Kapitel 3

Nicht-parametrische Regression

3.1 Nadaraya-Watson

Fragen? https://www.menti.com/alneo5om9sq5.

Flexible Modellierung mit parametrischen Regressionsmodellen

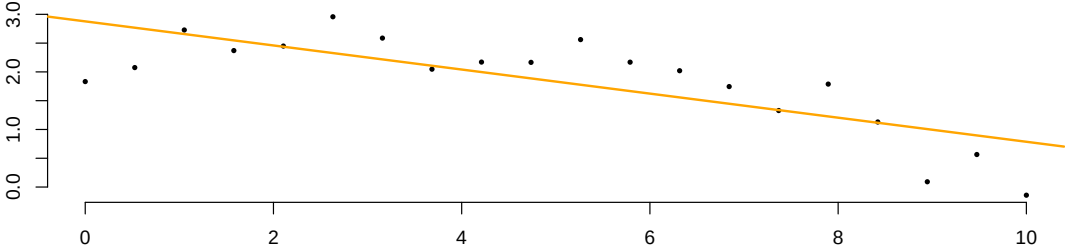
Parametrische Regressionsmodelle sind dergestalt, dass der Prädiktor durch **endlich viele Modellparameter** festgelegt ist, z.B.

- $\beta_0 + \beta_1 x_{i1}$
- $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$
- $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2$
- $\beta_0 + \beta_1 \sqrt{x_{i1}}$
- $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i1}^3 + \beta_4 x_{i1}^4$

Durch polynomiale Terme wie im letzten Beispiel kann man im Prinzip beliebig viel Flexibilität für die Form der Regressionsfunktion erhalten. Das schauen wir uns jetzt auf den folgenden Slides an.

Fragen? https://www.menti.com/alneo5om9sq5.

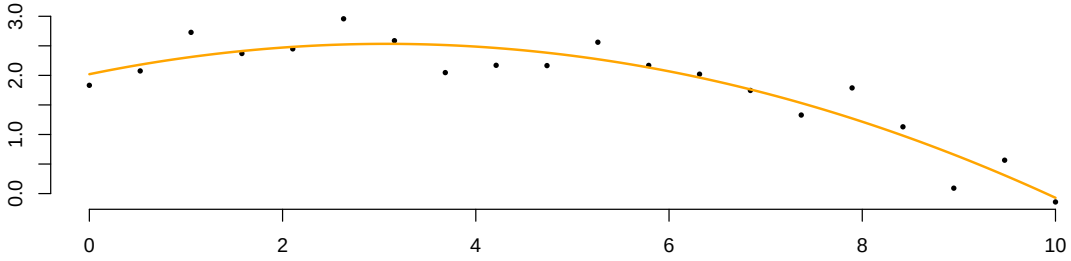
- wir simulieren Daten aus dem Modell $Y_i = 2 + 0.3x_{i1} - 0.05x_{i1}^2 + \epsilon_i, \quad \epsilon_i \sim N(0, 0.3^2)$
- wir passen ein lineares Modell **ohne** quadratischen Term an: $Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$



Klarer Fall von “Underfitting” (Unteranpassung): Modell ist zu unflexibel, um den quadratischen Effekt abzubilden.

Fragen? https://www.menti.com/alneo5om9sq5.

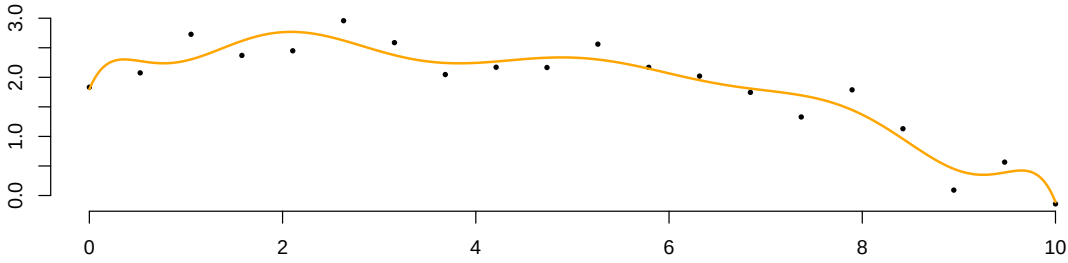
- wir betrachten dieselben simulierten Daten wie auf der vorherigen Folie, also wieder Daten aus dem Modell $Y_i = 2 + 0.3x_{i1} - 0.05x_{i1}^2 + \epsilon_i, \quad \epsilon_i \sim N(0, 0.3^2)$
- wir passen Regressionsmodell mit **quadratischem Prädiktor** an: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \epsilon_i$



