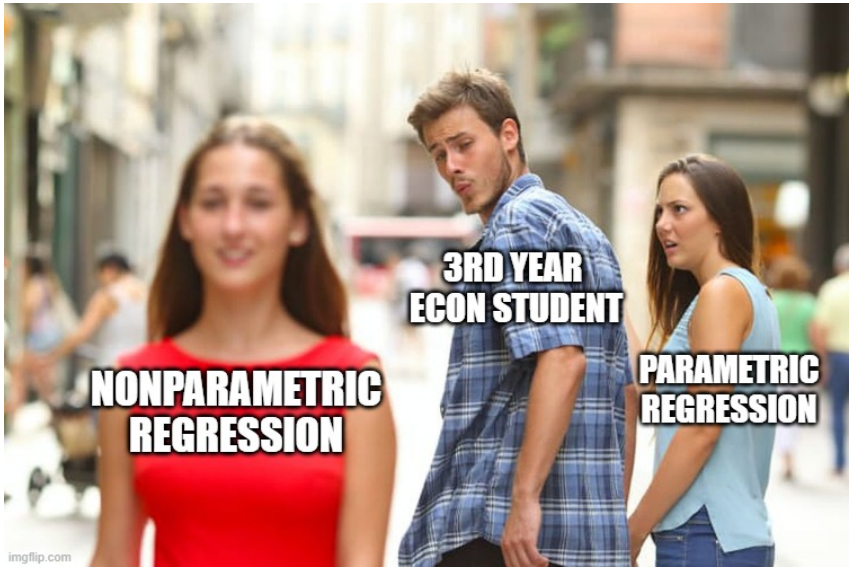
Woche 4 — Nichtparametrische Regression



Flexible Modellierung mit parametrischen Regressionsmodellen

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

- $\beta_0 + \beta_1 x_1$
- $\beta_0 + \beta_1 x_1 + \beta_2 x_2$
- $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$
- $\beta_0 + \beta_1 \sqrt{x_1}$
- $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$
- $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 x_1^4$

Durch polynomiale Terme wie im letzten Beispiel kann man im Prinzip beliebig viel Flexibilität für die Form der Regressionsfunktion erhalten. (s. nächste Slides)

