Weighted Spline Models

We can always assume all x_i's are different, otherwise, we just need to fit a weighted regression model due to the following argument. Suppose the first two obs have the same x value, i.e.,

$$(x_1, y_1), \quad (x_2, y_2), \qquad \text{where } x_1 = x_2.$$

Then

$$\begin{aligned}
\big[y_1 - g(x_1)\big]^2 + \big[y_2 - g(x_1)\big]^2 &= \sum_{i=1}^{2} \big[y_i - \frac{y_1+y_2}{2} + \frac{y_1+y_2}{2} - g(x_1)\big]^2 \\
&= \big(y_1 - \frac{y_1+y_2}{2}\big)^2 + \big(y_2 - \frac{y_1+y_2}{2}\big)^2 \\
&\quad +2\big[\frac{y_1+y_2}{2} - g(x_1)\big]^2
\end{aligned}$$

So we can replace the first two obs by one, $(x_1, \frac{y_1+y_2}{2})$, and its weight is 2 while the weights for other obs are 1.

Summary: Smoothing Splines

- Start with a model with the maximum complexity: NCS with knots at n (unique) x points.
- Fit a Ridge Regression model on the data. If we parameterize the NCS function space by the DR basis, then the design matrix is orthogonal and the corresponding coefficient is penalized differently: no penalty for the two linear basis functions, higher penalties for wigglier basis functions.
- How to do it in `R`?
- How to select the tuning parameter λ or equivalently the df?
- What if we have collected two obs at the same location x?