Modell passt gut zu Daten: quadratischer Term im Prädiktor trägt dem vorgefundenen Muster Rechnung

Fragen? https://www.menti.com/alneo5om9sq5.

- wir betrachten dieselben simulierten Daten wie auf den vorherigen Folien, also wieder Daten aus dem Modell $Y_i = 2 + 0.3x_{i1} - 0.05x_{i1}^2 + \epsilon_i, \quad \epsilon_i \sim N(0, 0.3^2)$
- wir passen Regressionsmodell mit **Prädiktor mit polynomialen Termen bis zur Ordnung 10** an: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i1}^3 + \ldots + \beta_{10} x_{i1}^{10} + \epsilon_i$

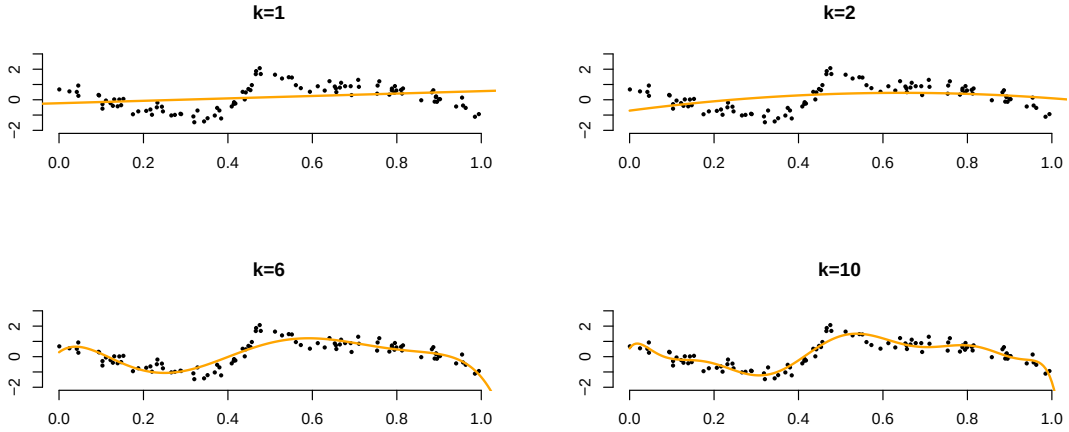


Klarer Fall von “Overfitting” (Überanpassung): Modell beschränkt sich nicht auf das wesentliche Muster, stattdessen werden vernachlässigbare Eigenschaften der Daten, die in Wirklichkeit nur zufälliges Rauschen sind, vom Modell abgebildet

Fragen? https://www.menti.com/alneo5om9sq5.

Bias-Varianz-Trade-off bei Nutzung polynomialer Prädiktoren

Wir simulieren Daten (x_{i1}, y_i), $i = 1, \ldots, 100$ und vergleichen verschiedene angepasste lineare Regressionsmodelle mit polynomialem Prädiktor $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \ldots + \beta_k x_{i1}^k$



Fragen? https://www.menti.com/alneo5om9sq5.

Wie sinnvoll sind Polynome höherer Ordnungen?

Durch die Idee der Variablentransformation — insbesondere durch polynomiale Prädiktoren — erreichen wir mit linearen Modellen im Prinzip beliebig viel Flexibilität!

Aber:

- die Schätzung von Polynomen höherer Ordnungen ist äußerst instabil
- insbesondere haben Ausreißer einen sehr starken Einfluss auf die Schätzung
- insgesamt besteht ein hohes Risiko des Overfitting[1]
- für jede Variable den optimalen Polynomgrad zu wählen ist umständlich

In der Praxis ist es daher unüblich, Polynome der Ordnung > 2 zu betrachten.

