

Nonparametric Regression

Crowding on Mt. Everest



Source: Nirma Purjal Project Possible Ltd.

Mt. Everest in the news

The BBC News website features a prominent red header bar. On the left is the BBC logo and a 'Sign in' button. To the right are navigation links for 'Home', 'News', 'Sport', 'Reel', and 'Worklife'. Below the header is a large, bold title 'NEWS'. Underneath the 'NEWS' title is a horizontal menu bar with links: 'Home | Coronavirus | Climate | Video | World | US & Canada | UK | Business | Tech | Science | Stories'. A secondary navigation bar below it includes 'Asia | China | India'. The main content area features a large, bold headline: 'Mount Everest: Why the summit can get so crowded'. Below the headline is the author's name, 'By Helier Cheung BBC News'.

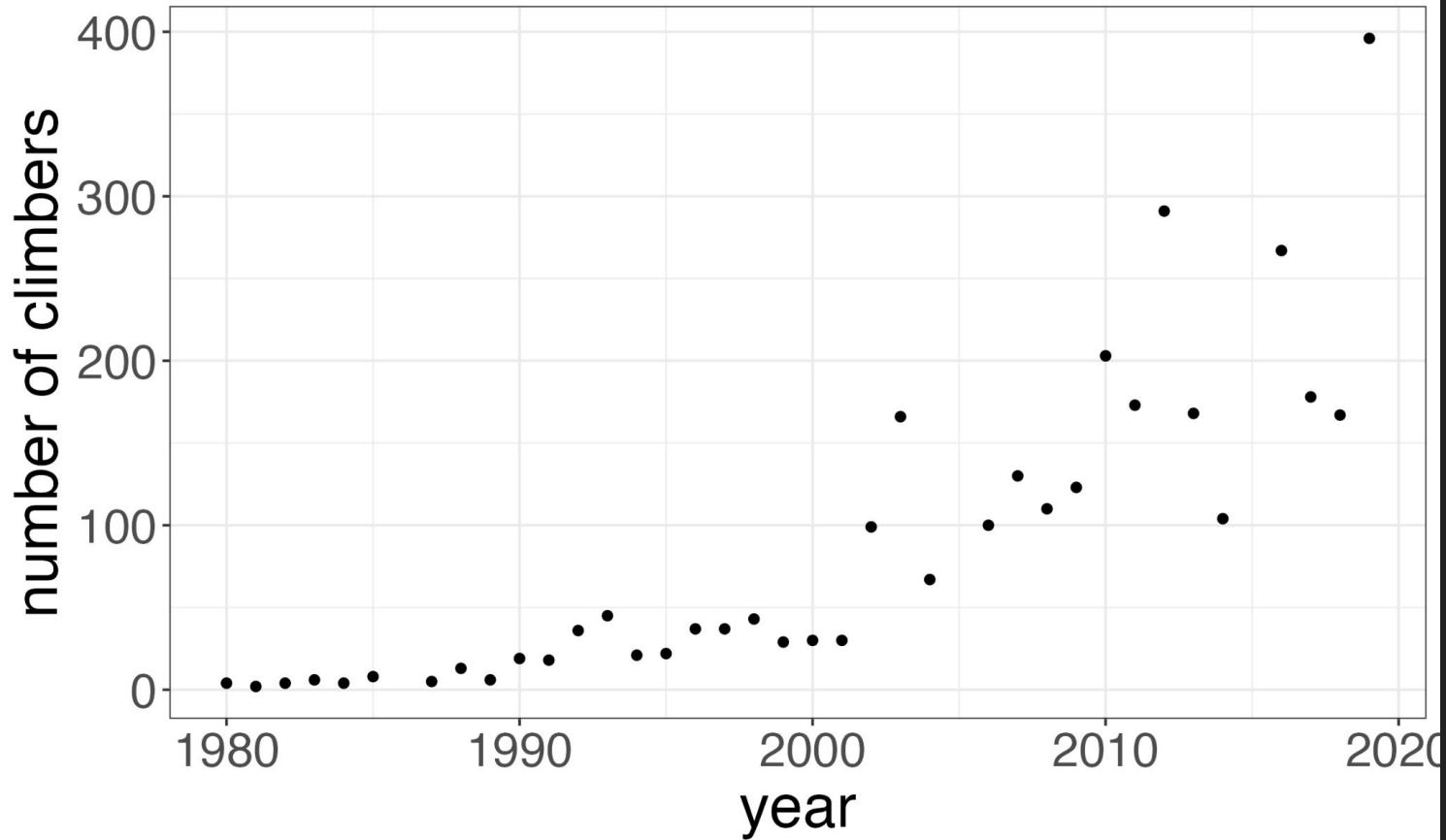
A GQ article titled 'Chaos at the Top of the World' is displayed. The article is categorized under 'Culture'. The main headline is 'Chaos at the Top of the World'. The source is cited as 'Source: GQ'.

The Washington Post website features a dark header bar with the publication's name and tagline 'Democracy Dies in Darkness'. A blue button on the right says '99¢ every fo...'. Below the header is a navigation bar with categories: 'World', 'Africa', 'Americas', 'Asia', 'Europe', 'Middle East', and 'Foreign Correspondents'. A link to 'Asia' is highlighted. The main content area features a large, bold headline: 'An often-overcrowded Everest has reopened to climbers. Some are questioning the decision.'

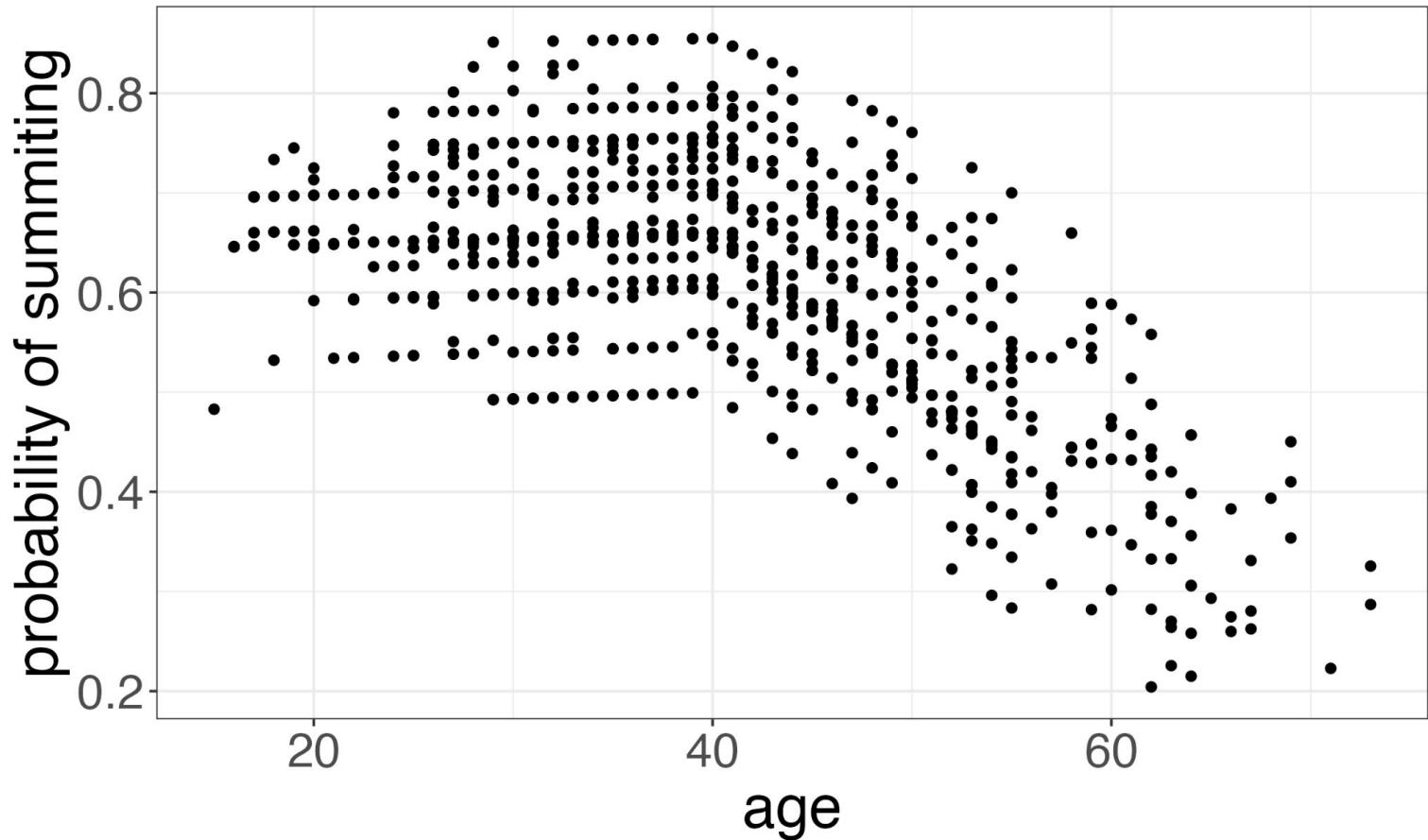
Policy Q: Should the Nepali government limit Everest permits?

- Safety concerns for climbers vs. economic benefits for Nepal
- How bad has the crowding gotten?

