

5 Semi- und Nonparametrische Regression (II)

Aufgabe 2

Im Folgenden sollen die bereits bekannten `fakesoep`-Daten (siehe Blatt 4) mittels generalisierter additiver Modelle (R-Package `mgcv`) analysiert werden.

- (a) Fitten Sie ein generalisiertes additives Modell mit gammaverteiltem Response (`beink`), log-Link und allen Kovariablen. Betrachten Sie Frauen und Männer getrennt, und modellieren Sie den Einfluss von Körpergröße, Alter und Dauer der Betriebszugehörigkeit als glatte Funktion. Verwenden Sie hierfür kubische Regressionssplines.
- (b) Fassen Sie Ihre beiden Modelle aus (a) nun in einem Modell mit gleicher Flexibilität zusammen.
- (c) Welche Verteilung bietet sich als Alternative zur Gamma-Verteilung an? Fitten Sie ein entsprechendes Modell und vergleichen Sie es mit Ihrem Gamma-Modell aus (b) an Hand von Diagnostik-Plots.
- (d) Vergleichen Sie Ihr(e) Modell(e) mit einem GLM (siehe Blatt 4). Sollten alle metrischen Prädiktoren nonparametrisch ins Modell aufgenommen werden?

Aufgabe 3

Man betrachte ein nonparametrisches Regressionsmodell folgender Form:

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2(x_i))$$

Einen *lokal linearen Schätzer* $\hat{f}(x)$ für $f(x)$ erhält man, indem man den Ausdruck

$$Q(\beta_0, \beta_1) := \sum_{i=1}^n [y_i - \beta_0 - \beta_1(x_i - x)]^2 K_h(x_i - x) \quad (1)$$

minimiert, wobei $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$, mit einer Kernfunktion K .

- (a) Überlegen Sie sich mittels einer Taylorentwicklung der Funktion f um den Punkt x , welche Interpretation die Parameter β_0 und β_1 in (1) haben. Was also ist $\hat{f}(x)$?
- (b) Lösen Sie obiges Minimum-Quadrat-Problem im speziellen Fall $\beta_1 \equiv 0$. Dies führt zum sogenannten *Nadaraya-Watson-Schätzer*. Zeigen Sie, dass dieser lokale Schätzer tatsächlich ein Minimum (und kein Maximum!) von $Q(\beta_0, 0)$ darstellt. Berechnen Sie Bias und Varianz des Nadaraya-Watson-Schätzers.
- (c) Wir betrachten nun den Fall $h \rightarrow 0$ und $nh \rightarrow \infty$. Dann lassen sich folgende asymptotische Ausdrücke für Bias und Varianz des lokal linearen Schätzers ($\beta_1 \neq 0$) an der Stelle x herleiten (bitte nicht versuchen...):

$$\begin{aligned} \text{Bias}(\hat{f}(x)) &\approx \frac{h^2 \mu_2}{2} f''(x), \\ \text{Var}(\hat{f}(x)) &\approx \frac{\nu_0 \sigma^2(x)}{nhg(x)}. \end{aligned}$$

mit $\mu_2 = \int u^2 K(u) du$, $\nu_0 = \int K^2(u) du$ und $\sigma^2(x) = \text{Var}(Y|X = x)$; $g(x)$ bezeichne die Datendichte an der Stelle x . Interpretieren Sie diese Formeln.