[1] solche zu flexiblen Modelle liegen nahe an den Daten, liefern aber schlechte Vorhersagen

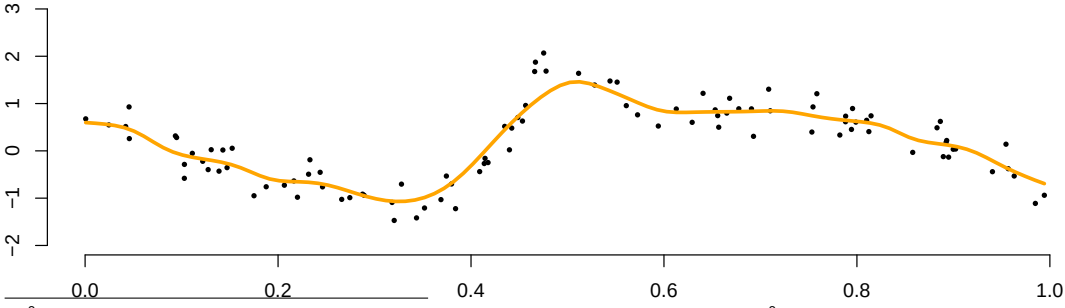
Fragen? https://www.menti.com/alneo5om9sq5.

Idee der nichtparametrische Regression

Idee der **nichtparametrischen Regression**: schätze das Modell

$$Y_i = m(x_{i1}) + \epsilon_i, \quad i = 1, \ldots, n$$

wobei m eine nicht näher spezifizierte glatte Funktion ist $\rightsquigarrow$ keine bestimmte parametrische Form, daher keine restriktiven Annahmen![2]



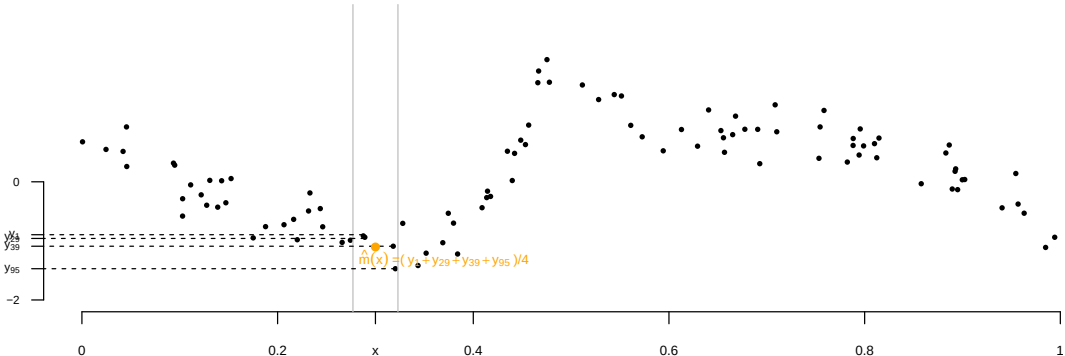
[2] wir nehmen allerdings weiterhin an, dass die ϵ_i unabhängig sind mit $E(\epsilon_i) = 0$ und $\text{Var}(\epsilon_i) = \sigma^2$

Fragen? https://www.menti.com/alneo5om9sq5.

Kernregression

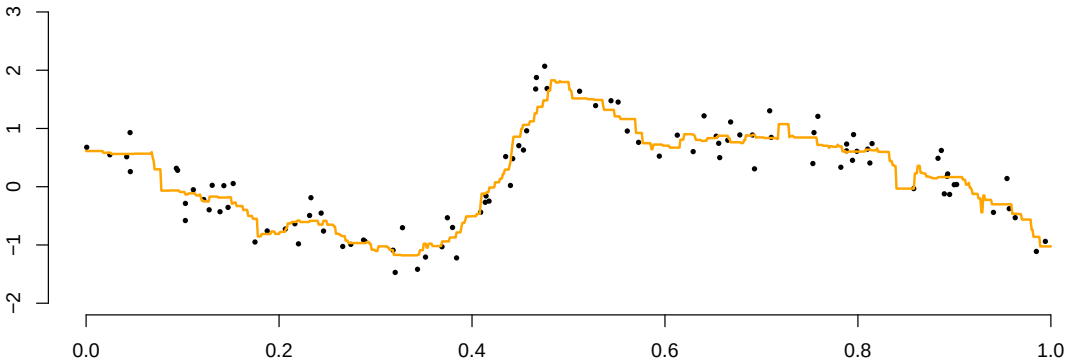
Es gibt eine Vielzahl an Methoden, wie die glatte Funktion m geschätzt werden kann. Wir betrachten zunächst die sogenannte **Kernregression**.