Crowding on Everest



Chance of summitting vs. age



The regression problem

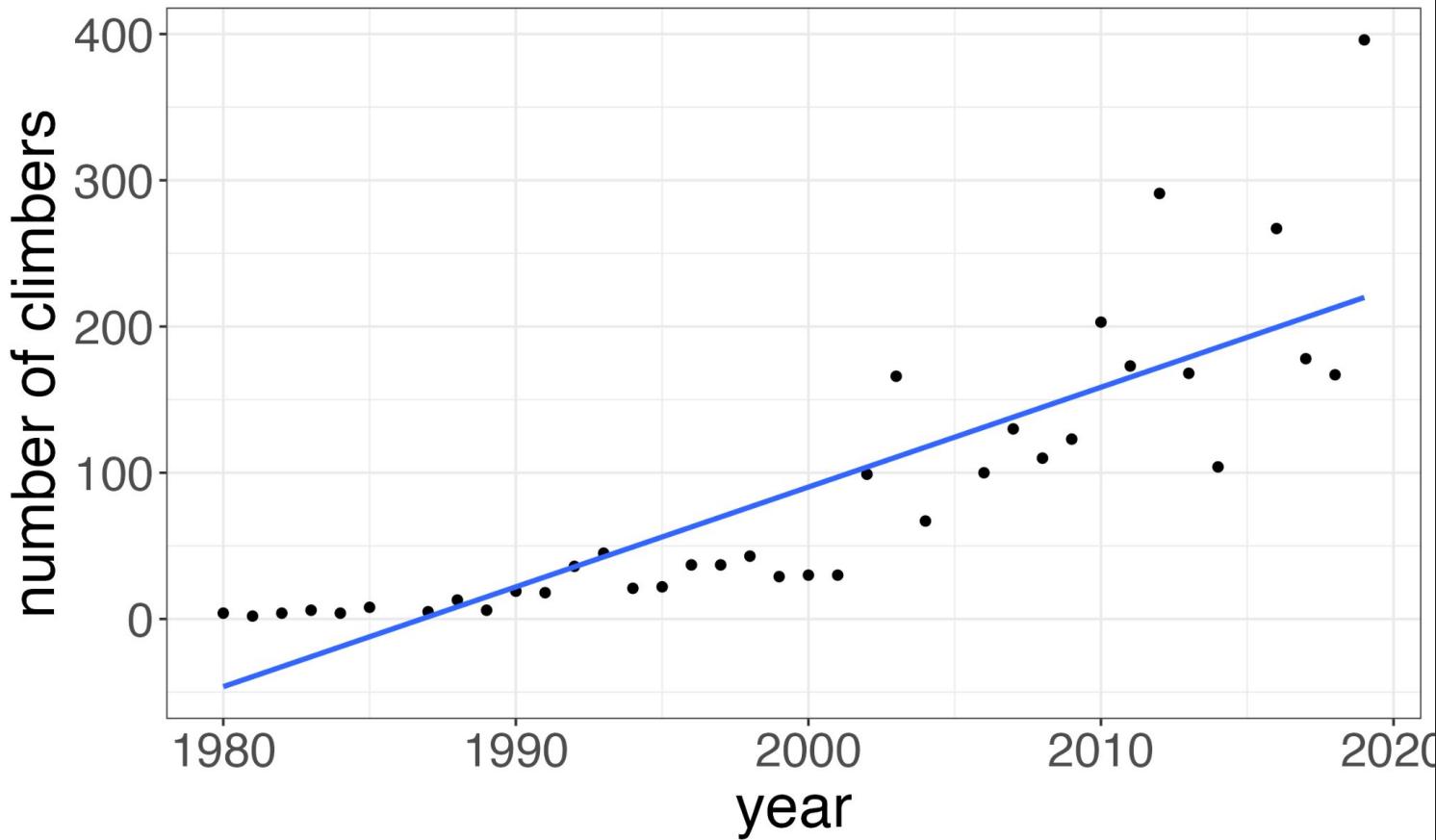
$$y = g(x) + \varepsilon$$

$$\mathbb{E}(\varepsilon) = 0$$

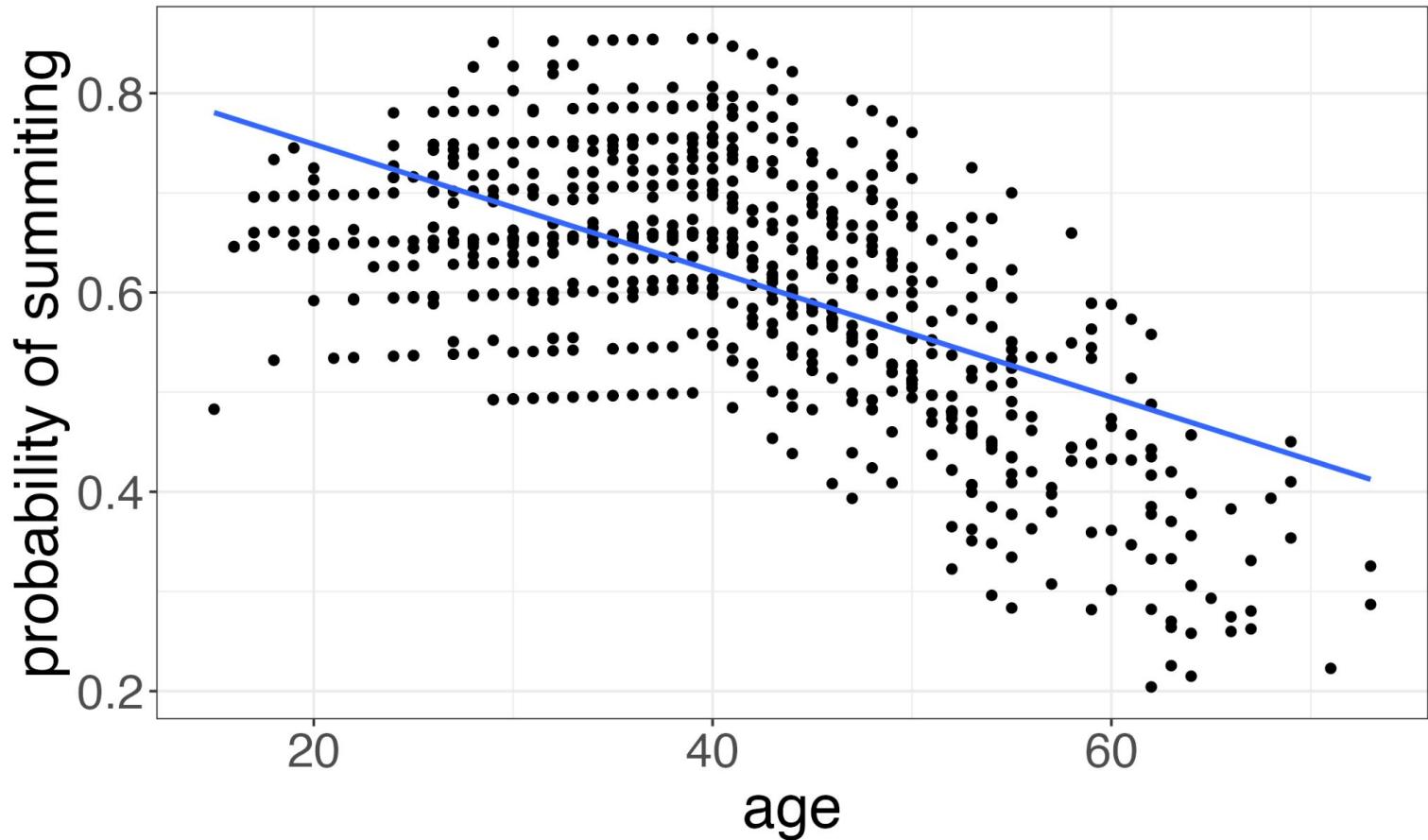
$$y = g(x) + \varepsilon$$

$$g(x) = \beta_0 + \beta_1 x$$

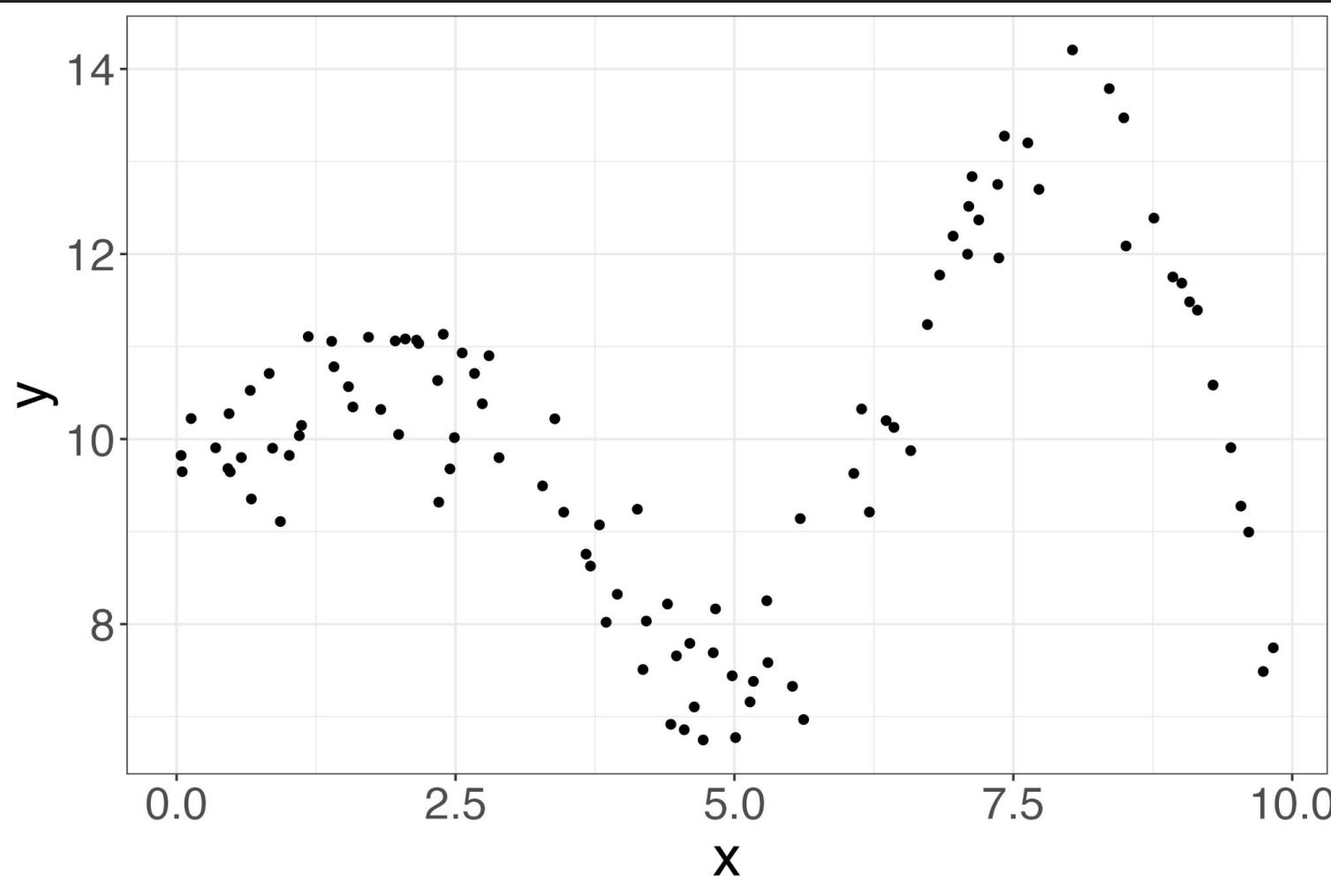
Crowding on Everest



Chance of summitting vs. age



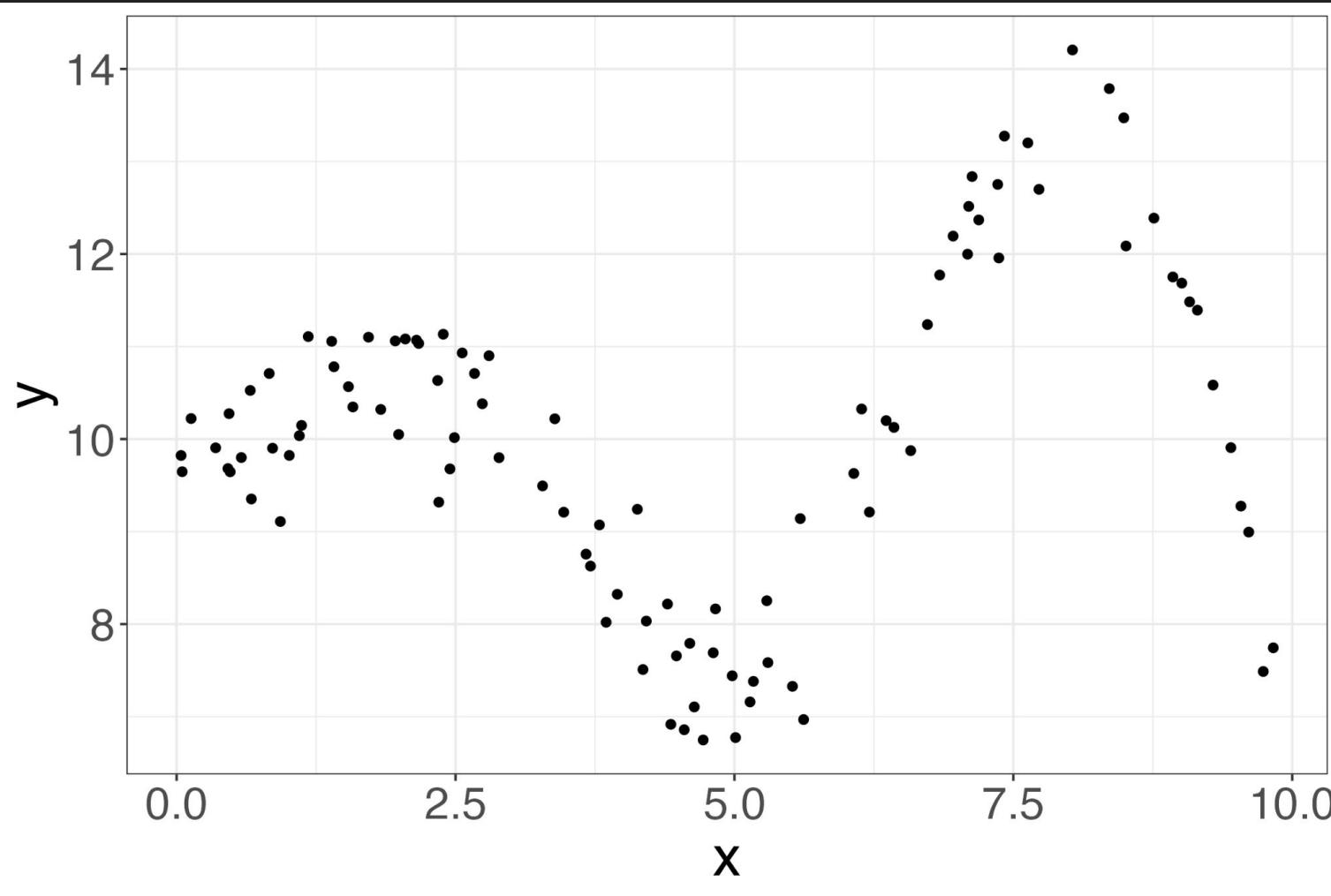
Need a flexible alternative to OLS



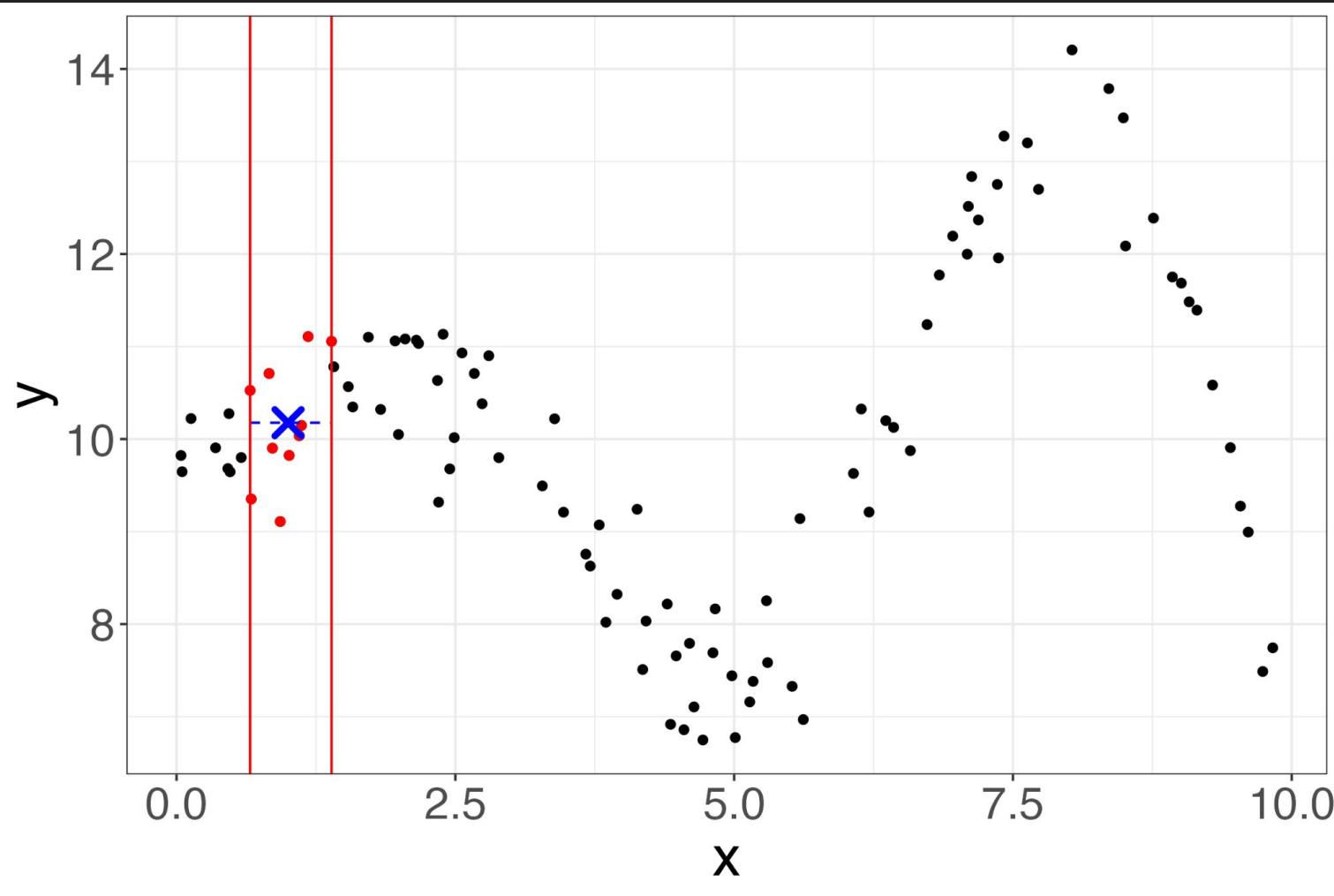
Could do polynomial regression...

$$\text{Ex: } g(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_{p-1} x^{p-1}$$

