Nonparametric regression – curve fitting

The relationship between the effect and the cause is not necessarily linear.

Nonlinear regression requires numerical optimization, therefore it is more difficult and computationally more intensive to fit than linear regression.

$(X_i, Y_i)_{i=1,\dots,n}$ observations i.i.d.
$(x_i, y_i)_{i=1,\dots,n}$ – the statistical data.

Nonparametric regression – curve fitting

interpolation and smoothing techniques are used to estimate values between our data points $(x_i, y_i)_{i=1,\dots,n}$ and then, to smooth the data.

The conditional expectation is a function with "smoothing" properties:

$$E(X \mid Y = y) = g(y), \text{ with } g : [a,b] \to \mathbb{R}$$

$$\exists\, g'(y),\ g''(y) \text{ continuous functions}$$

The nonparametric regression curve is:

$$X = g(y) + \xi$$

Nonparametric regression – curve fitting

The sum of square deviations with penalty is:

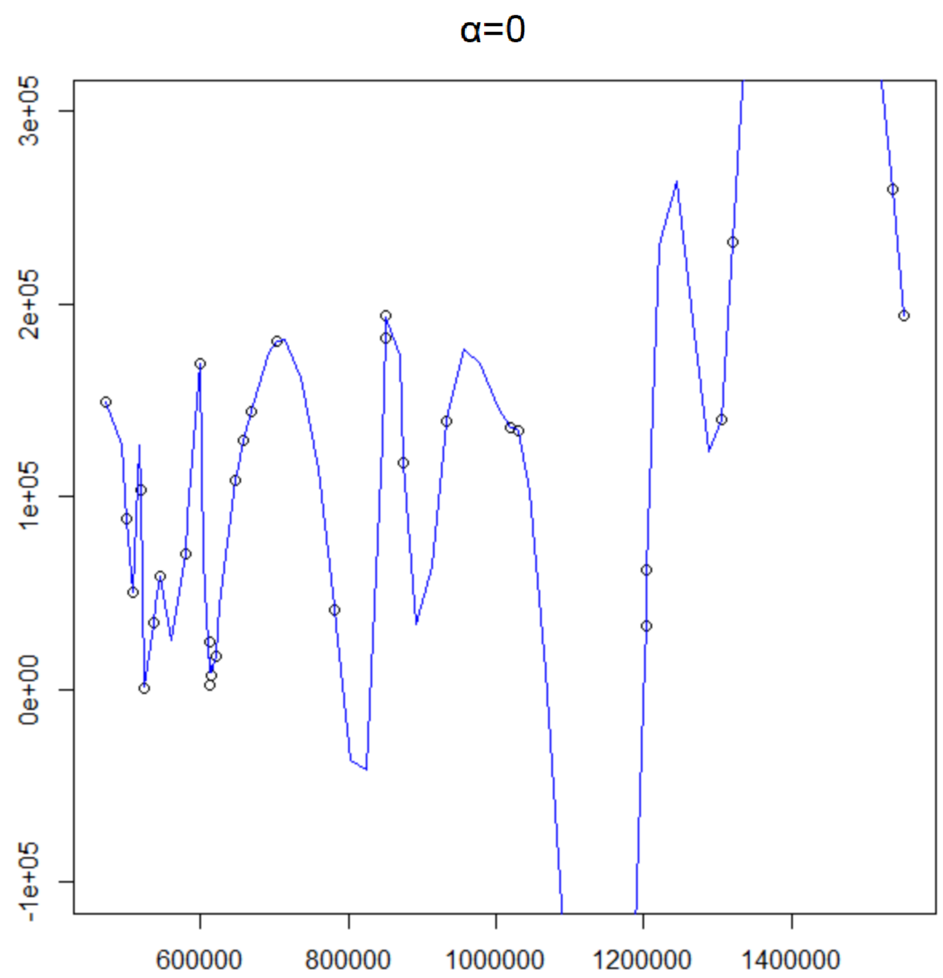
$$SS(g) = \sum_{i=1}^{n} \left(x_i - g(y_i)\right)^2 + \alpha \cdot \int_a^b \underbrace{\left(g''(y)\right)^2}_{\text{it measures the curvature in } g \text{ at } y.} dy$$