Grundidee: für gegebenes x, bestimme $\hat{m}(x)$ als (lokalen) Mittelwert aller y_i deren zugehörige x_{i1} nahe bei x liegen — tue dies für alle x.



Fragen? https://www.menti.com/alneo5om9sq5.

So wandert man Schritt für Schritt über die x-Achse und verbindet schließlich alle lokalen Mittelwerte:



Offensichtliches Problem: der Schätzer ist **nicht stetig** — unschön!

Fragen? https://www.menti.com/alneo5om9sq5.

Formalisierung der Idee

Die lokalen Mittelwerte lassen sich wie folgt formalisieren:

$$\hat{m}(x) = \frac{1}{\sum_{i=1}^{n} \mathcal{I}\left(x_i \in [x - \frac{b}{2}, x + \frac{b}{2}]\right)} \sum_{i=1}^{n} \mathcal{I}\left(x_i \in [x - \frac{b}{2}, x + \frac{b}{2}]\right) \cdot y_i$$

Alternative Schreibweise:

$$\hat{m}(x) = \frac{1}{\sum_{i=1}^{n} K(\frac{x - x_i}{b})} \sum_{i=1}^{n} K\left(\frac{x - x_i}{b}\right) \cdot y_i$$

mit $K(y) = \begin{cases} 1 & \text{falls } -1/2 \leq y \leq 1/2; \\ 0 & \text{sonst.} \end{cases}$

Fragen? https://www.menti.com/alneo5om9sq5.

Zur Erinnerung: allgemeine Kernfunktionen:

Eine Funktion $K(y)$ heißt Kernfunktion, falls

(i) $K(y) \geq 0$ für alle y

(ii) $K(y) = K(-y)$ für alle y

(iii) $\int_{-\infty}^{\infty} K(y)dy = 1$

Beispiele für Kernfunktionen:

- Rechteckkern
- Dreieckkern
- Gaußkern
- Epanechnikovkern

Fragen? https://www.menti.com/alneo5om9sq5.

Nadaraya-Watson-Schätzer

Für beliebige Kernfunktion $K(y)$ und feste Bandweite b heißt

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{b}) y_i}{\sum_{i=1}^{n} K(\frac{x-x_i}{b})}$$

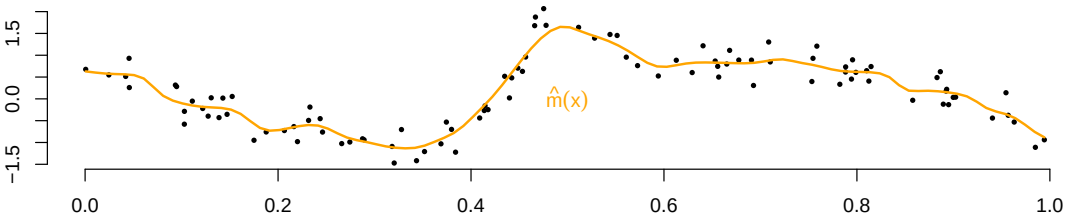
Nadaraya-Watson-Schätzer (NWS) (oder auch: Kernregressionsschätzer).

In R kann dafür die Funktion `ksmooth()` verwendet werden.

Fragen? https://www.menti.com/alneo5om9sq5.