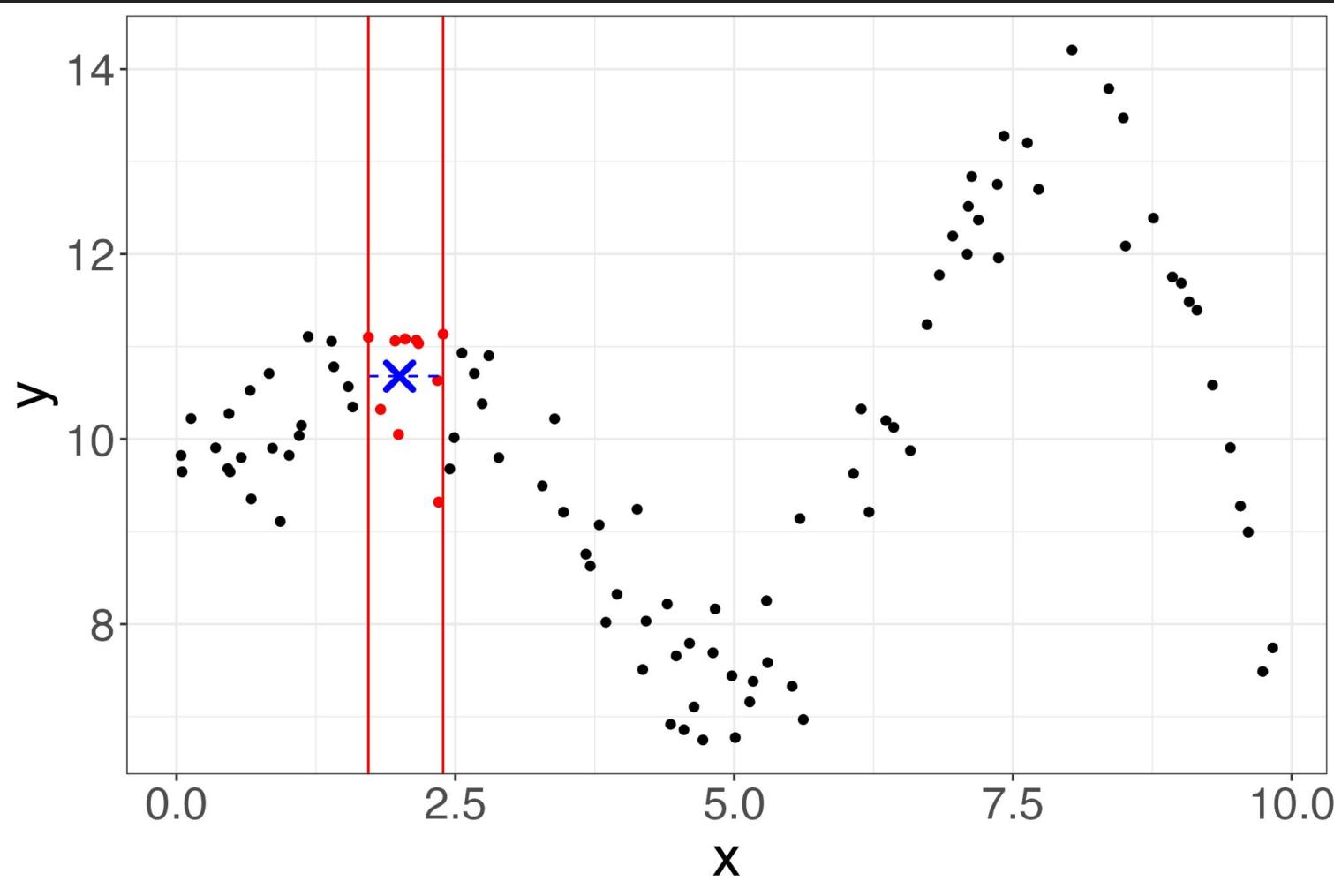
But let's try something simpler



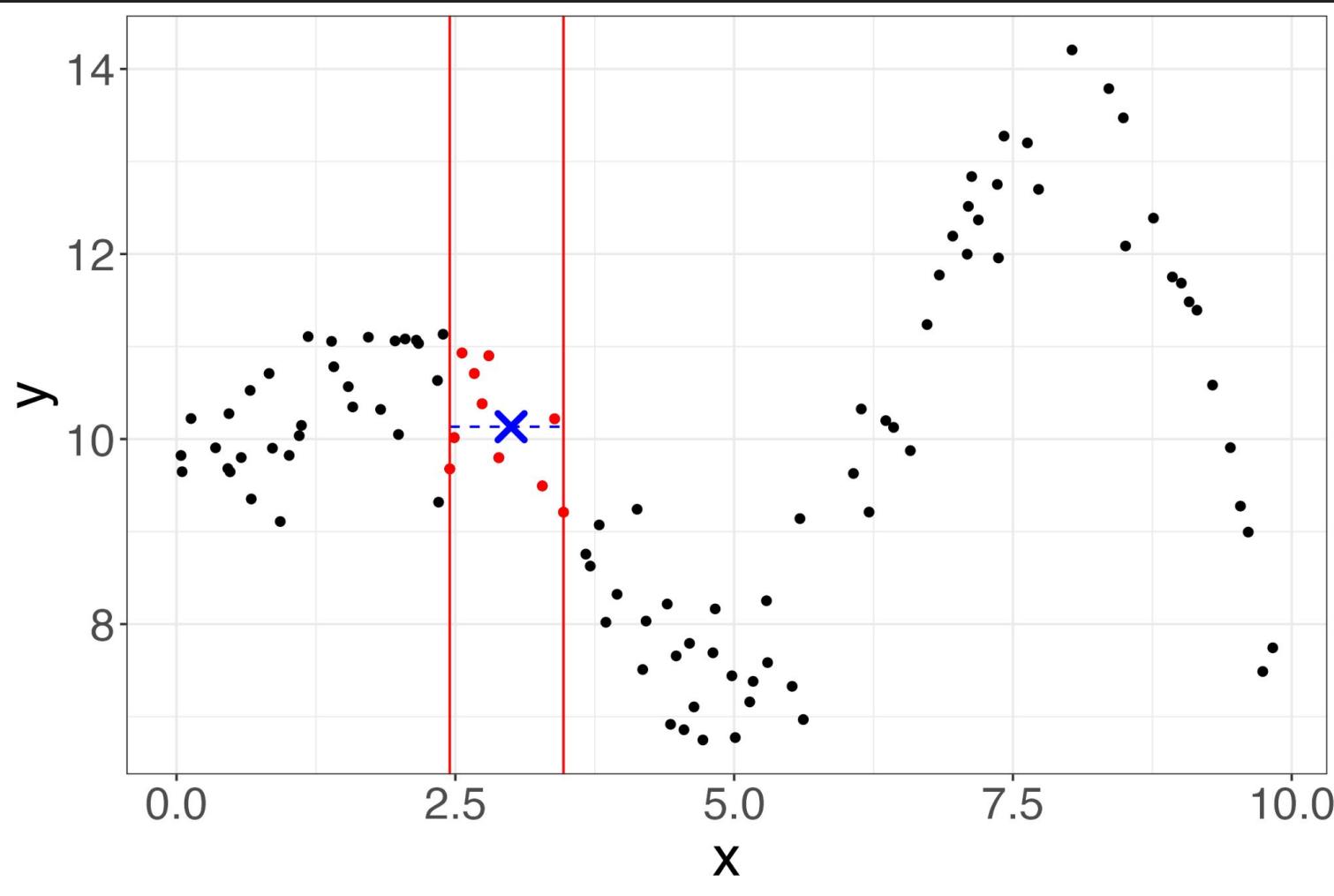
Prediction when $x=1$

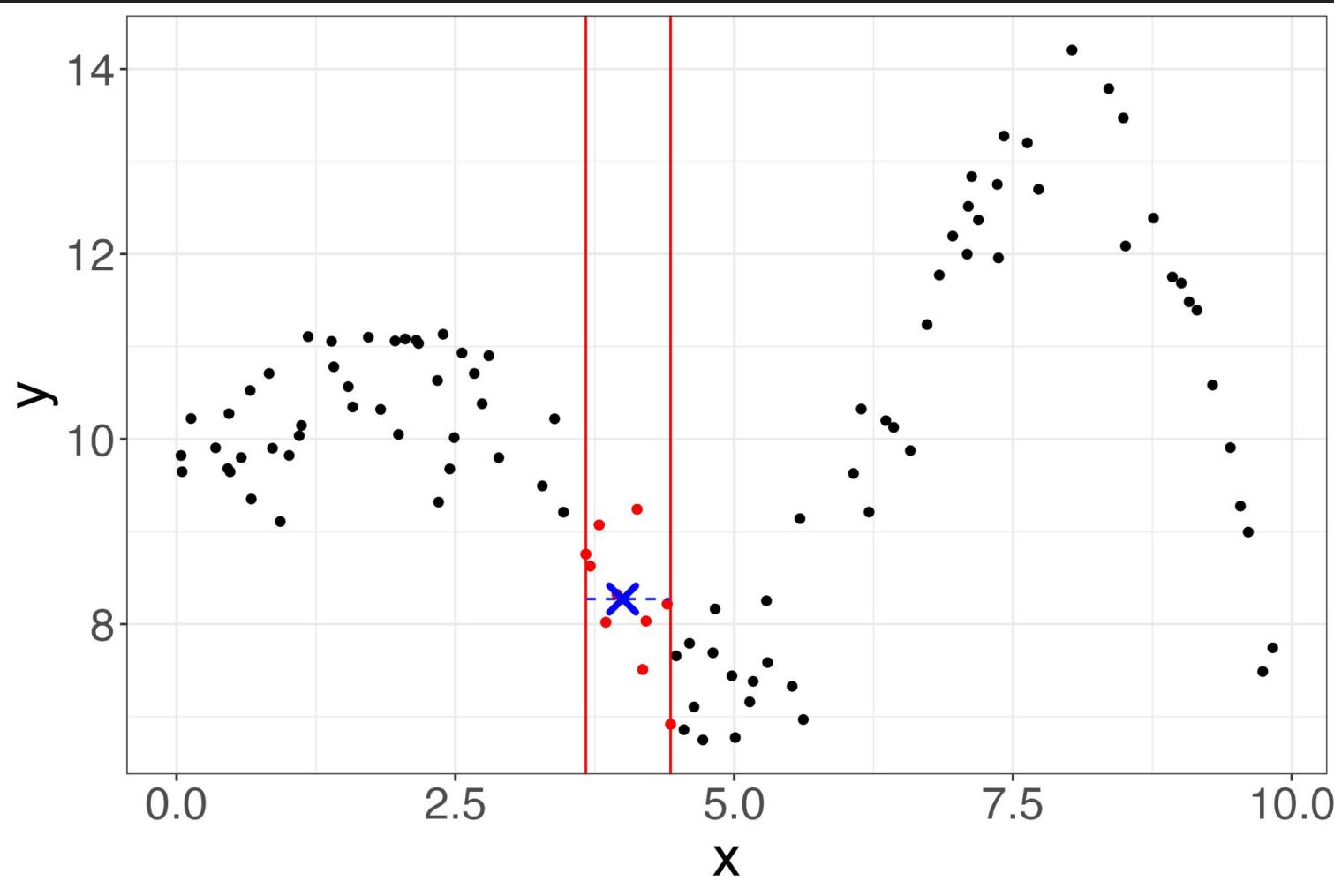


