be seen quite clearly, while beyond this level the number of knots did not significantly affect the plot. I would assess their influence as strong below a certain level and weak above it.