Tutorial 5

Nicht-parametrische Regression, Nadaraya-Watson

Aufgabe 1: Nadaraya-Watson „zu Fuß“ und mit `ksmooth()`

Auf diesem Übungszettel arbeiten wir mit Mietdaten aus München. Wir betrachten dabei nur Wohnungen, die höchstens 200 m² groß sind:

```
miete = read.table("https://tinyurl.com/yn9ttfcu", header = TRUE)
miete = miete[miete$wfl <= 200, ]
```

- a) Erstellen Sie ein Streudiagramm zwischen den Variablen Wohnfläche (`wfl`) und Nettomiete pro Quadratmeter (`nmqm`). Wir betrachten anschließend das nicht-parametrische Regressionsmodell:

$$\text{nmqm}_i = m(\text{wfl}_i) + \varepsilon_i, \quad i = 1, \dots, n.$$

Bestimmen Sie „zu Fuß“ — d.h. ohne Zuhilfenahme einer R-Funktion zur nicht-parametrischen Regression — den Wert des Nadaraya-Watson-Schätzers $\hat{m}(\text{wfl})$ an der Stelle `wfl = 150` unter Verwendung des Rechteckkerns und mit Bandweite $b = 20$. Ergänzen Sie anschließend den berechneten Schätzwert wie folgt im Streudiagramm:

```
points(150, ..., col = "orange", pch = 16)
```

- b) Definieren Sie nun mit den folgenden Befehlen eine Funktion in R, welche unserem Rechteckkern entspricht:

```
K = function(z) { ifelse(abs(z) > 0.5, 0, 1) }
```

Probieren Sie die Funktion einmal aus, indem Sie sie auf verschiedene Werte von `z` anwenden. Bestimmen Sie anschließend noch einmal den Wert des Nadaraya-Watson-Schätzers $\hat{m}(\text{wfl})$ an der Stelle `wfl = 150` (mit $b = 20$), diesmal unter Verwendung der Funktion `K()` (siehe Slide 13 der Vorlesung).

- c) Sie können den Nadaraya-Watson-Schätzer mit Gaußkern in komplett analoger Art und Weise bestimmen, indem Sie statt der von uns definierten Funktion `K()` (Rechteckkern) einfach `dnorm()` (Gaußkern) nutzen. Probieren Sie dies einmal aus, wieder an der Stelle `wfl = 150` und wieder mit Bandweite `b = 20`. Ergänzen Sie anschließend mit:

```
points(150, ..., col = "lightblue", pch = 16)
```

den berechneten Schätzwert in dem Streudiagramm. Fügen Sie mit:

```
NWS <- ksmooth(miete$wfl, miete$nmqm, kernel = "normal", bandwidth = 55)
lines(NWS$x, NWS$y, col = "lightblue", lwd = 2)
```

die komplette, mit Nadaraya-Watson geschätzte Regressionsfunktion $\hat{m}(x)$ dem Streudiagramm hinzu.

Hinweis: Die Funktion `ksmooth()` benutzt eine Umrechnung der Bandweiten, welche hier gerade so ist, dass die Wahl von `bandwidth = 55` etwa unserer Wahl `b = 20` entspricht.

Aufgabe 2: Bandweitenwahl und Kreuzvalidierung

- a) Schauen Sie sich mit `ksmooth()` den Nadaraya-Watson-Schätzer für das Modell aus Aufgabe 1 mit dem Gaußkern für verschiedene Bandweiten an und wählen Sie eine Ihres Erachtens geeignete Bandweite aus.
- b) Wir wollen nun eine Kreuzvalidierung umsetzen. Gehen Sie wie folgt vor:

1. Bestimmen Sie mit `ksmooth()` den Ausdruck

$$(\text{nmqm}_i - \hat{m}_{b,-i}(\text{wfl}_i))^2$$

aus dem Kreuzvalidierungskriterium für $i = 1$, d.h. wir betrachten zunächst nur die erste Beobachtung zur Validierung. Verwenden Sie den Gaußkern und eine Bandweite von 10. Zur Anpassung müssen Sie die erste Beobachtung unberücksichtigt lassen, z.B. durch:

```
ksmooth(miete$wfl[-1], miete$nmqm[-1], ...)
```

Durch Ergänzung der Option `x.points = miete$wfl[1]` wird dann die Vorhersage für eben jene erste Beobachtung bestimmt.

2. Bestimmen Sie nun $CV(10) = \sum_{i=1}^n (\text{nmqm}_i - \hat{m}_{10,-i}(\text{wfl}_i))^2$, d.h. den Wert des Kreuzvalidierungskriteriums für $b = 10$, nun aber für die Summe aller Beobachtungen — dazu eignet sich eine `for`-Schleife.
 3. Erweitern Sie ihre `for`-Schleife aus 2., sodass Sie $CV(b)$ für die Bandweiten 10, 10.1, 10.2, ..., 19.9, 20 ausrechnen. Plotten Sie dann $CV(b)$ als Funktion von b . Wählen Sie die gemäß des Kreuzvalidierungskriteriums beste Bandweite und vergleichen Sie den Wert mit Ihrer subjektiven Wahl aus a).
- c) Inwiefern äußert sich in dem Plot aus 3. der Bias-Varianz-Trade-off?