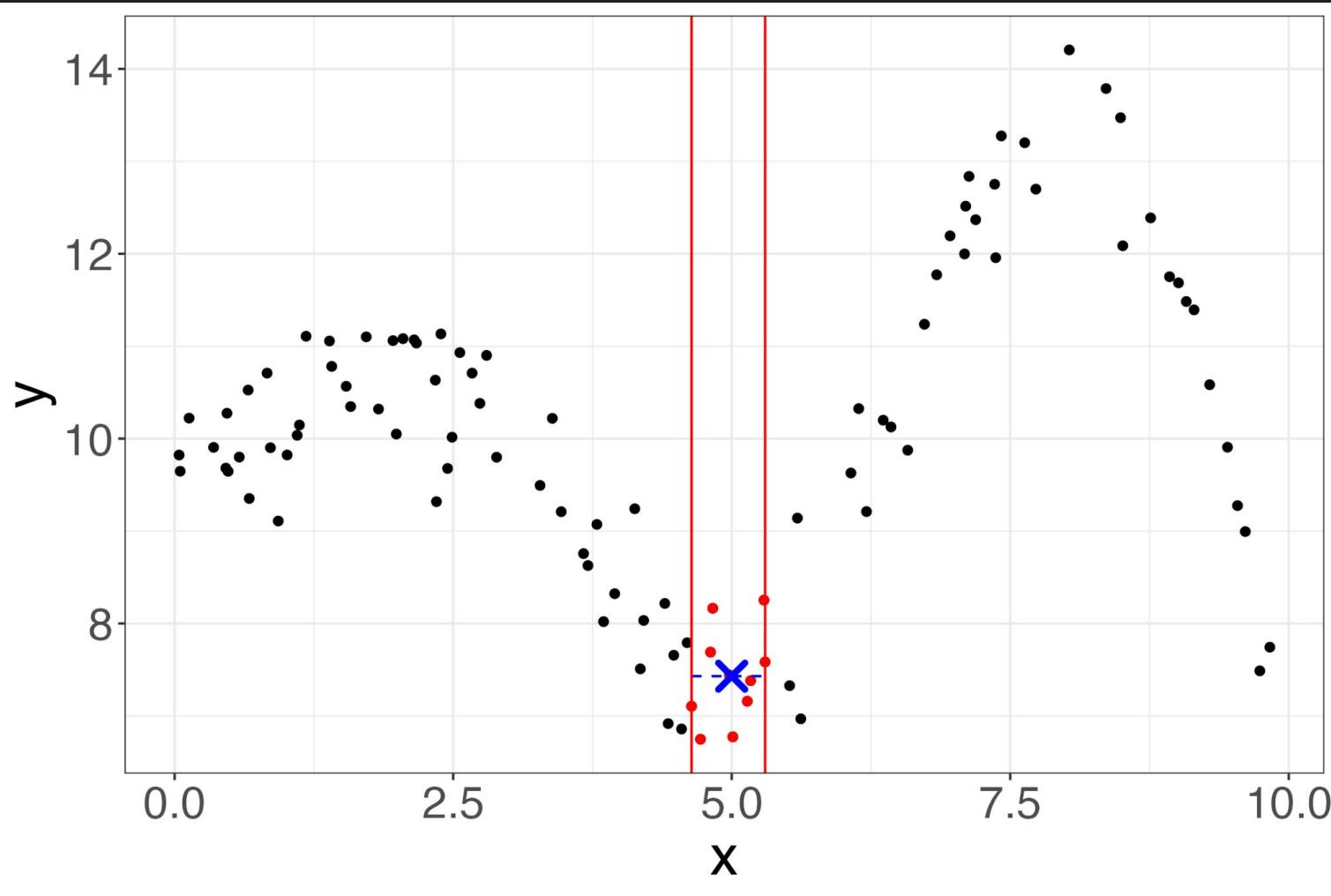
Prediction when $x=2$

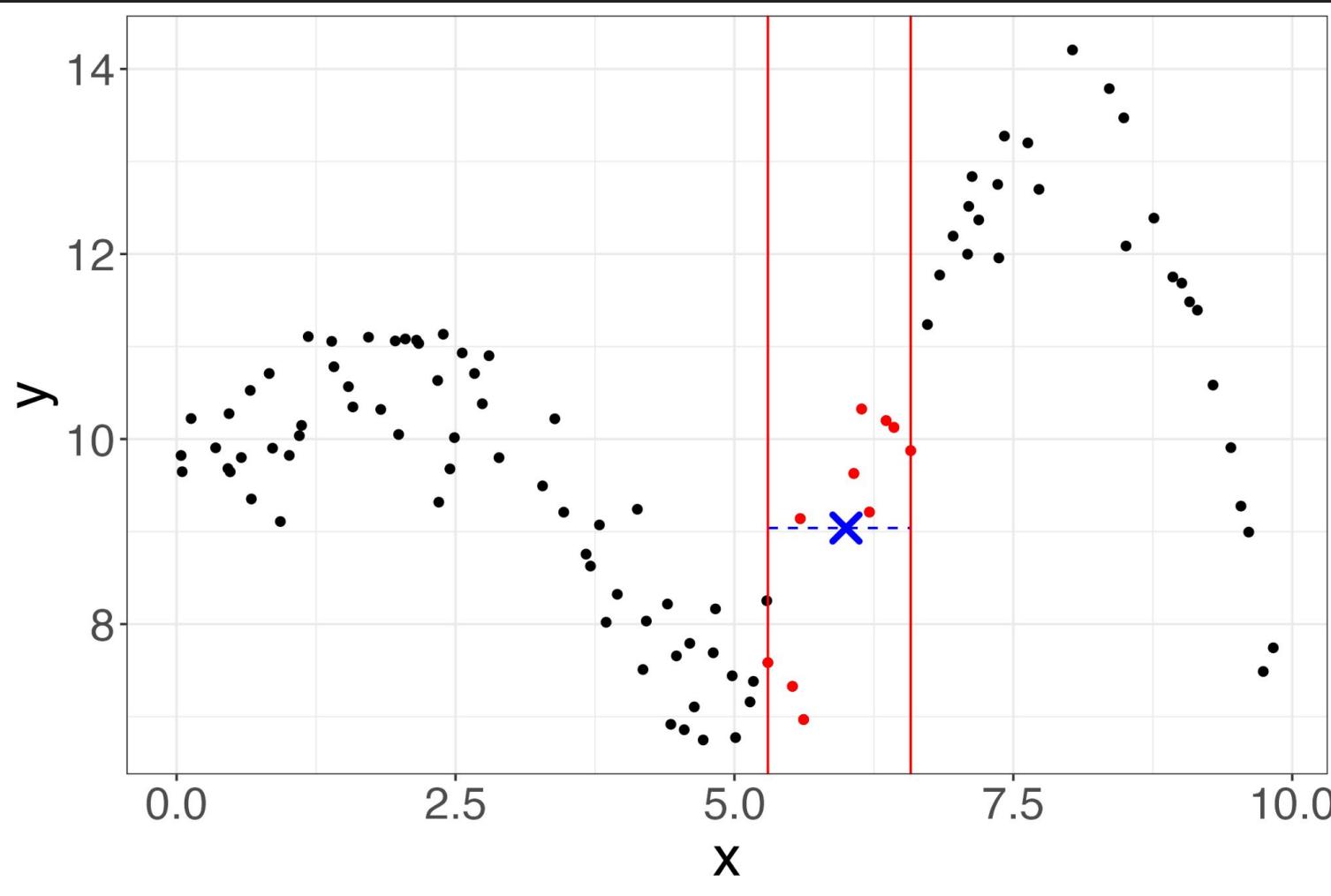


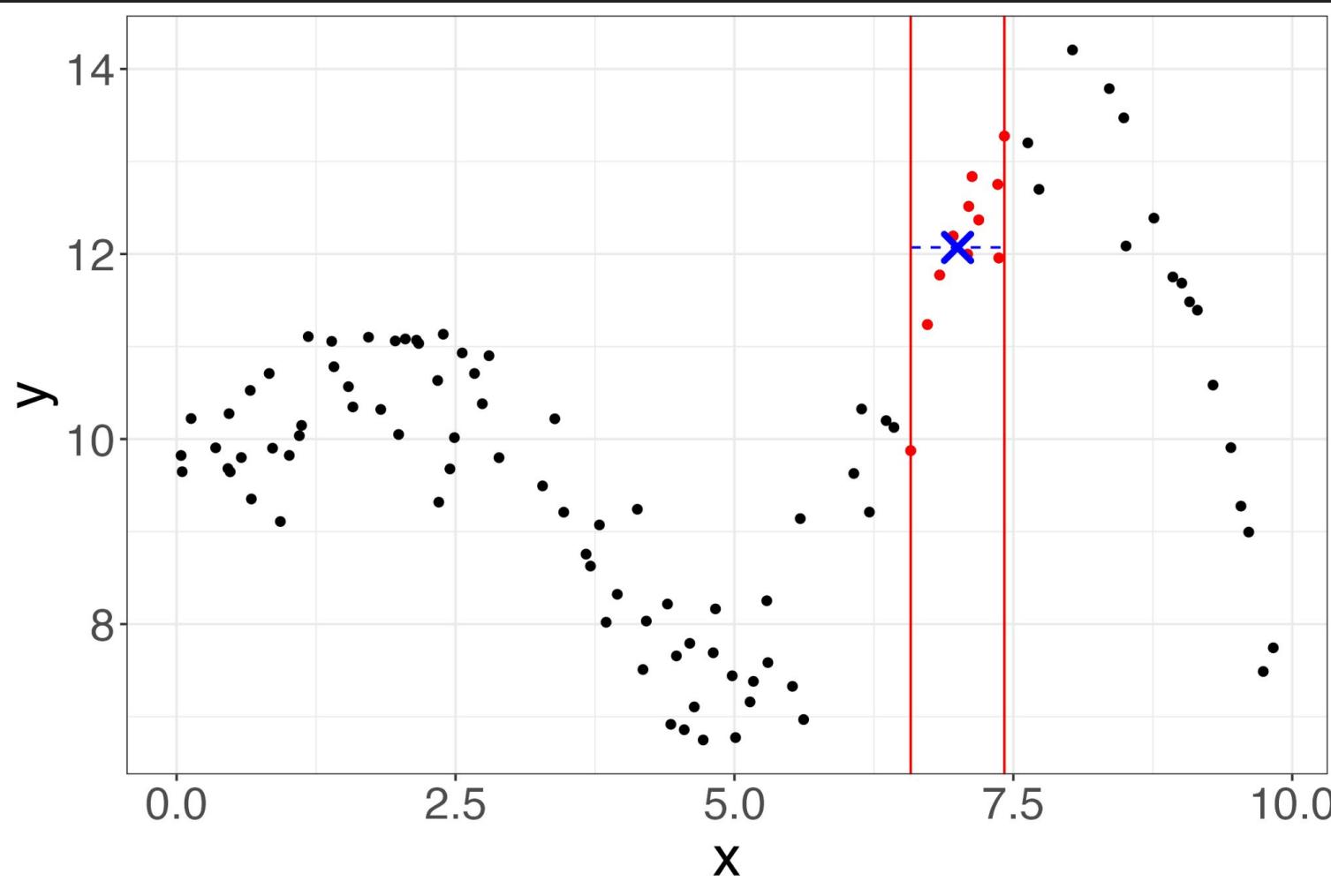
Keep going...

