Arbeitsblatt 8

A 1 Arbeiten Sie weiter mit dem Datensatz mcycle.

a) Schätzen Sie $\hat{f}$ jeweils über eine quadratische und eine kubische Regression Spline mit einem Knoten im Median der Daten: Erstellen Sie dazu jeweils eine geeignete Designmatrix mithilfe der trunkierten Potenzfunktionen als Basiserweiterung in X und wenden Sie dann die Funktion lm an.

b) Bestimmen Sie für jedes Modell die optimale Anzahl an Knoten K_{CV} mit Kreuzvalidierung. Dabei sollen die Knotenpunkte jeweils an den entsprechenden Quantilen liegen. Sie können hier die Basisfunktionserweiterung über B-Splines nutzen, die in der Funktion bs aufgestellt wird.

c) Berechnen Sie die Anzahl an freien Parametern in den vier Modellen (quadratisch/kubisch, 1 Knoten/K_{CV} Knoten)

d) Wir definieren nun k als die Anzahl freier Parameter der kubischen Regression Spline mit K_{CV} Knoten. Schätzen Sie $\hat{f}$ auch als natürliche kubische Regression Spline mit k freien Parametern

e) Schätzen Sie $\hat{f}$ auch als Smoothing Spline mit k "effective degrees of freedom"

f) Schätzen Sie $\hat{f}$ auch als Smoothing Spline indem Sie den Smoothing Parameter λ über Kreuzvalidierung schätzen. Sind die resultierenden "effective degrees of freedom" kleiner, größer oder gleich k?

g) Fügen Sie die kubische Regression Spline und die natürliche kubische Regression Spline mit k freien Parametern sowie die Smoothing Spline
mit k "effective degrees of freedom" in das Streudiagramm des letzten Arbeitsblatts.

h) Beschreiben Sie Unterschiede und Gemeinsamkeiten der polynomialen, der Spline- und der Smoothing-Spline-Schätzer von f. Für welches Modell würden Sie sich abschließend entscheiden?

A 2 In dieser Aufgabe soll gezeigt werden, dass eine natürliche kubische Spline mit K Knotenpunkten K freie Parameter hat. Da jede kubische Spline über eine Basis der trunkierten Potenzfunktionen darstellbar ist, schreiben wir f dazu zunächst als

$$f(x) = \sum_{j=0}^{3} a_j x^j + \sum_{k=1}^{K} b_k (x - \xi_k)_+^3$$

Dabei werden die ersten vier Parameter mit a_0, a_1, a_2, a_3 bezeichnet, damit der Index von b in der zweiten Summe bei 1 beginnen kann. Das vereinfacht die Darstellung später und hat sonst keine Bedeutung.

a) Welche zusätzlichen Eigenschaften hat eine natürliche kubische Spline im Gegensatz zu einer allgemeinen kubischen Spline?

b) Zeigen Sie, dass diese Eigenschaften den vier Bedingungen an die Parameter

$$a_2 = 0, \quad a_3 = 0, \quad \sum_{k=1}^{K} b_k = 0, \quad \sum_{k=1}^{K} b_k \xi_k = 0$$

entspricht.

c) Begründen Sie damit, dass die Anzahl freier Parameter K ist.

d) Zeigen Sie mit diesen Ergebnissen weiter, dass jede natürliche kubische Spline über die folgende Basisfunktionserweiterung darstellbar ist

$$f(x) = a_0 + a_1 x + \sum_{k=1}^{K-2} \tilde{b}_k \cdot \left(\frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_k} - \frac{(x - \xi_{K-1})_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_{K-1}} \right)$$

($\tilde{b}_k$ statt b_k, da der Parameter nicht dem Wert b_k aus den vorangegangenen Aufgabenteilen entsprechen muss.)