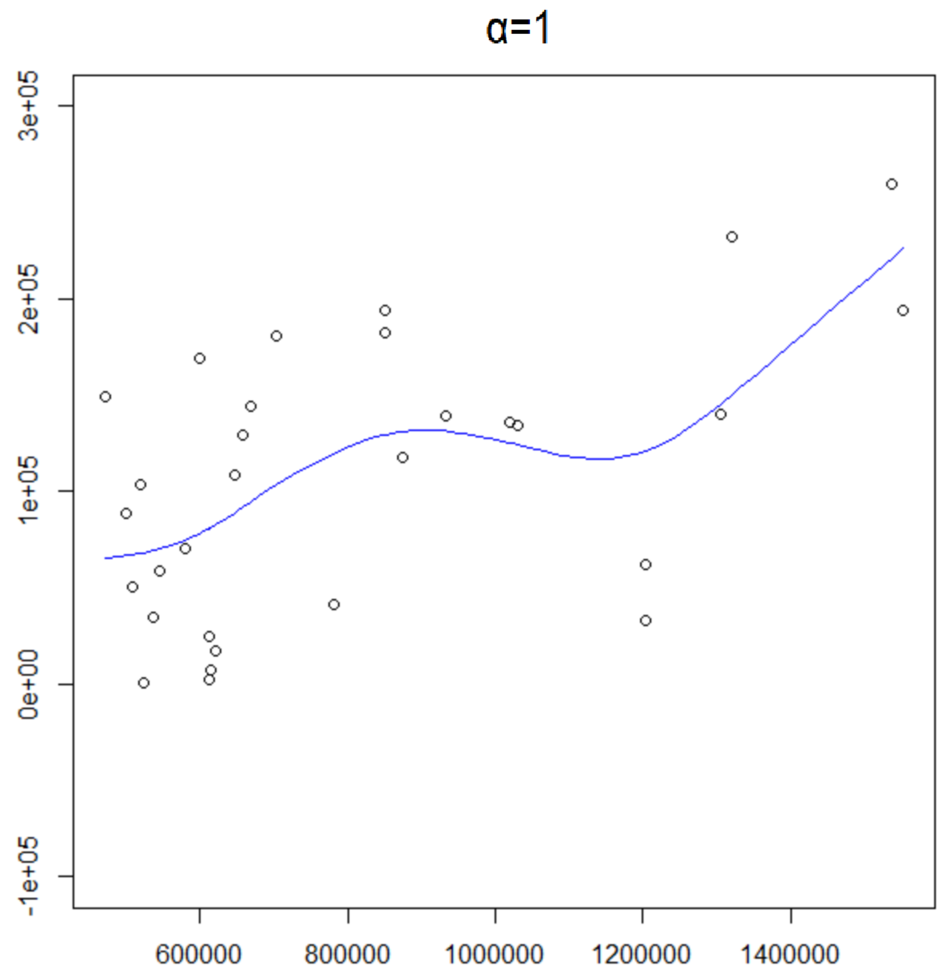
$\alpha > 0$ is the smoothing parameter

The "Least Squares with penalty" estimator is the solution of the optimization problem $\inf_g SS(g)$

Nonparametric regression – curve fitting

Obs. if α is small then the regression curve will "follow" closely the statistical data.
if α is high, the regression curve will be smoother.





Nonparametric regression – curve fitting

Cubic splines

Splines provide a way to smoothly interpolate between fixed points.

Def Suppose that $a < y_1 < y_2 < \dots < y_n < b$ and $g : [a,b] \to \mathbb{R}$.
The function g is called cubic spline if:

- g is a 3rd degree polynomial on each subinterval
$$(a, y_1), (y_1, y_2), \dots, (y_n, b)$$
- g is continuous, with g' and g'' continuous on $[a,b]$.

Nonparametric regression – curve fitting

$\{y_i, i=1,\dots,n\}$ are called <u>knots</u>.