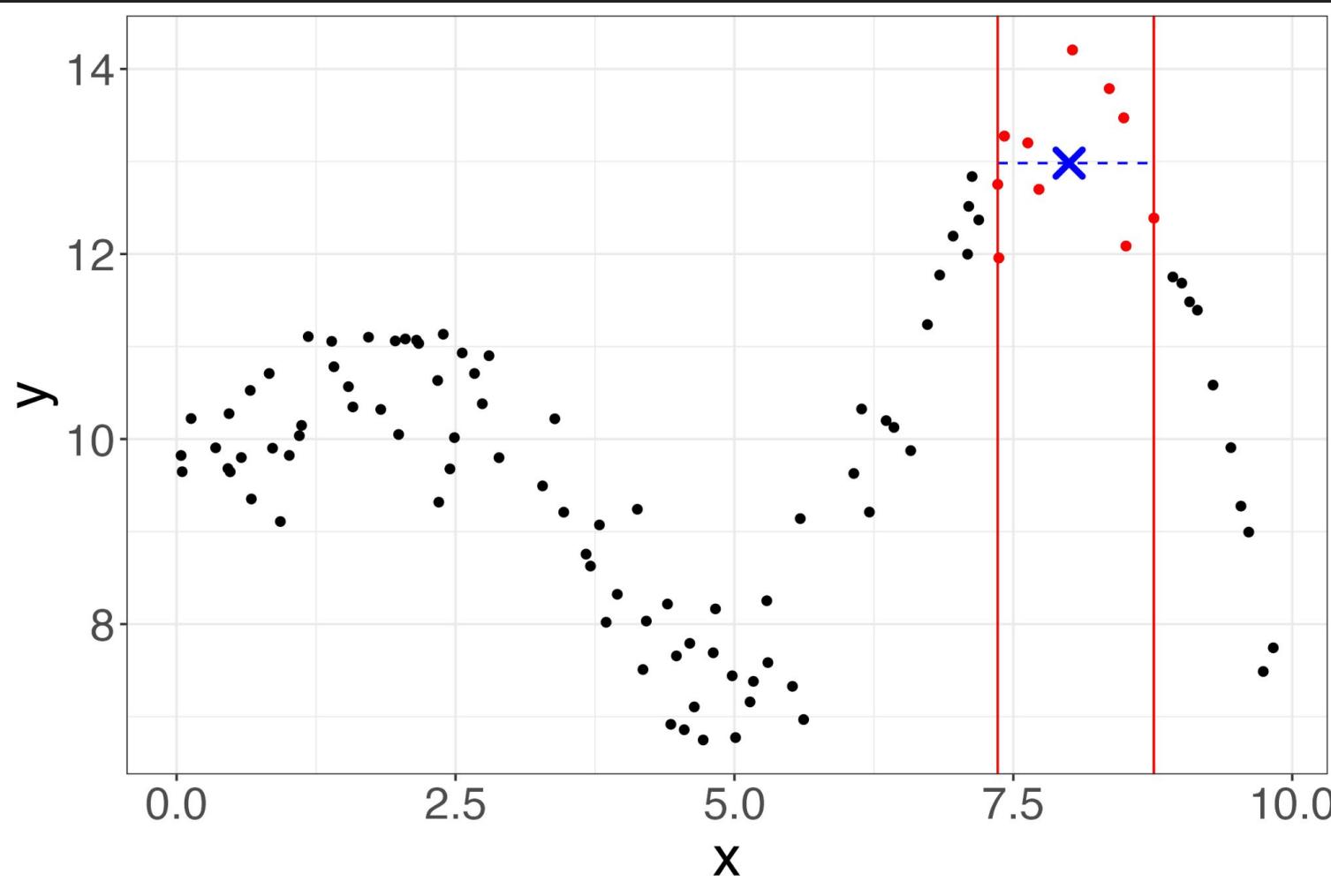


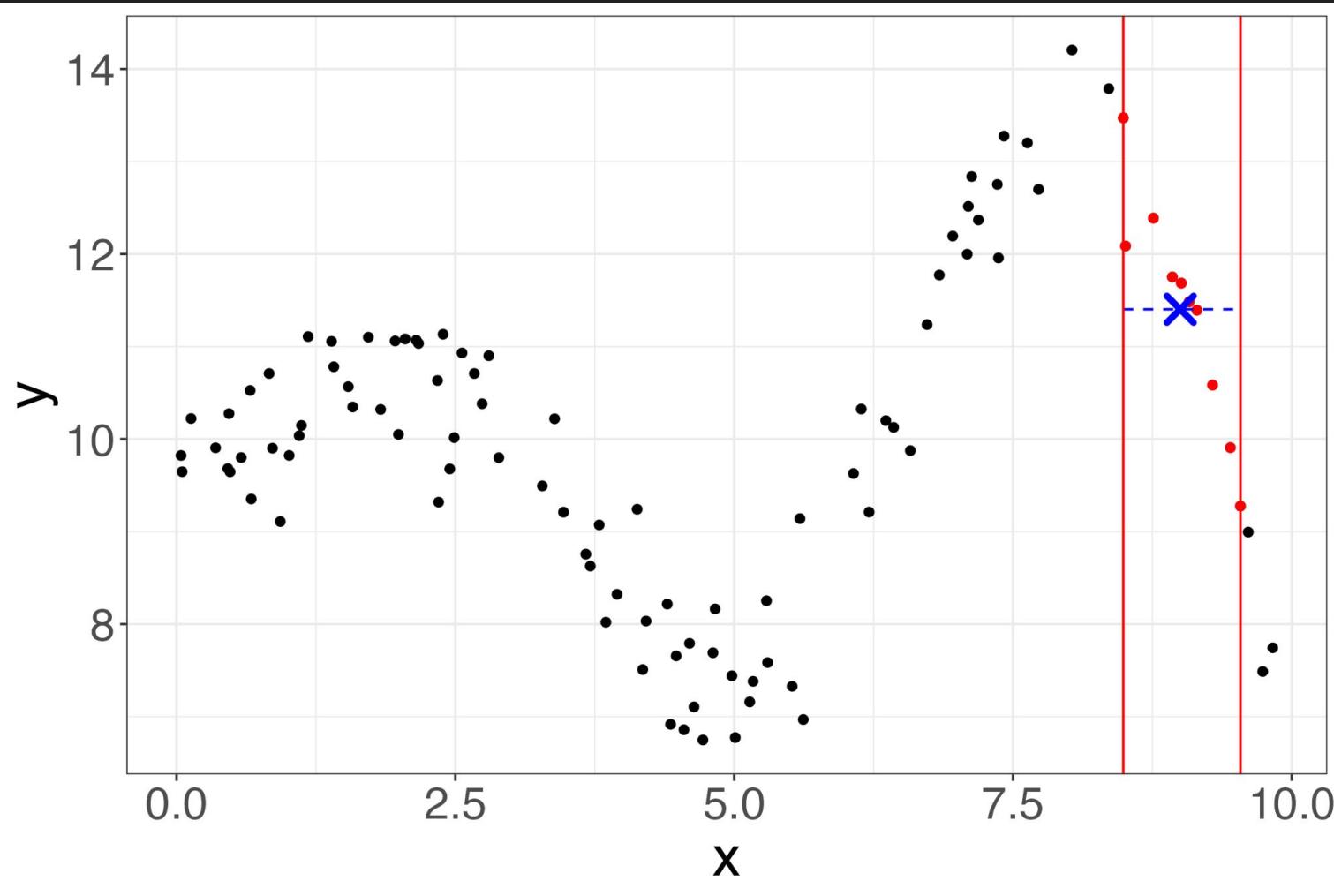




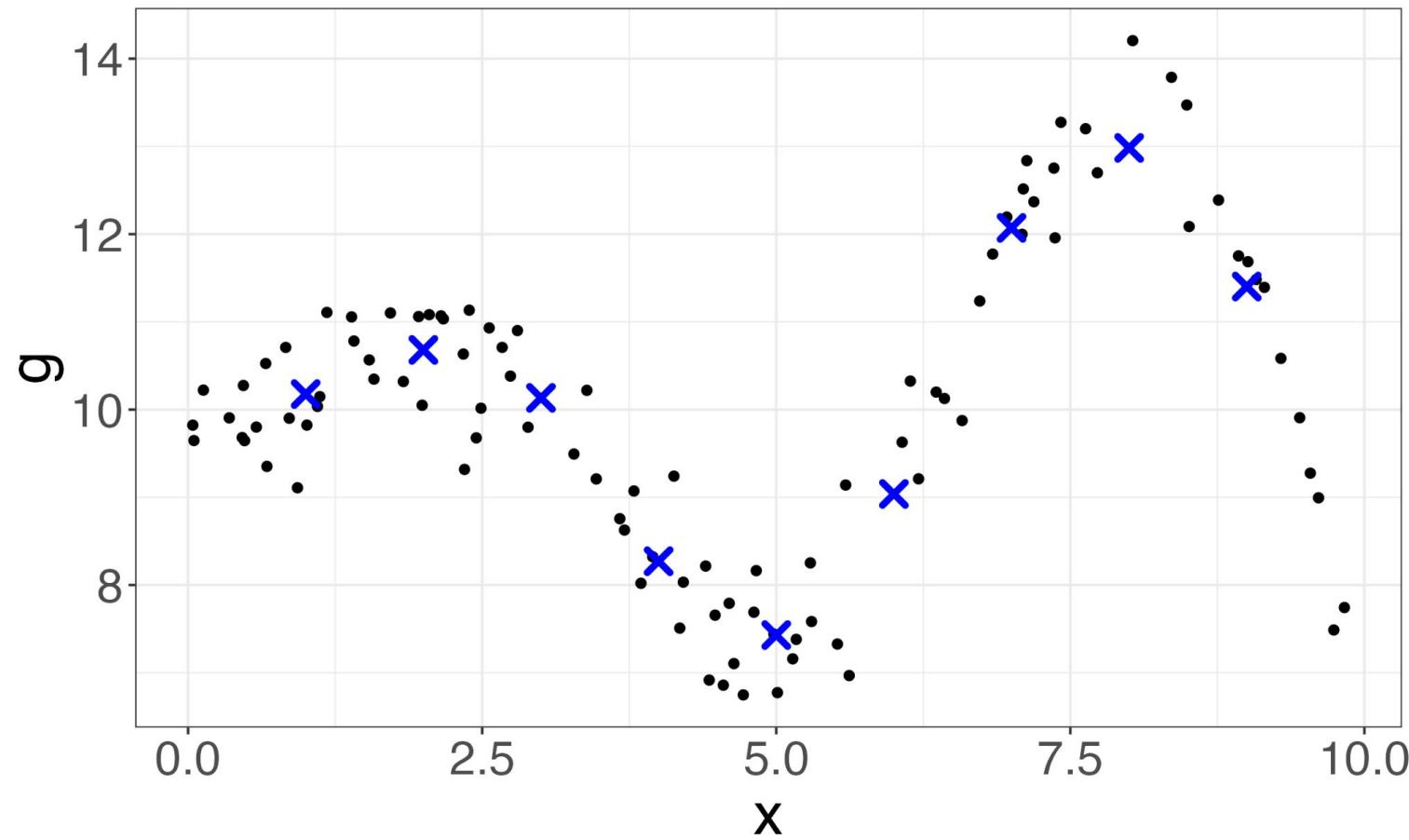




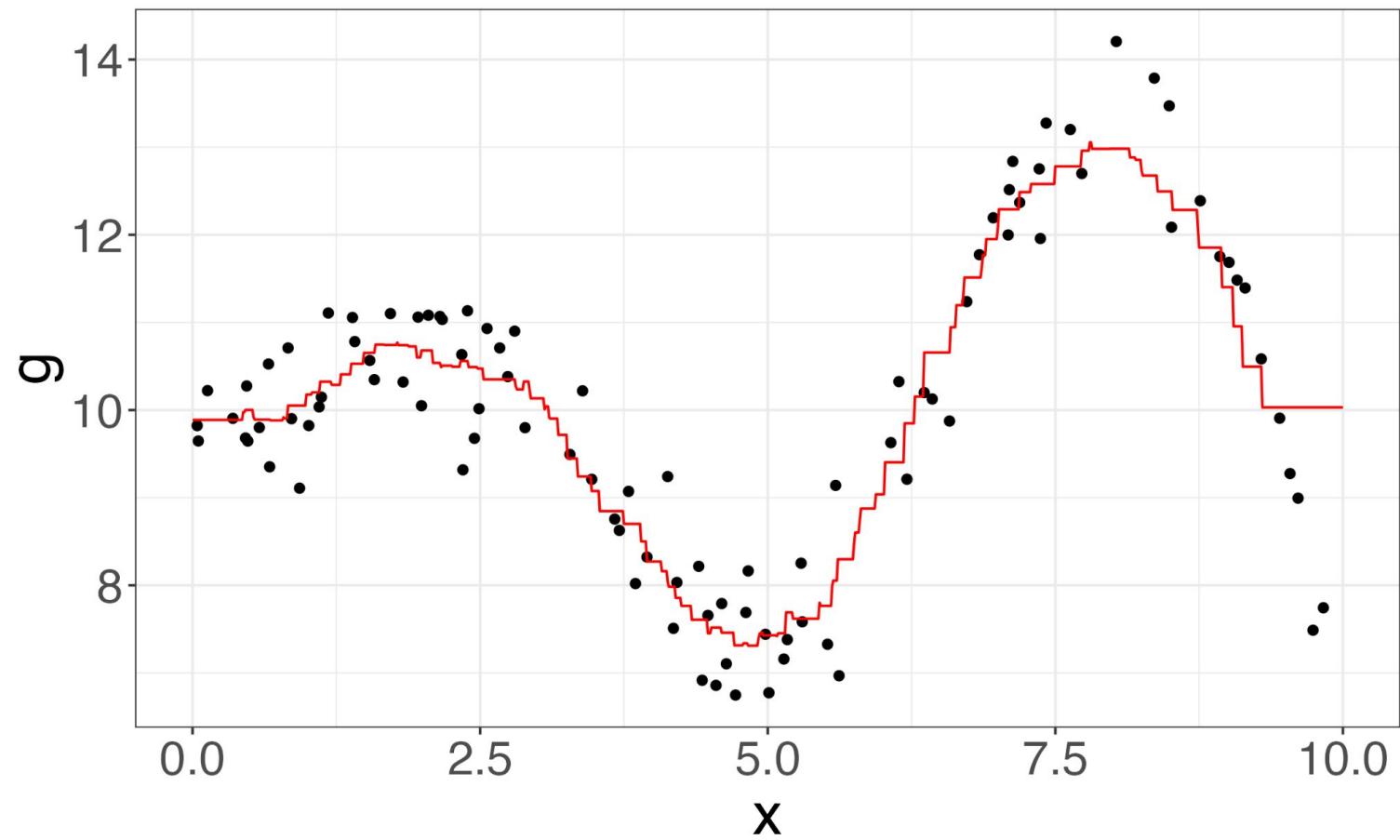




$K=10$



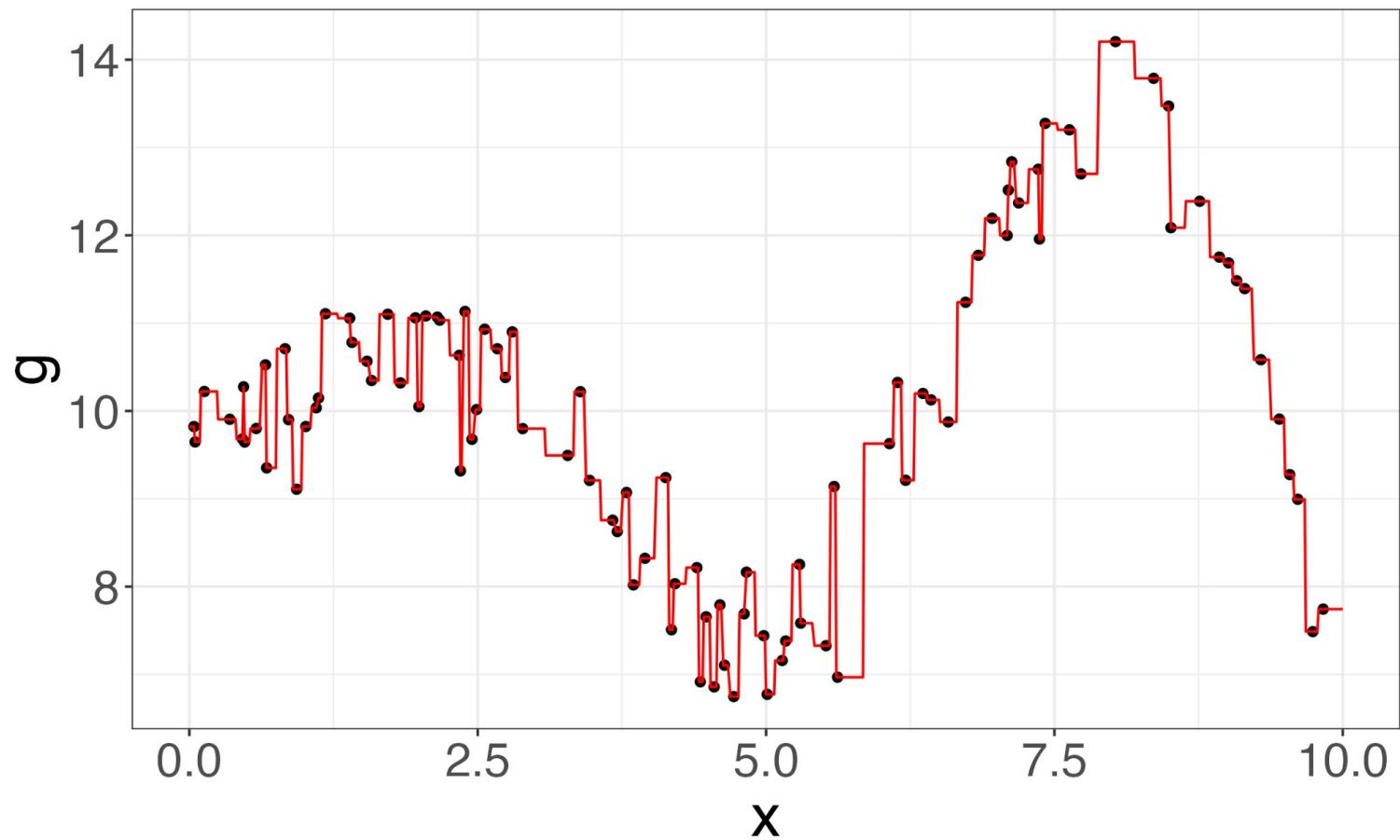
$K = 10$



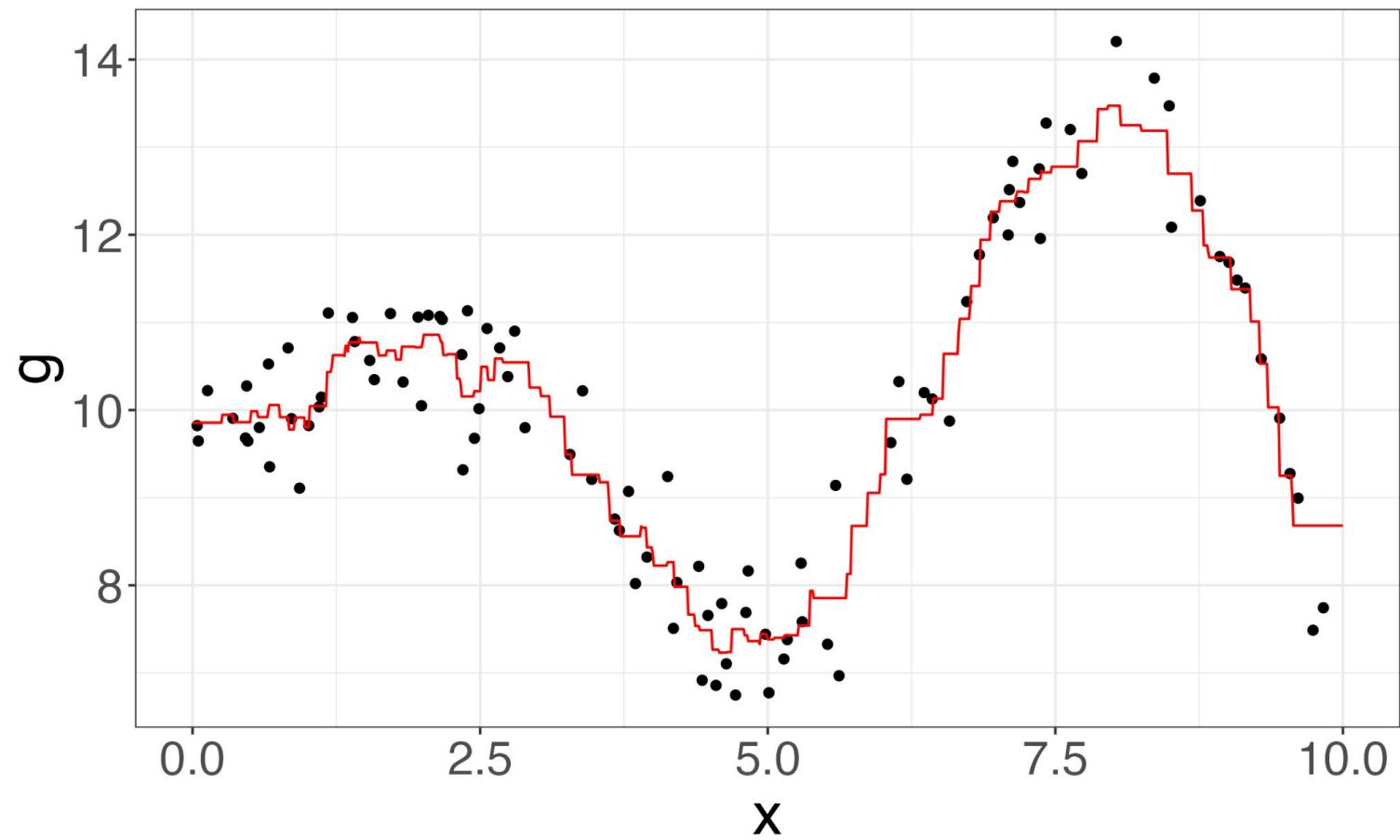
K-nearest neighbor regression (Fix and Hodges 1951)

Why 10 neighbors?