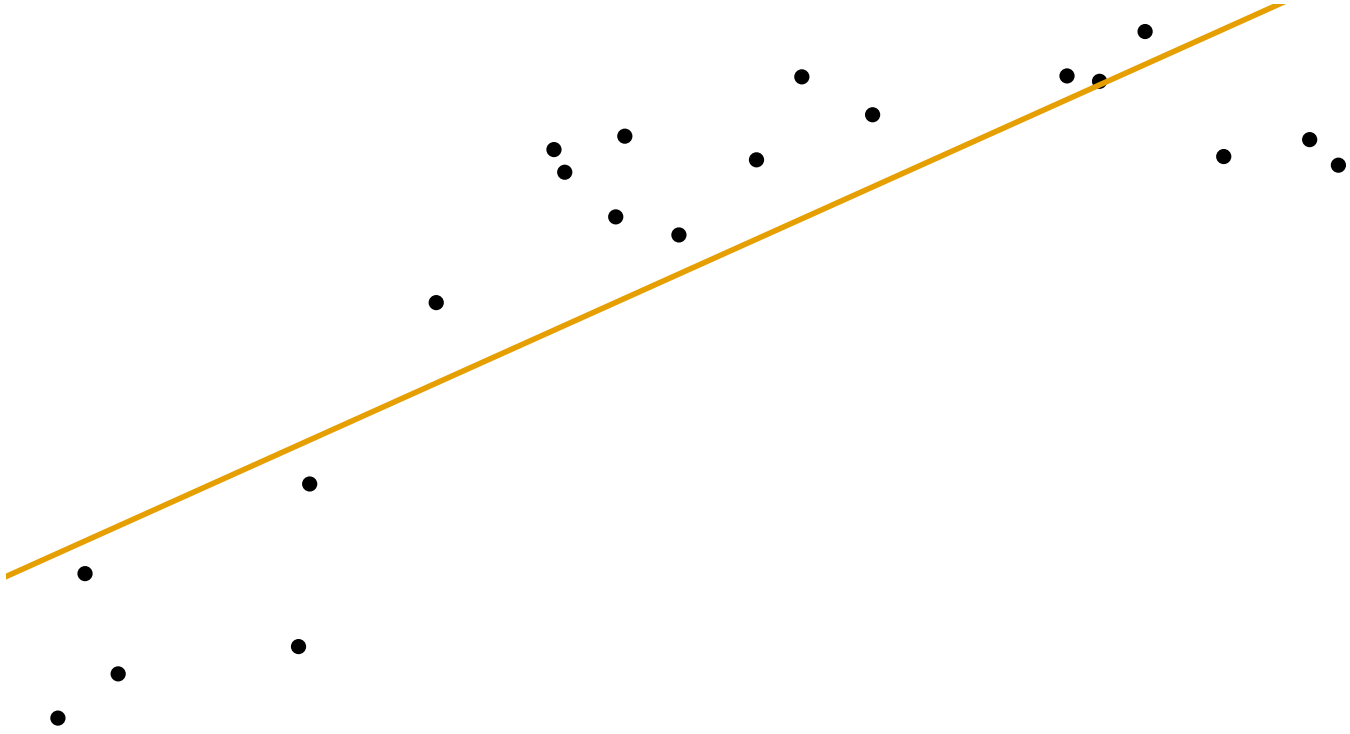


Abbildung: Simulierte Daten und angepasstes Regressionsmodell mit **linearem Prädiktor**.

Klarer Fall von “Underfitting” (Unteranpassung):

- Modell ist zu unflexibel, um den quadratischen Effekt abzubilden

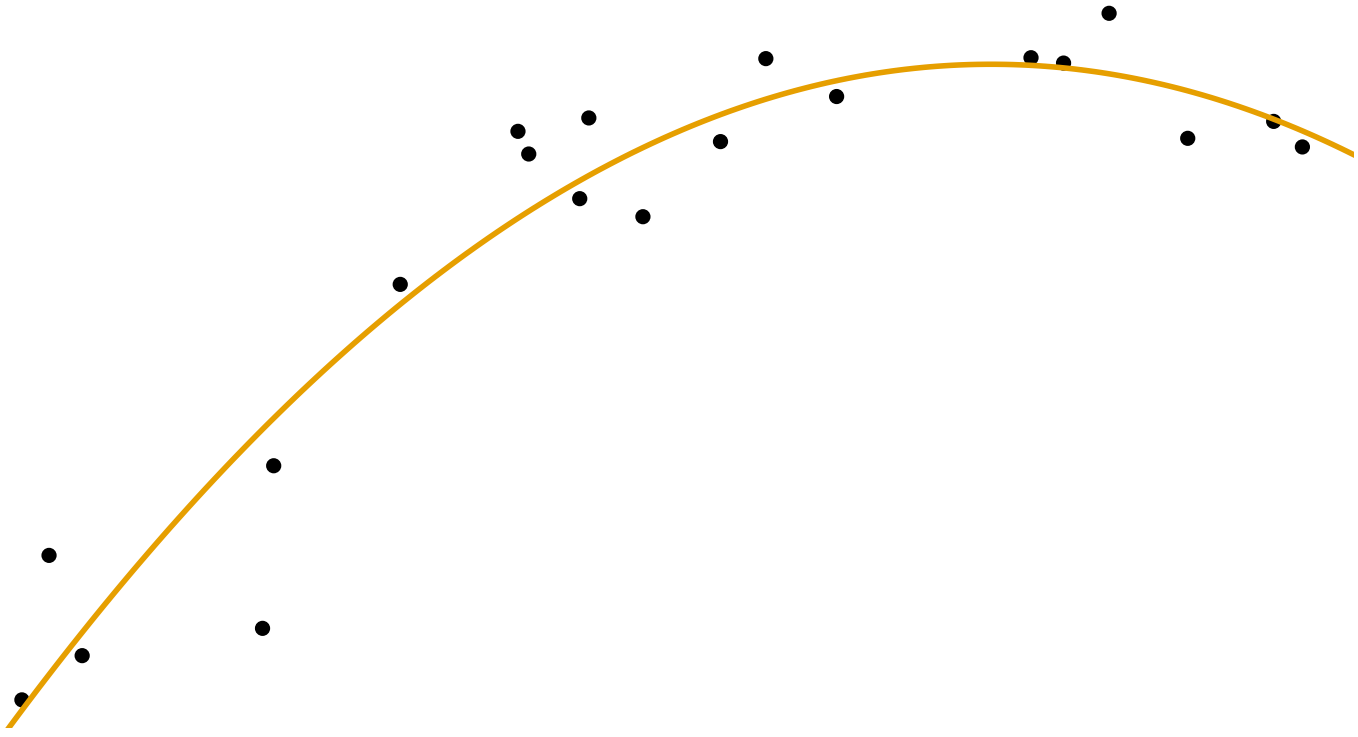


Abbildung: Simulierte Daten und angepasstes Modell mit **quadratischem Prädiktor**.