<u>Convention:</u> $y_0 = a$, $y_{n+1} = b$

Using the base $\{y^3, y^2, y, 1\}$, a cubic spline has the form:

$$g(y) = d_i(y-y_i)^3 + c_i(y-y_i)^2 + b_i(y-y_i) + a_i \,, \quad y_i \le y \le y_{i+1}, \quad i=0,\dots,n$$

with the condition for continuity in y_{i+1}, $i=0,\dots,n-1$

$$d_i(y_{i+1}-y_i)^3 + c_i(y_{i+1}-y_i)^2 + b_i(y_{i+1}-y_i) + a_i = a_{i+1}$$

Nonparametric regression – curve fitting

<u>Def</u>. A cubic spline over $[a,b]$ is called <u>natural</u> <u>cubic</u> <u>spline</u> (NCS) if its 2^{nd} and 3^{rd} order derivatives are null in a and b.

<u>Obs</u>. From
$$\left.\begin{array}{l} g''(a) = g''(b) = 0 \\ g'''(a) = g'''(b) = 0 \end{array}\right\} \Rightarrow d_0 = c_0 = d_n = c_n = 0$$

$\Rightarrow g$ is liniar on $[a, y_1]$ and $[y_n, b]$

Nonparametric regression – curve fitting

Notations: $g_i = g(y_i), \quad i = 1, \ldots, n$

$$\underline{g} = (g_1, \ldots, g_n)'$$

$$\gamma_i = g''(y_i), \quad i = 1, \ldots, n \quad (\gamma_1 = \gamma_n = 0 \text{ because } g \text{ is linear on } [a, y_1], [y_n, b])$$

$$\underline{\gamma} = (\gamma_2, \ldots, \gamma_{n-1})'$$

$$h_i = y_{i+1} - y_i, \quad i = 1, \ldots, n-1$$

$$q_{ij} = \begin{cases} \frac{1}{h_{j-1}}, & i = j-1 \\ -\frac{1}{h_{j-1}} - \frac{1}{h_j}, & i = j \\ \frac{1}{h_j}, & i = j+1 \\ 0, & |i-j| \geq 2 \end{cases} \qquad Q = \|q_{ij}\|_{\substack{i=1,\ldots,n \\ j=2,\ldots,n-1}}$$

$$\begin{cases} r_{ii} = \frac{1}{3}(h_{i-1} + h_i), & i = 2, \ldots, n-1 \\ r_{i,i+1} = r_{i+1,i} = \frac{1}{6} h_i, & i = 2, \ldots, n-1 \\ r_{ij} = 0, & |i-j| \geq 2 \end{cases} \qquad R = \|r_{ij}\|_{\substack{i=2,\ldots,n-1 \\ j=2,\ldots,n-1}}$$

Nonparametric regression – curve fitting

Proposition g can be specified by using $\underline{g}, \underline{\gamma}$ and $h_i, i=1,\dots,n-1$ as following:

$$
\begin{cases}
g(y) = g_1 - (y_1 - y)\left(\dfrac{g_2 - g_1}{h_1} - \dfrac{1}{6} h_1 \gamma_2\right), & a \le y \le y_1 \\[2ex]
g(y) = \dfrac{1}{h_i}\left[(y-y_i)\, g_{i+1} + (y_{i+1}-y)\cdot g_i\right] - \\
\quad - \dfrac{1}{6}(y-y_i)(y_{i+1}-y)\left[\left(1+\dfrac{y-y_i}{h_i}\right)\gamma_{i+1} + \left(1+\dfrac{y_{i+1}-y}{h_i}\right)\gamma_i\right], & y_i \le y \le y_{i+1}, \\
 & i = 1, \dots, n-1 \\[2ex]
g(y) = g_n + (y-y_n)\left(\dfrac{g_n - g_{n-1}}{h_{n-1}} + \dfrac{1}{6} h_{n-1}\gamma_{n-1}\right), & y_n \le y \le b
\end{cases}
$$

Theorem Given the knots $a < y_1 < y_2 < \dots < y_n < b$ and $\underline{g}, \underline{\gamma}$ and $h_i, i=1,\dots,n-1$, the function g built as described in the Proposition above is NCS iff $Q'\underline{g} = R\cdot\underline{\gamma}$