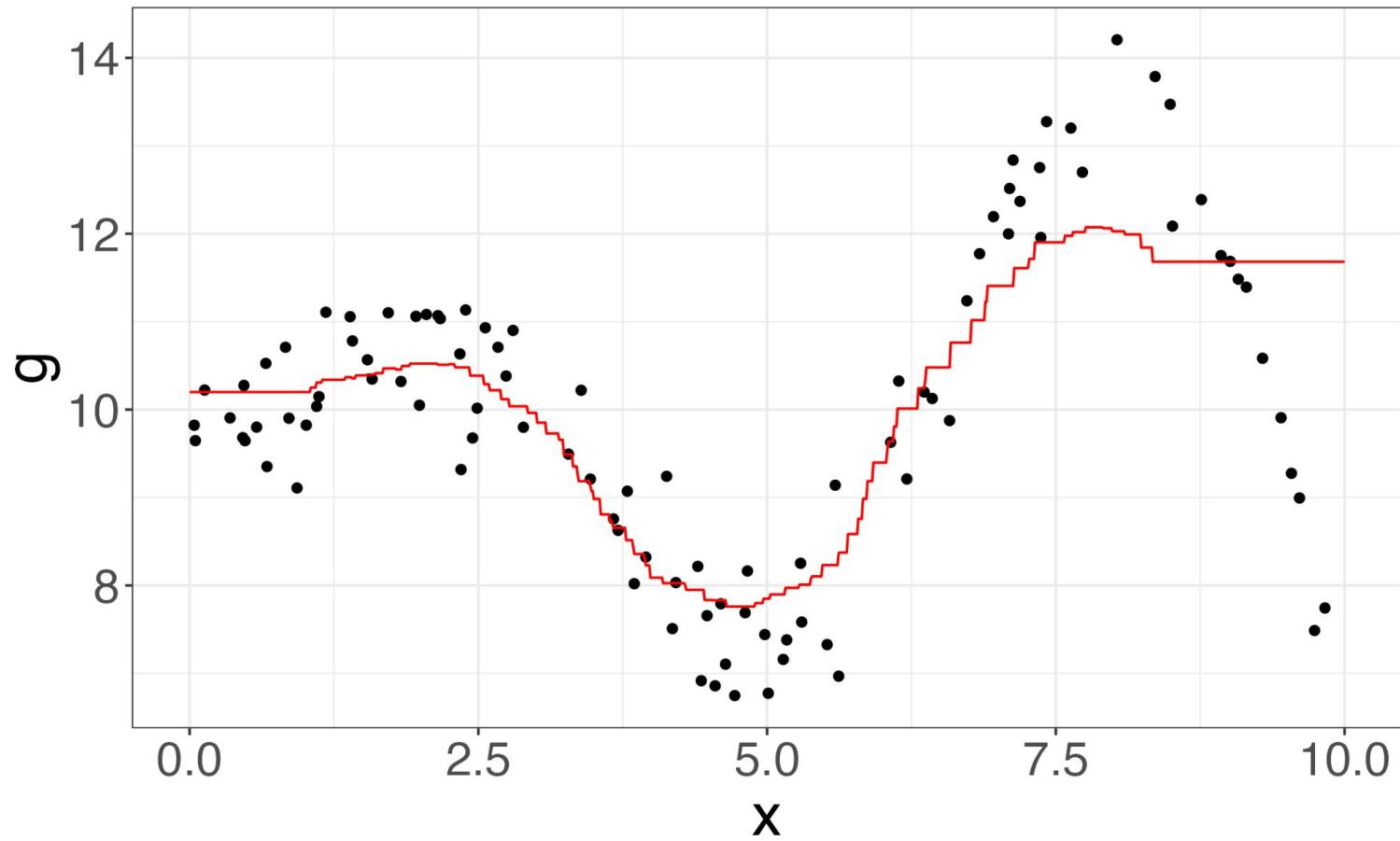
$K = 1$



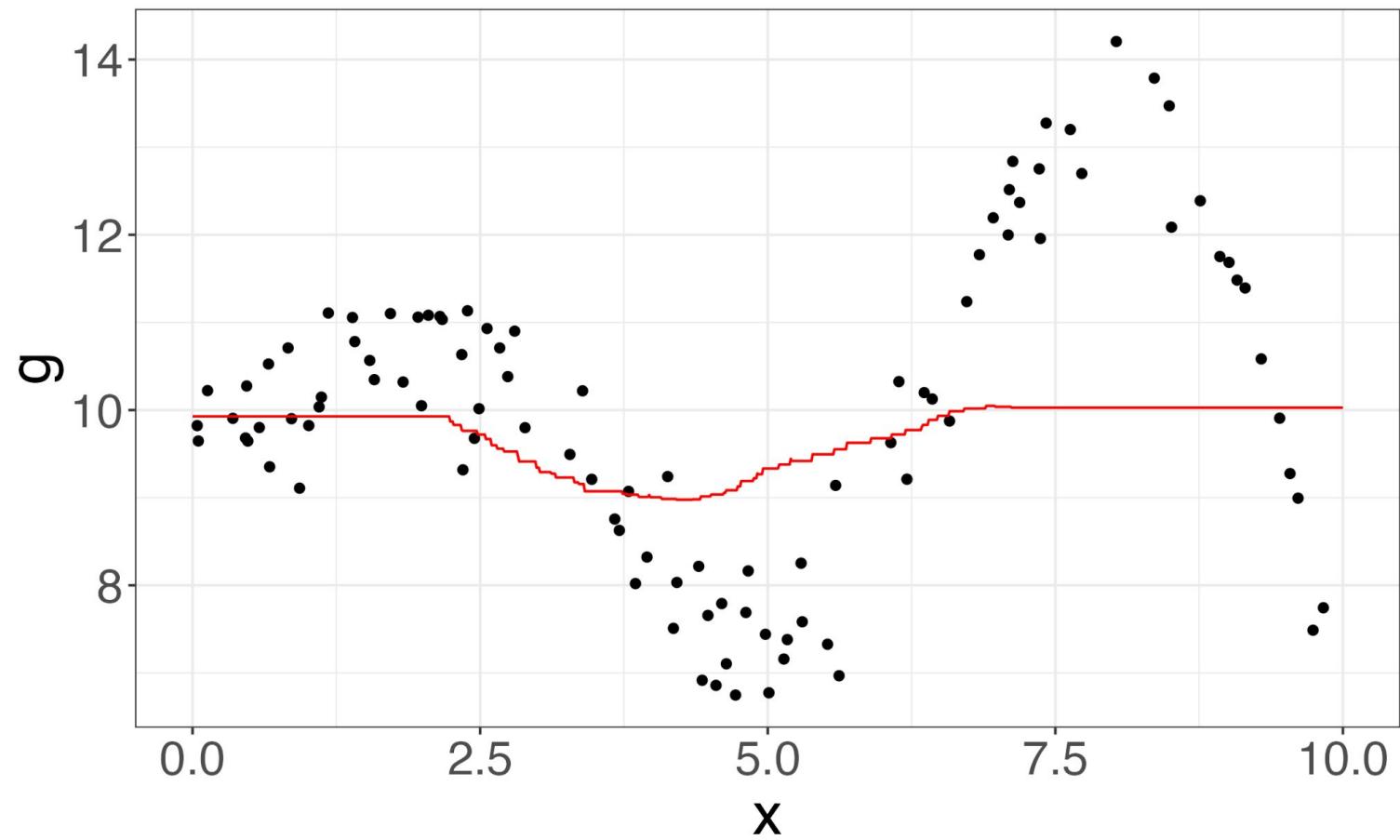
$K = 5$



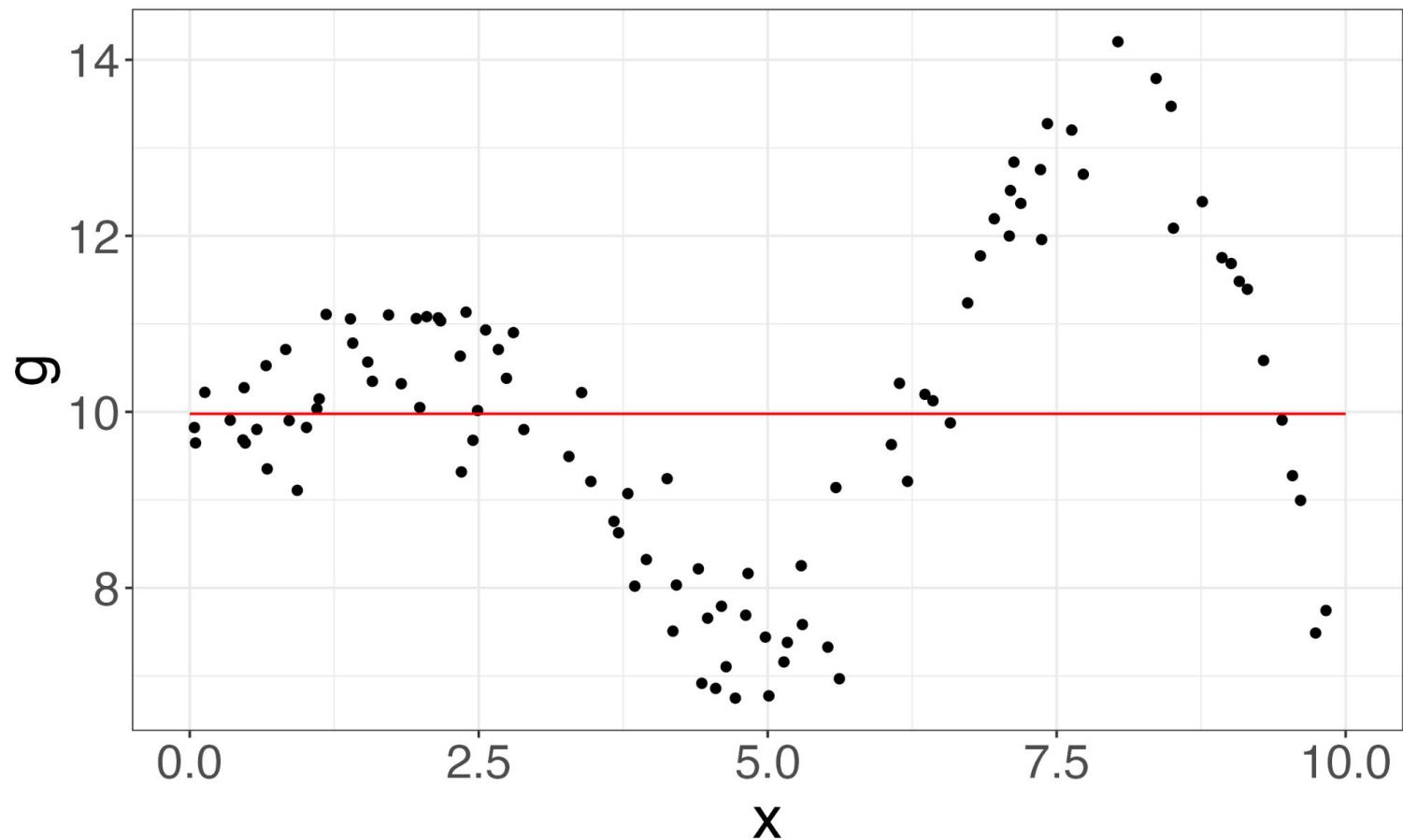
$K = 25$



$K = 50$



$K = 100$

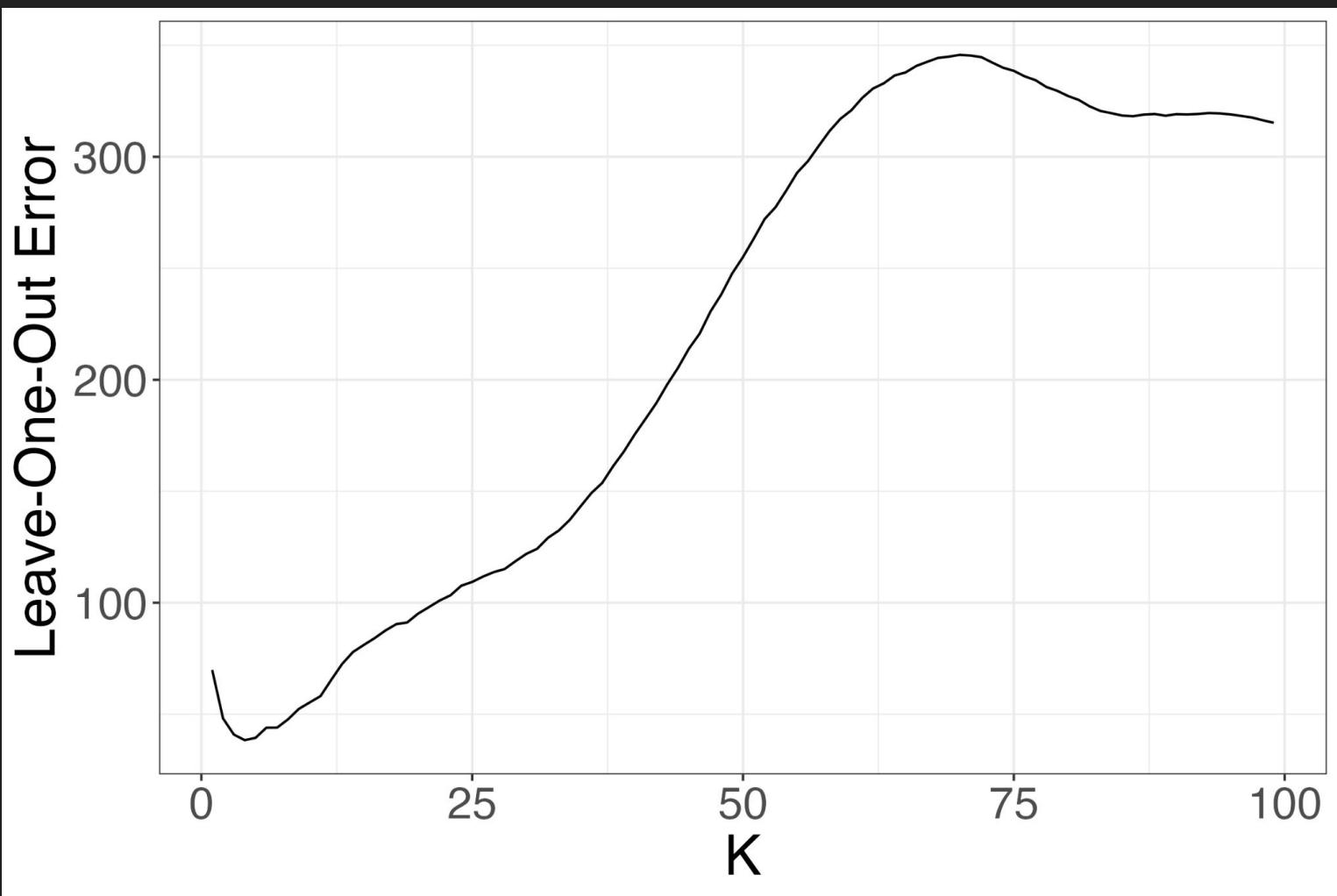


Overfitting
vs.
Oversmoothing

Bias-Variance Tradeoff

Choosing the right K

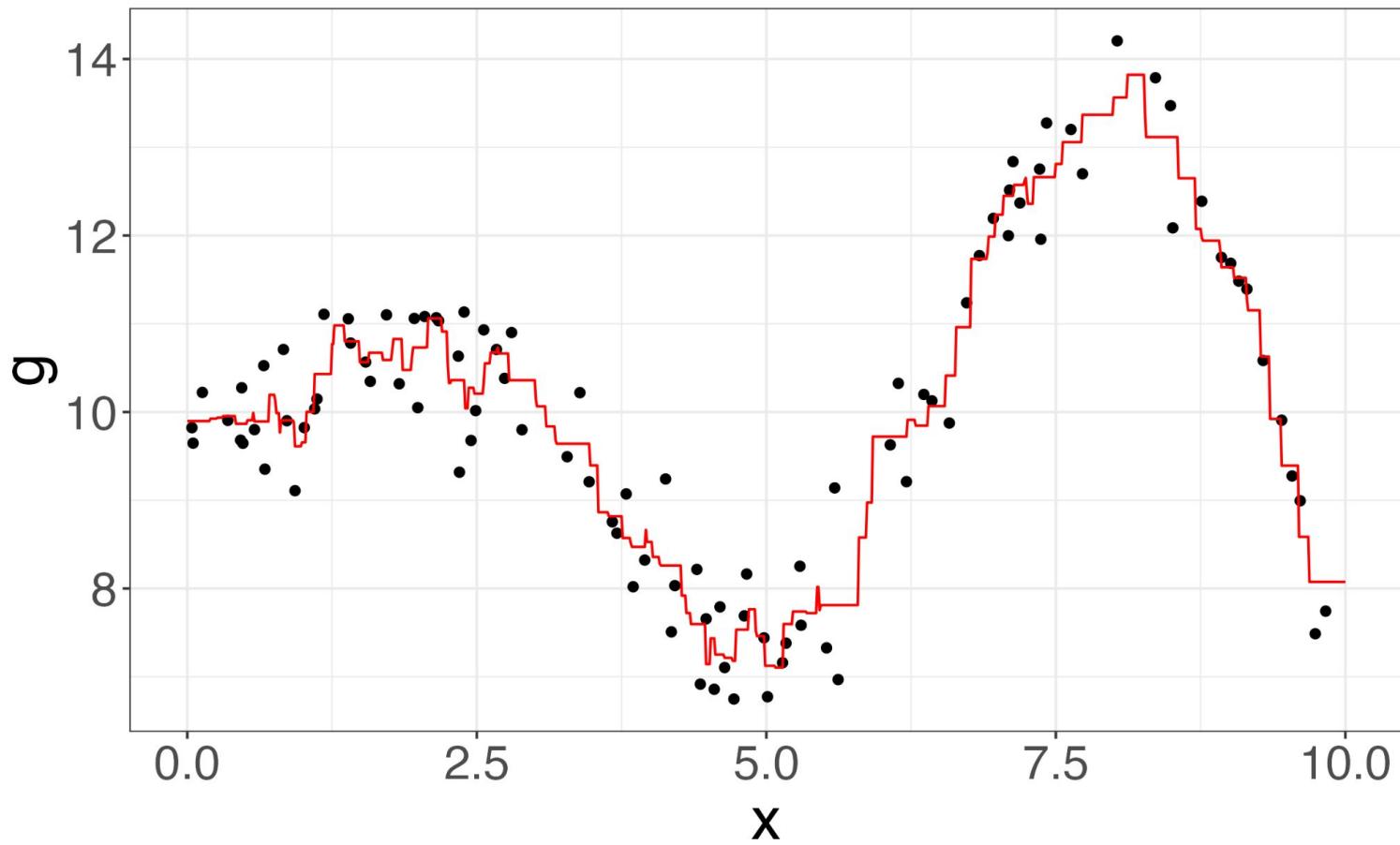
- Visual inspection
- Cross-validation: minimizing error on a validation set