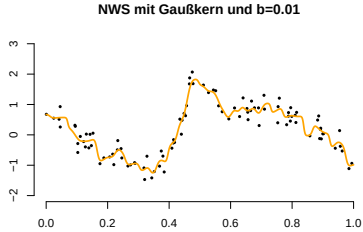
Nadaraya-Watson-Schätzer in R mit Gaußkern und $b = 0.05$:

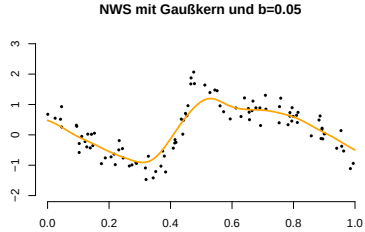
```
plot(x, y, pch = 16, bty = "n", main = "", xlab = "", ylab = "", cex = 0.6)
NW <- ksmooth(x, y, kernel = "normal", bandwidth = 0.05)
lines(NW, lwd = 2, col = "orange")
text(0.5, 0, expression(hat(m)(x)), col = "orange")
```

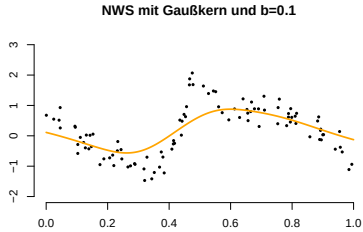


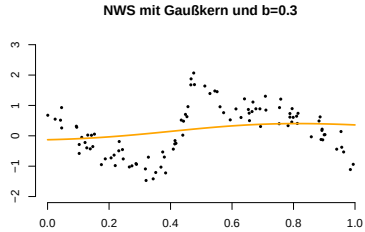
Die geschätzte Funktion $\hat{m}(x)$ ist glatt, da der Gaußkern glatt ist.

Einfluss der Bandweite









Fragen? https://www.menti.com/alneo5om9sq5.

Was die Bandweite angeht, so stellen wir fest:

- je höher b, desto höher der Bias
  (vor allem im Bereich von “Tälern” und “Gipfeln”)
- je kleiner b, desto höher die Varianz
  (da die lokalen Mittelwerte dann von nur wenigen Punkten abhängen)
- es ergibt sich also wieder ein **Bias-Varianz-Trade-off**

(Der Einfluss der Kernfunktion ist eher unbedeutend.)

Konkret kann man zeigen:

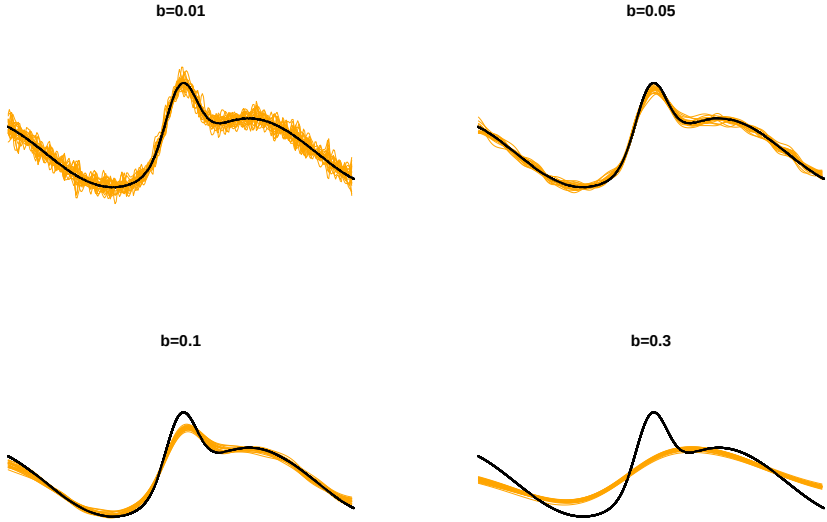
$$\text{Bias}\big(\hat{m}(x)\big) = E(\hat{m}(x)) - m(x) \approx b^2 \cdot m''(x) \cdot \text{konst.}$$

$$\text{Var}\big(\hat{m}(x)\big) \approx \frac{1}{nb} \cdot \text{konst.}$$

Der Einfluss von b ist im Wesentlichen also analog wie beim Kerndichteschätzer.

Fragen? https://www.menti.com/alneo5om9sq5.

- wir betrachten ein wahres Modell (schwarze Linie), aus dem wir 20 Mal Daten simulieren
- anschließend werden Regressionsfunktionen an jede der 20 Stichproben angepasst



Fragen? https://www.menti.com/alneo5om9sq5.