Nonparametric regression – curve fitting

<u>Corollary</u> if $Q'\underline{g} = R\cdot\underline{\gamma}$ then

$$\underbrace{\int_a^b (g''(y))^2\,dy}_{} = \underline{\gamma}'R\cdot\underline{\gamma} = \underline{g}'K\underline{g}, \text{ where } K = QR^{-1}Q'$$

this is the measure of the total curvature - in $SS(g)$

<u>Interpolation using NCS</u>

$(x_i, y_i)_{i=1,\dots,n}$ the statistical data $n \geq 2$

$a < y_1 < y_2 < \dots < y_n < b$ the knots

We search for a curve to fit $g(y_i) = x_i$, $i = 1, \dots, n$

Nonparametric regression – curve fitting

<u>Proposition</u> There is a unique NCS g, with the knots $y_1, \dots, y_n$ that satisfies $g(y_i) = x_i$, $i = 1, \dots, n$.

<u>Algorithm for interpolation</u>

<u>Input</u>: (x_i, y_i), $i = 1, \dots, n$

1. $g_i = x_i$, $i = 1, \dots, n$
2. with the previous notations, compute $\mathbf{x} = Q' g$
3. solve the system $R \cdot \underline{\gamma} = \mathbf{x} \Rightarrow \underline{\gamma}$
4. having g, $\underline{\gamma}$ and the knots, we built the NCS interpolant g according to the Proposition on slide 9.

<u>Output</u>: g

Nonparametric regression – curve fitting

<u>Proposition</u> The NCS interpolant g is the solution of

$$\inf_{\substack{\tilde{g} \text{ interpolant} \\ \tilde{g}(y_i)=x_i,\ i=1,\dots,n \\ \tilde{g} \text{ is differentiable}}} SS(\tilde{g})$$

<u>Obs</u>. For any interpolant $\tilde{g}$, $SS(\tilde{g}) = \alpha \int_a^b (\tilde{g}''(y))^2 dy$

(because $\tilde{g}(y_i) = x_i$, $i=1,\dots,n$)

Nonparametric regression – curve fitting

Smoothing using NCS

$(x_i, y_i)_{i=1,\dots,n}$ – the statistical data $n \geq 3$

$a < y_1 < \dots < y_n < b$

The regression curve is $X = g(y) + \xi$.

Proposition. if $\hat{g}$ is the solution of the optimization problem

$$\inf_{\substack{g \text{ is differentiable} \\ \text{on } [a,b]}} SS(g)$$

then $\hat{g}$ is NCS.

Nonparametric regression – curve fitting

<u>Obs</u>. Unlike the interpolation problem, here we don't know whether $\hat{g}(y_i) = x_i$, $i = 1, \dots, n$.

<u>The Reinsch algorithm for smoothing using NCS</u>

<u>Input</u>: $(x_i, y_i)_{i=1,\dots,n}$, $\alpha > 0$

1. Compute $z = Q' \cdot (x_1, \dots, x_n)'$
2. $R + \alpha Q'Q$ is symmetric and positive definite – it admits a Cholesky decomposition
$$R + \alpha Q'Q = L D L',$$
where D is diagonal, positive definite and L has the properties
$$\begin{cases} l_{ii} = 1 & \forall i = 1, n-2 \\ l_{ij} = 0 & \text{for } \forall j,\ 2 < j < i-2,\ i = 6, \dots, n-2 \end{cases}$$

Nonparametric regression – curve fitting

3. Solve the system $LDL'\underline{\gamma} = z$ and we get $\underline{\gamma}$

4. Compute $\hat{g} = (x_1, \dots, x_n)' - \alpha Q \underline{\gamma}$

5. Having $\hat{g}$, $\underline{\gamma}$ and the knots $y_1, \dots, y_n$, we built the NCS smoothing function $\hat{g}$ as indicated in the Proposition on slide 9.

<u>Output</u>: $\hat{g}$ - the smoothing function of the regression curve

<u>Obs</u>. R and Q are the matrices defined at slide 8.
$R + \alpha Q'Q$ is decomposed into LDL' for computational purpose.