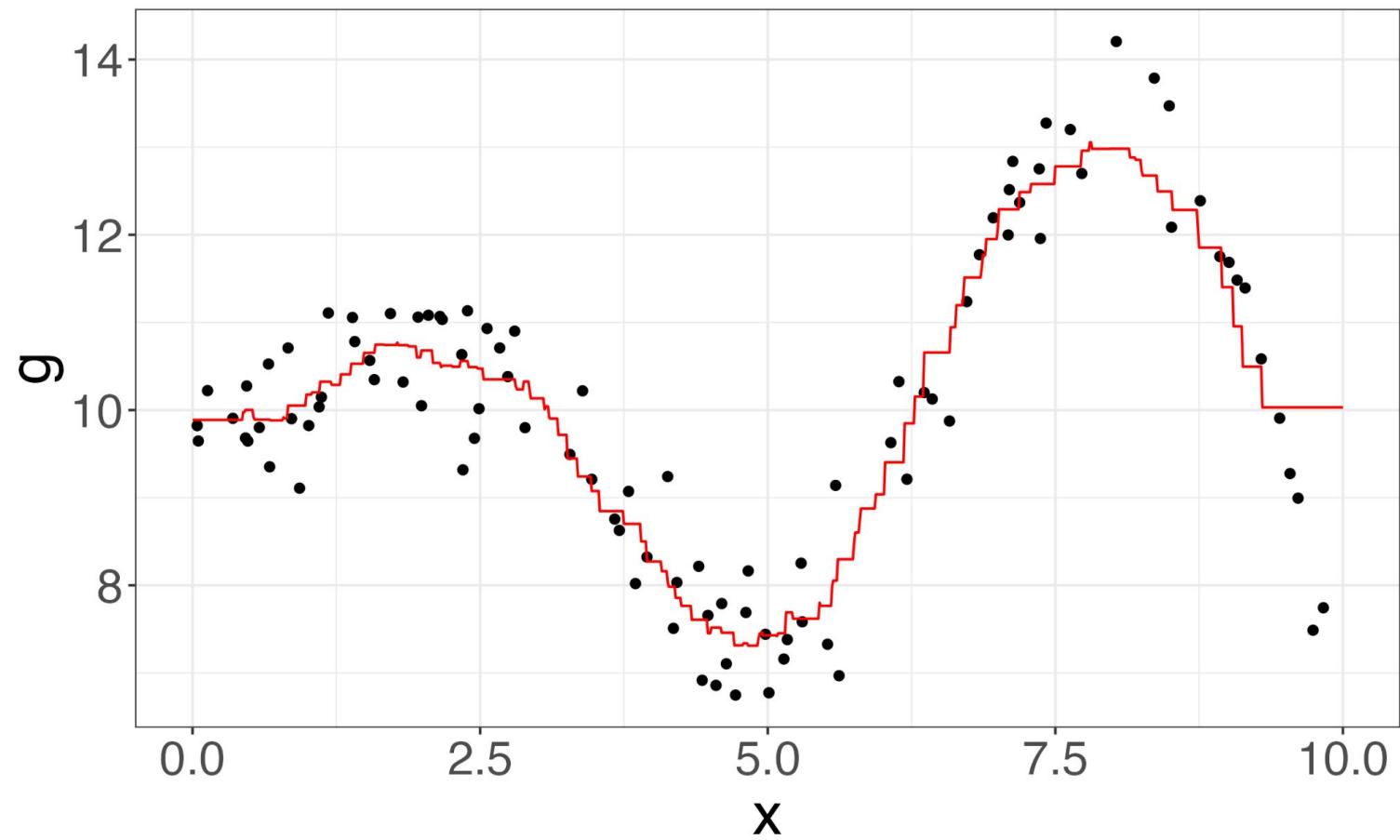
--Coding Demo--

Results of K-fold CV

$K = 3$

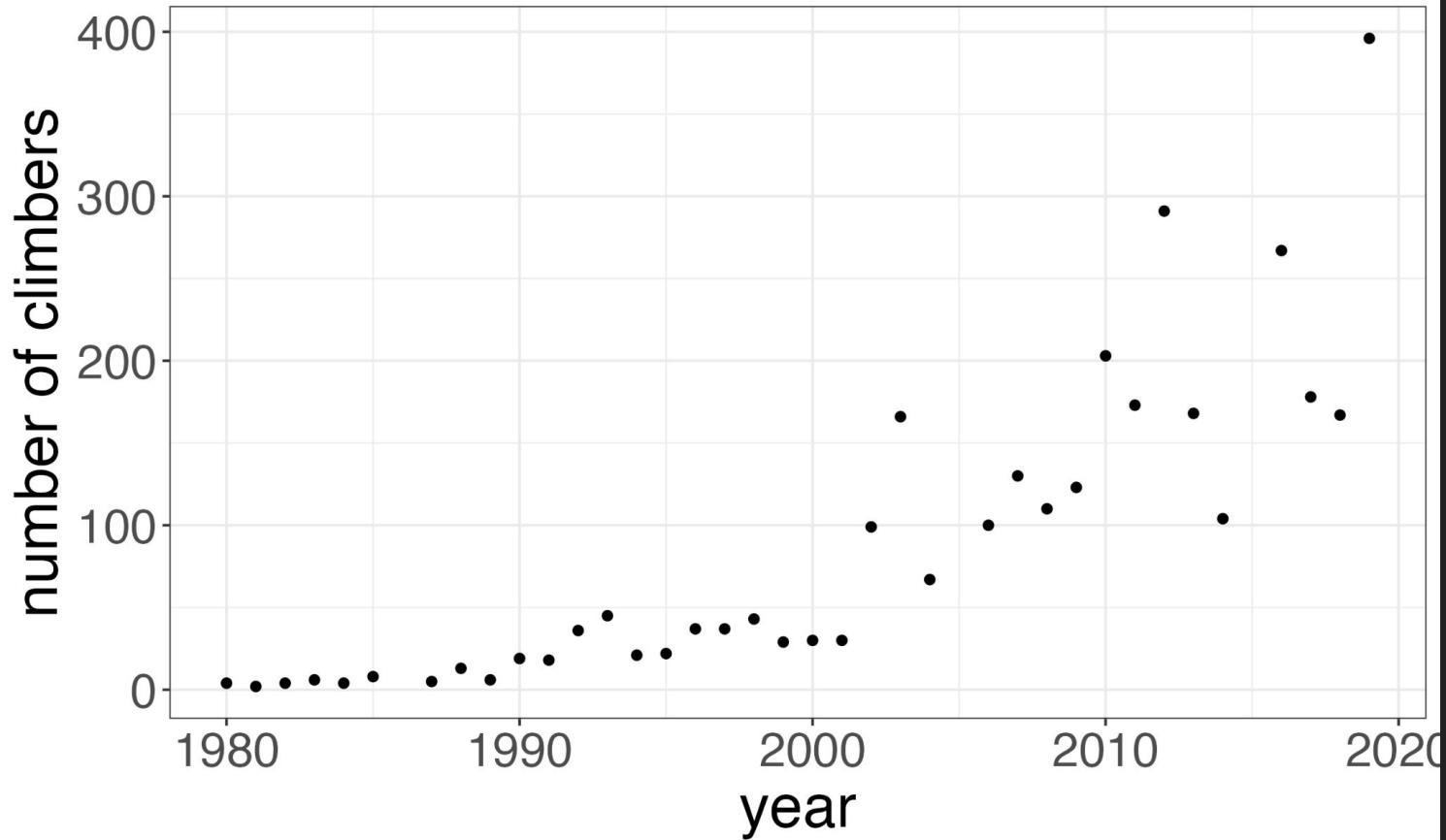


$K = 10$

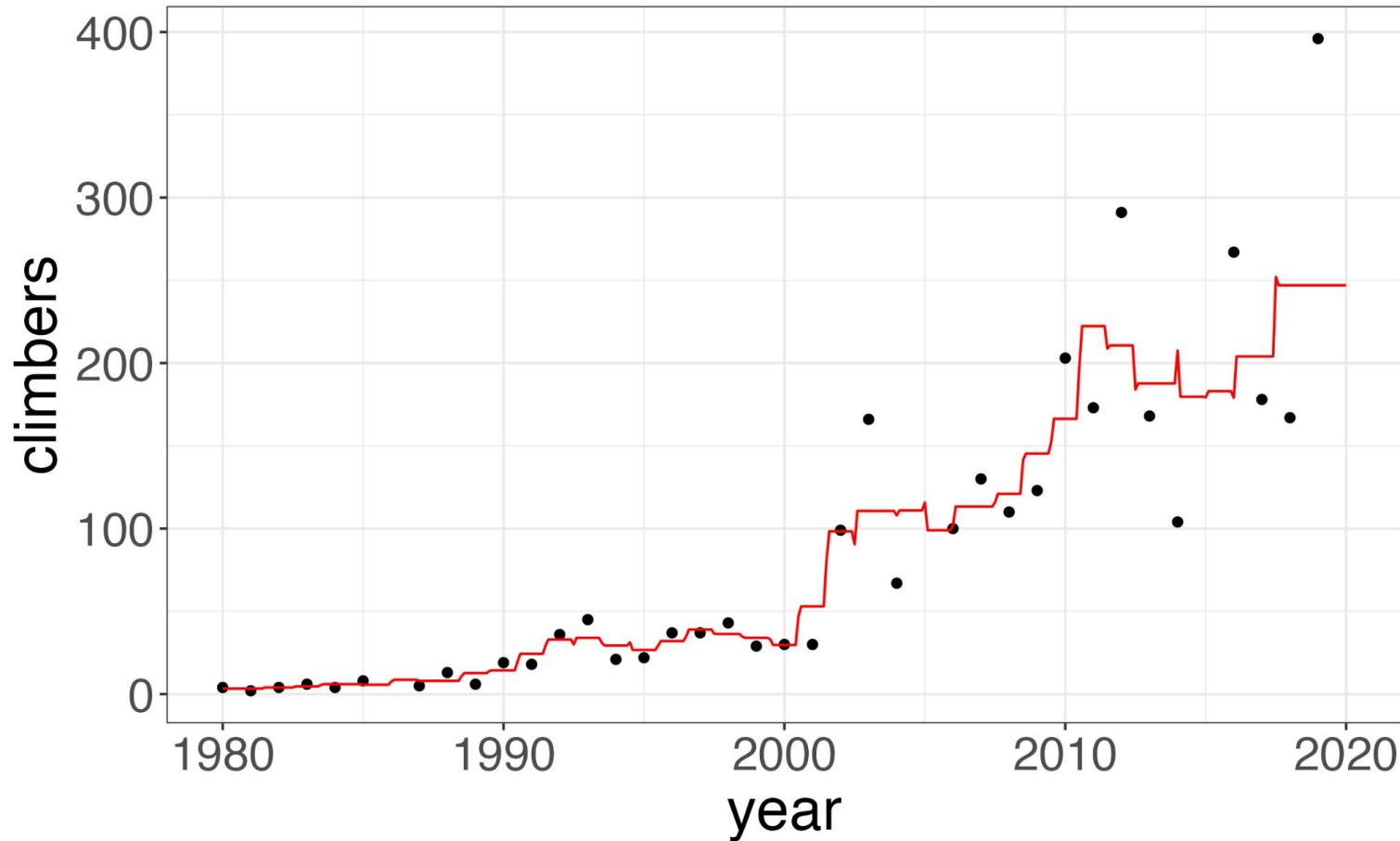


Back to Everest

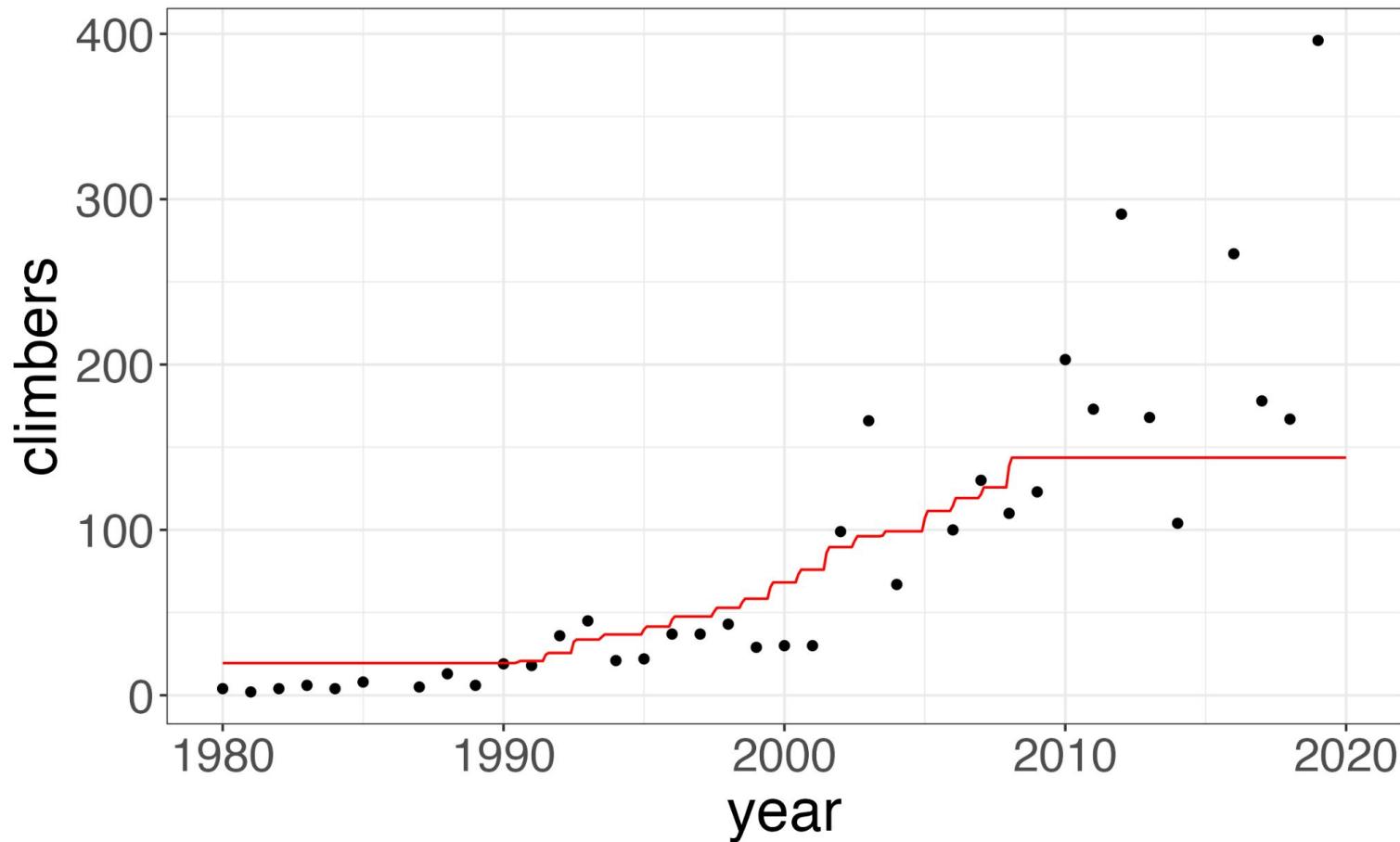
Crowding on Everest



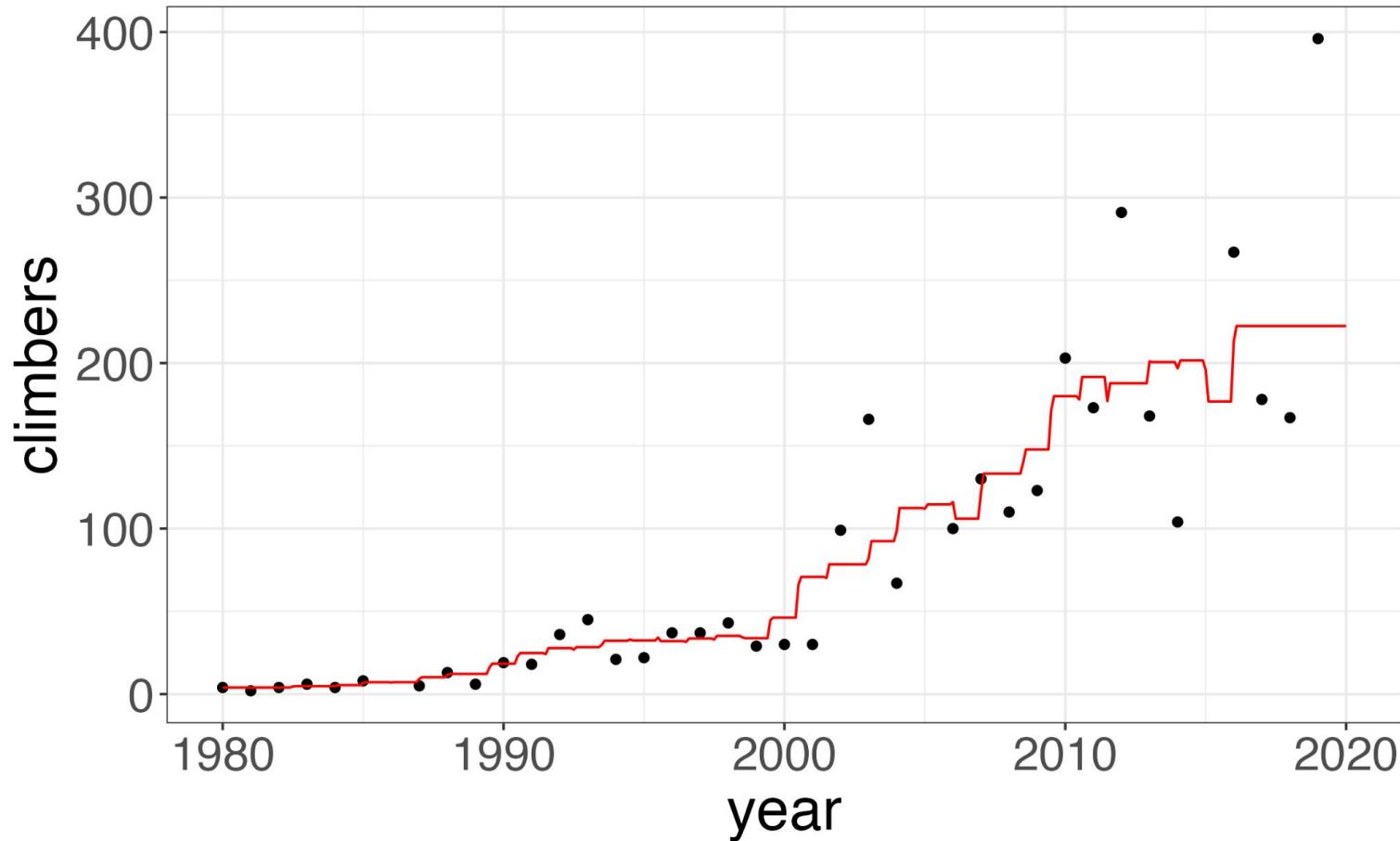
$K = 3$



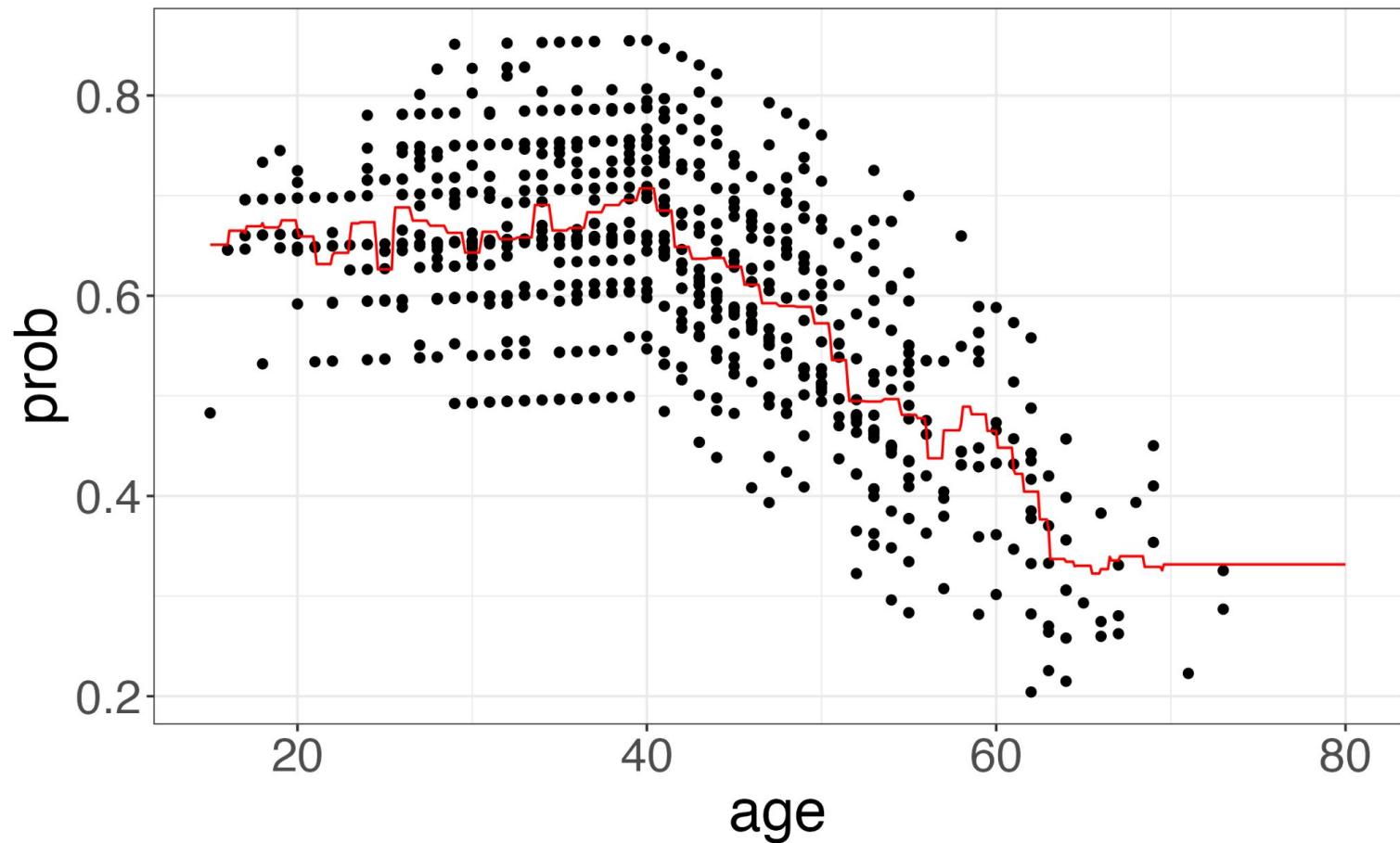
$K = 20$



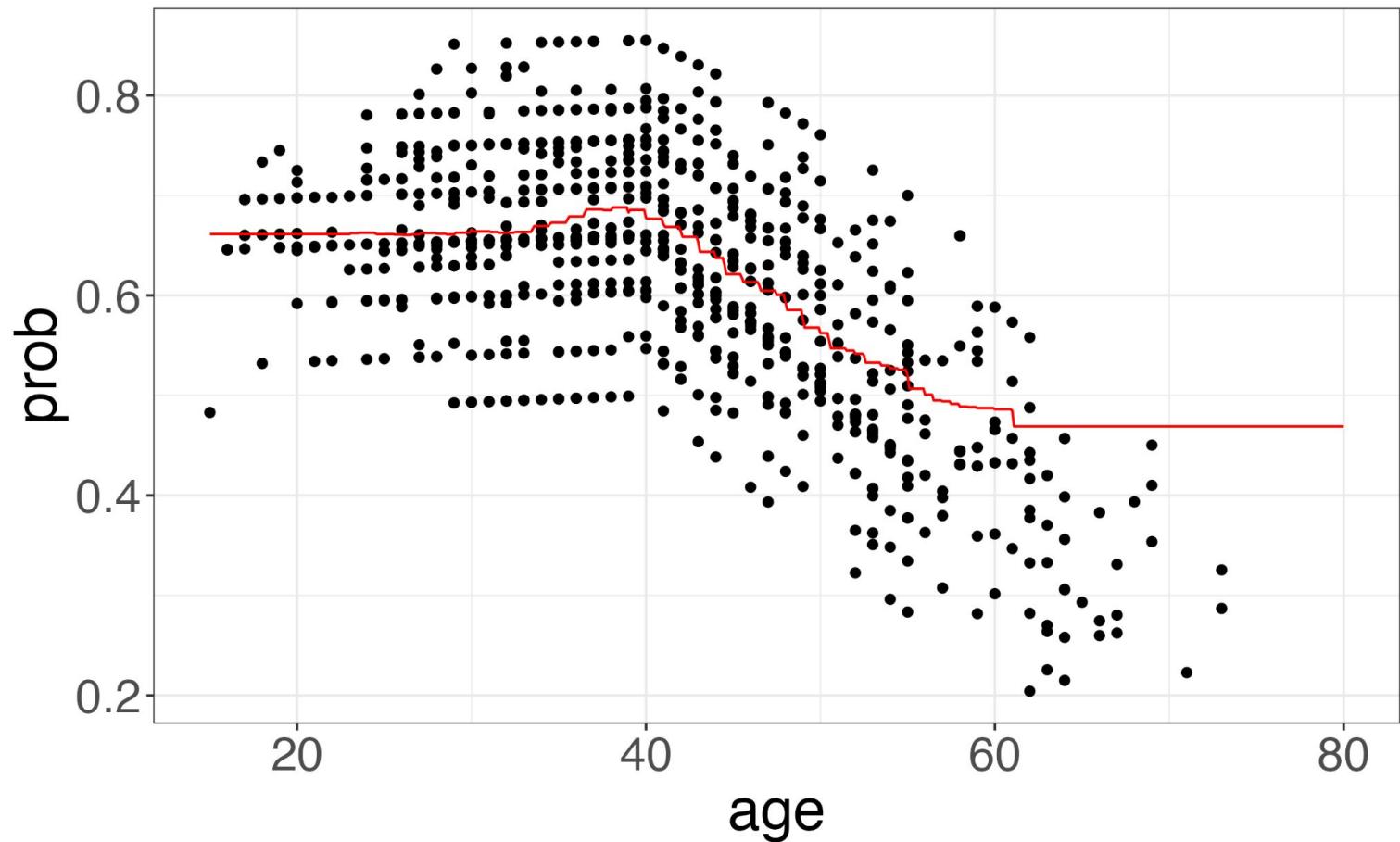
$K = 5$



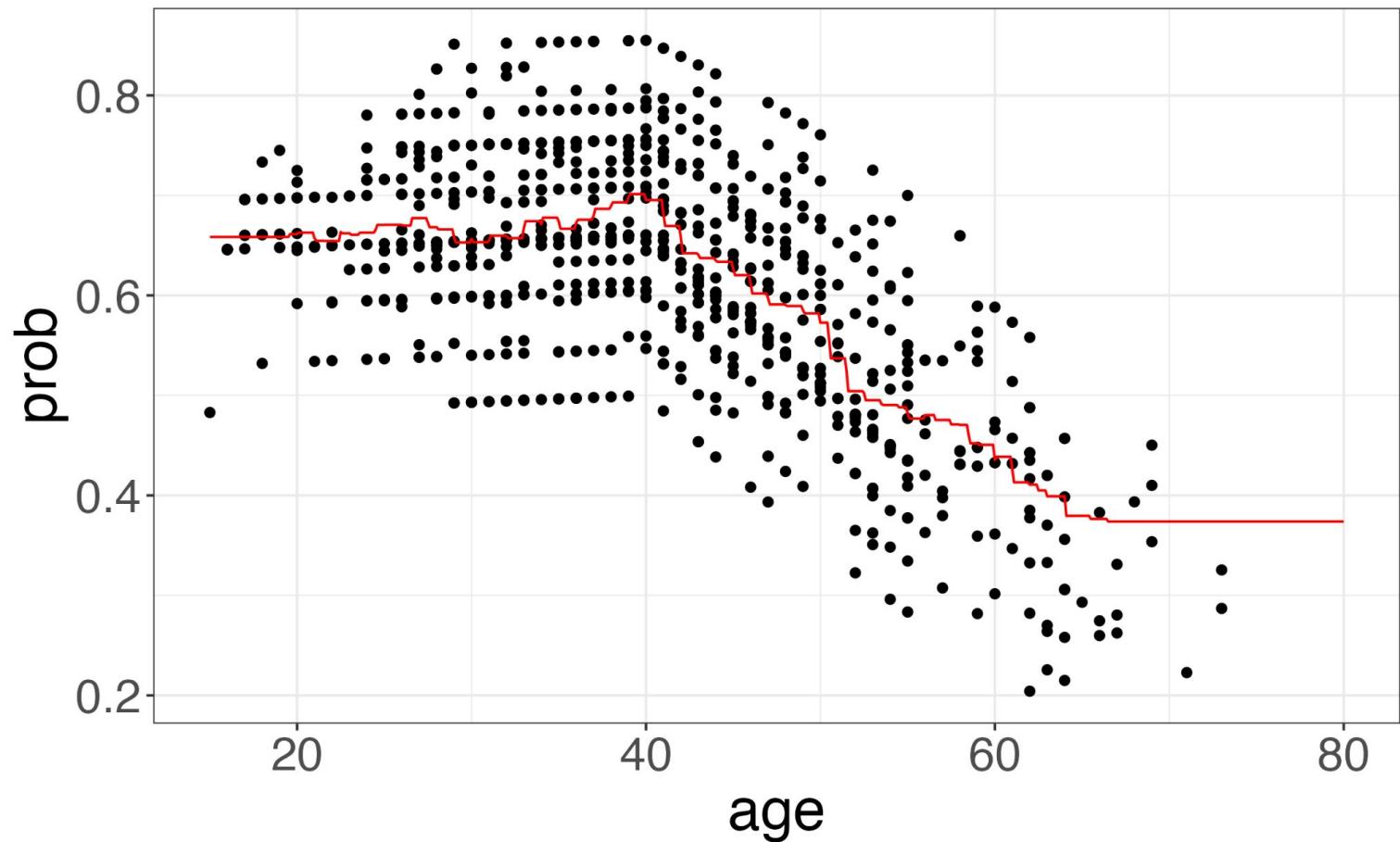
$K = 10$



$K = 200$



$K = 50$



Some things to notice

- KNN regression is a **weighted average**
- Estimate minimizes a **special loss function**

Some things to notice

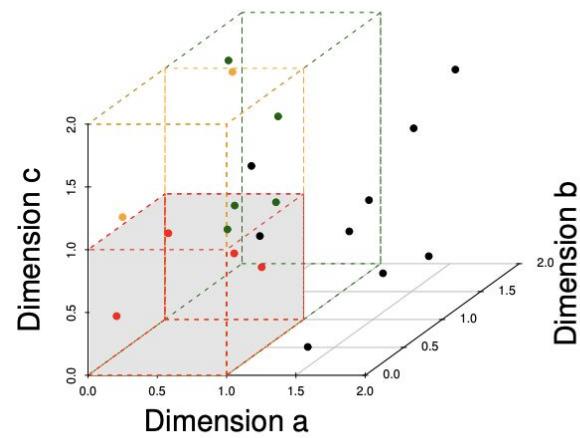