Wahl der Bandweite durch Kreuzvalidierung

Inuitiv einleuchtend: eine Bandweite ist dann geeignet, wenn der dazugehörige Schätzer zuverlässige Vorhersagen zukünftiger Werte liefert. Um abzuschätzen, wie gut der Schätzer zukünftige Werte vorhersagen kann, betrachten wir das sogenannte **Kreuzvalidierungskriterium**:

$$\mathrm{CV}(b) = \sum_{i=1}^{n} \left(y_i - \hat{m}_{b,-i}(x_i) \right)^2,$$

wobei $\hat{m}_{b,-i}$ der NWS für Bandweite b angepasst an alle Daten **außer** (x_i, y_i) ist.

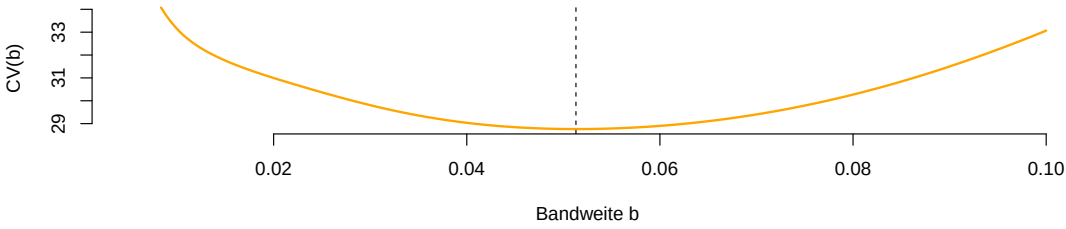
Was ist hier die Idee?

- es wird so getan, als sei die i-te Beobachtung ein zukünftiger Wert
- der NWS wird dementsprechend an die anderen $n - 1$ Beobachtungen angepasst
- dann wird geschaut, wie gut dieser NWS die $i - te$ Beobachtung vorhersagt
- dies wird für jede Beobachtung gemacht

Dann wird das b ausgewählt, welches $\mathrm{CV}(b)$ minimiert — sinnvoll, aber sehr rechenaufwändig!

Fragen? https://www.menti.com/alneo5om9sq5.

Kreuzvalidierung im vorherigen Beispiel mit simulierten Daten: optimal ist hier $b = 0.051$

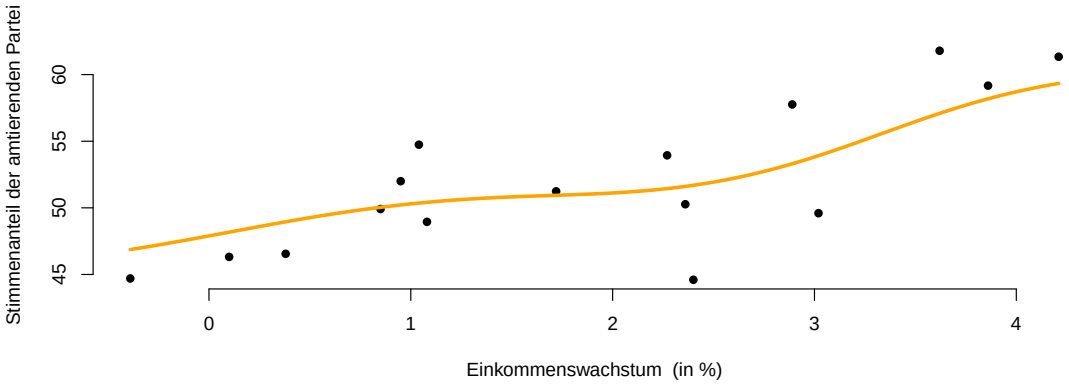


Kreuzvalidierung ist konzeptuell ansprechend, aber in der Praxis ist “**Wahl durch Augenmaß**” meist völlig ausreichend — und spart Rechenzeit!

Fragen? https://www.menti.com/alneo5om9sq5.