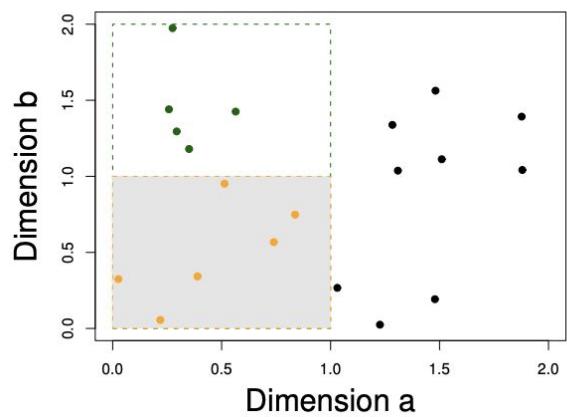
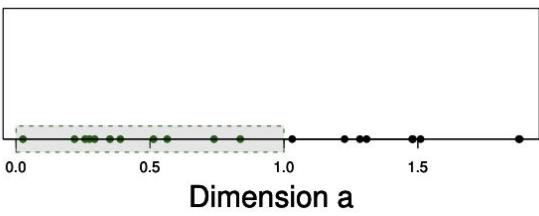
- Size of “neighborhood” changes with x
- Estimated curve is not continuous

Strengths of KNN Reg

- Intuitive
- Local
- No parametric form assumed
- Easy to implement in software

Drawbacks of KNN Reg

- Sensitive to outliers
- Noticeable bias at boundaries (asymmetric neighbors)
- Less easy to interpret
- Many predictors → curse of dimensionality (dissimilar neighbors)



Source: Parsons et al (2004)

Other local nonparametric regression techniques

- Moving averages / kernel regression (e.g. Nadaraya and Watson 1964)
- Local linear / polynomial regression [e.g. LOESS] (Gram 1879)

Kahoot!