Nadaraya-Watson-Schätzer von m im Modell

$$\text{Stimmenanteil}_i = m(\text{Einkommenswachstum}_i) + \epsilon_i$$

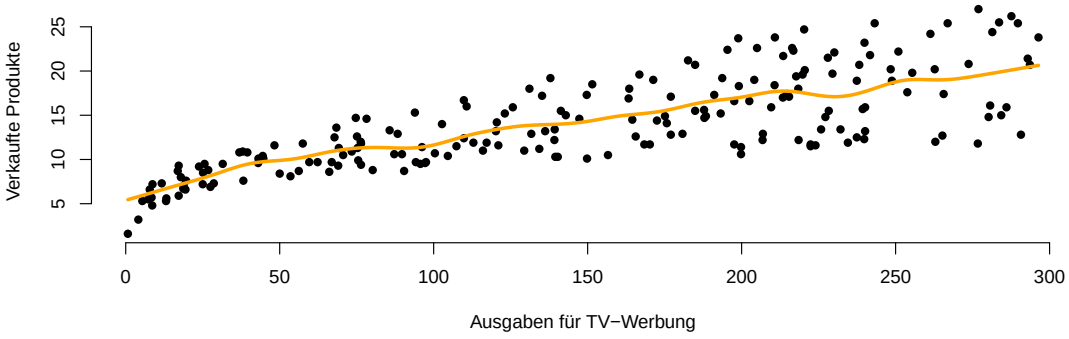


Hier scheint eine lineare Regressionsfunktion ausreichend zu sein.

Fragen? https://www.menti.com/alneo5om9sq5.

Nadaraya-Watson-Schätzer von m im Modell

$$\text{Produktverkauf}_i = m(\text{Werbung}_i) + \epsilon_i$$



$\rightsquigarrow$ geringerer Effekt ab 200.000 Euro Ausgaben für Werbung?

Fragen? https://www.menti.com/alneo5om9sq5.

Nadaraya-Watson-Schätzer — einige Feststellungen

- der NWS wird oft als **exploratives Werkzeug** eingesetzt, um einen zu schätzenden funktionalen Zshg. zwischen x und Y zu visualisieren
- oft zeigt sich, dass ein linearer oder quadratischer Prädiktor ausreichend ist; in einem solchen Fall ist das parametrische Modell i.d.R. vorzuziehen, da
  - stabiler
  - einfacher zu interpretieren
  - einfacher handzuhaben
- wenn der NWS **nicht** auf einen linearen oder quadratischen Zusammenhang hindeutet, so ist der nichtparametrische Schätzer i.d.R. vorzuziehen

Fragen? https://www.menti.com/alneo5om9sq5.

Nadaraya-Watson im Falle multipler erkl. Variablen

Es gibt verschiedene Wege, Nadaraya-Watson auf den Fall multipler erklärender Variablen zu erweitern. Einfachste Möglichkeit, hier illustriert für den Fall $p = 2$:

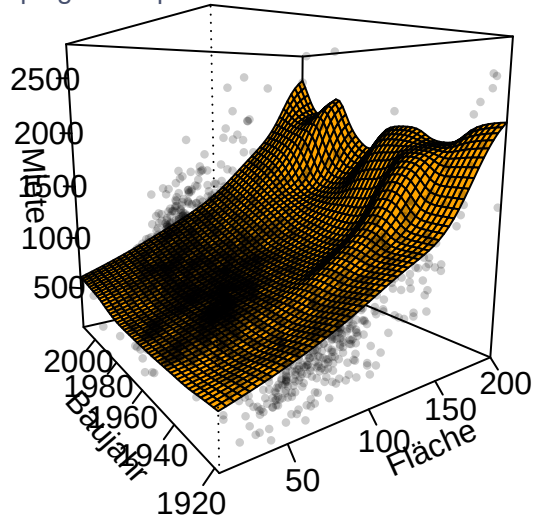
$$\hat{m}(x_1, x_2) = \frac{\sum_{i=1}^{n} K(\frac{x_1 - x_{i1}}{b_1}) K(\frac{x_2 - x_{i2}}{b_2}) y_i}{K(\frac{x_1 - x_{i1}}{b_1}) K(\frac{x_2 - x_{i2}}{b_2})}$$

Man spricht dann von einem **multiplikativen Kern**.

Was hier passiert:

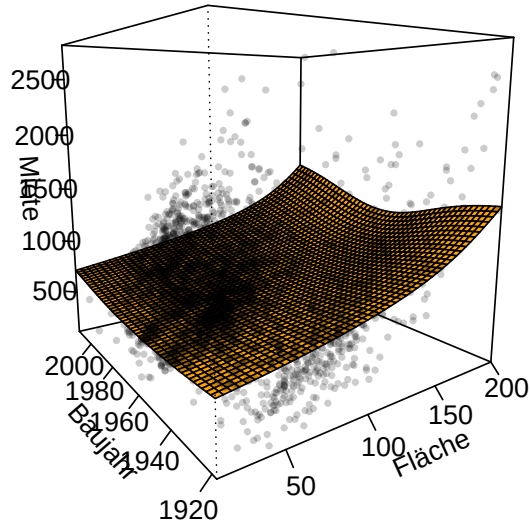
- letztendlich ist $\hat{m}(x_1, x_2)$ wieder ein lokaler Mittelwert
- in beiden Dimensionen wird geprüft, welche Daten nahe bei (x_1, x_2) liegen

Multivariater NWS im Mietspiegel-Beispiel



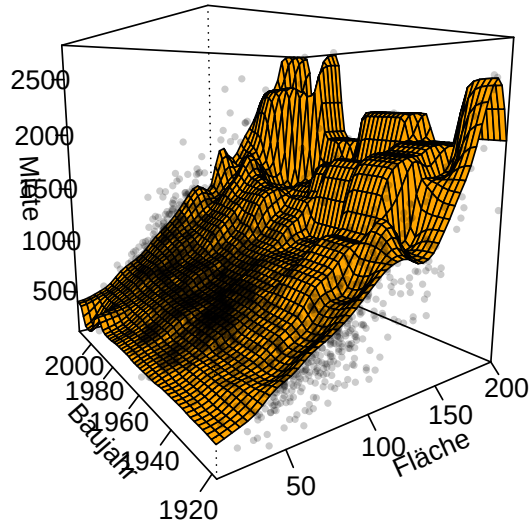
- große und eher neue (ab 1990 gebaute) aber auch alte (vor 1940) Wohnungen sind auffällig teuer
- in der Nachkriegszeit gebaute Wohnungen sind eher günstiger

Fragen? https://www.menti.com/alneo5om9sq5.



falls b_1 und b_2 zu groß $\rightsquigarrow$ Underfitting

Fragen? https://www.menti.com/alneo5om9sq5.



falls b_1 und b_2 zu klein $\rightsquigarrow$ Overfitting

Fragen? https://www.menti.com/alneo5om9sq5.

“Curse of dimensionality”

Kann man ein Modell der Form

$$Y_i = m(x_{i1}, \ldots, x_{ip}) + \epsilon_i, \quad i = 1, \ldots, n$$

auch mit $p > 2$ sinnvoll nichtparametrisch schätzen?

Antwort: ja, aber der Schätzer ist dann sehr “datenhungrig”. Illustration: (in der Vorlesung)

Fragen? https://www.menti.com/alneo5om9sq5.

Für $n = 1000$, $b = 0.1$ und $x_{i1}, \ldots, x_{ip} \sim U[0, 1]$ wird jeder lokale Mittelwert...

- ... für $p = 1$ aus etwa 100 beobachtungen gebildet
- ... für $p = 2$ aus etwa 10 Beobachtungen gebildet
- ... für $p = 3$ aus etwa einer Beobachtung gebildet

Um den 3D-Raum ähnlich gut abzudecken wie den 1D-Raum braucht man also 100 Mal so viele Daten (und nicht etwa nur 3 Mal so viele).

Kurzum: die nötige Stichprobengröße, um m sinnvoll zu schätzen steigt rasant mit steigendem p. Man spricht vom “curse of dimensionality”.

Daher wird der NWS wie hier besprochen quasi nur bei $p \leq 2$ angewendet — wir werden aber noch einen Ansatz sehen, welcher mit größerem p zurechtkommt.

Fragen? https://www.menti.com/alneo5om9sq5.