Modell passt gut zu Daten:

- quadratischer Term im Prädiktor trägt dem vorgefundenen Muster Rechnung

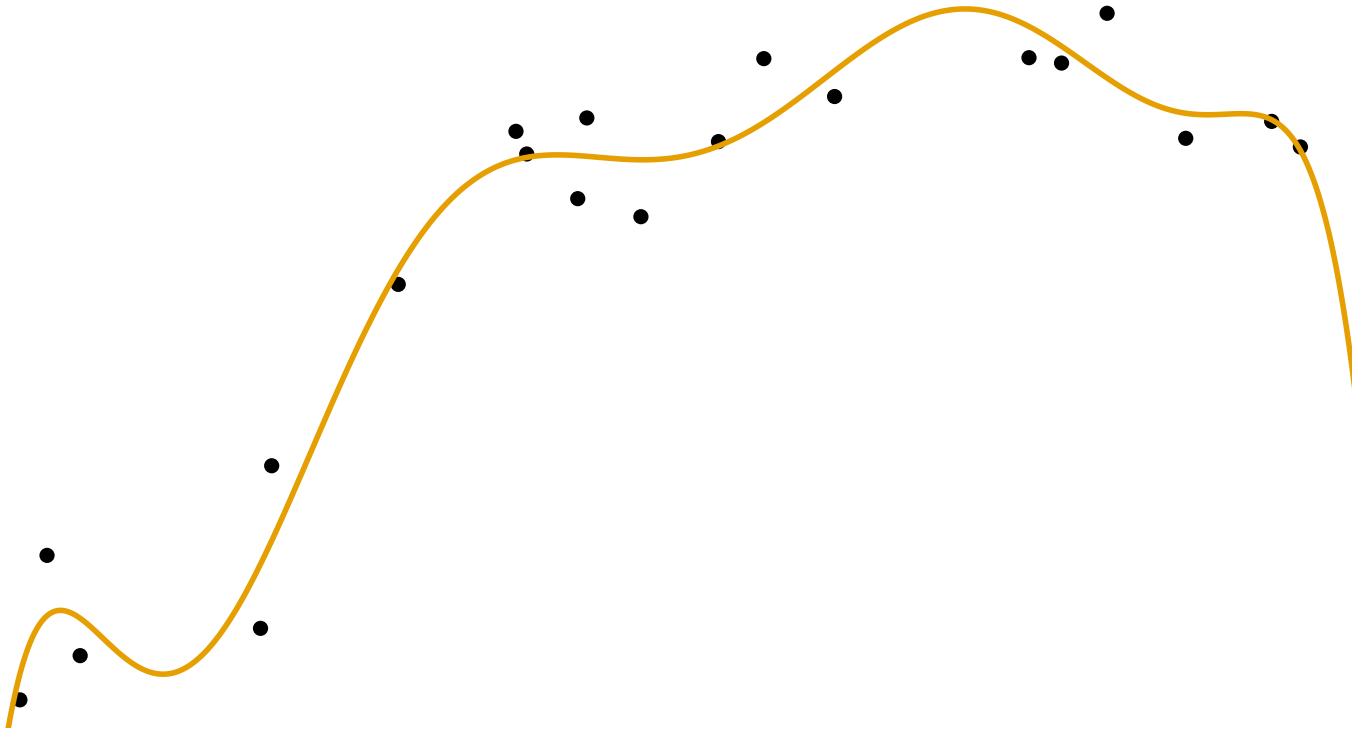


Abbildung: Simulierte Daten und Modell mit **polynomialem Prädiktor der Ordnung 10**.

Klarer Fall von “Overfitting” (Überanpassung):

- Modell bildet wesentliches Muster, aber leider auch zufälliges Rauschen ab

Bias-Varianz-Trade-off

Man kann Polynome von beliebiger Ordnung k anpassen:

$$\beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k$$

Dadurch erhält man Kontrolle darüber, wie präzise das Modell die Daten erklärt.

Er ergibt sich allerdings der sogenannte **Bias-Varianz-Trade-off** — im Beispiel:

- Modell zu einfach, $k = 1 \rightsquigarrow$ system. Fehlschätzung durch Inflexibilität[1]
- Modell zu komplex, $k = 10 \rightsquigarrow$ Überanpassung, zu genaues Abbild der Daten[2]
- das beste Modell liegt in der Regel irgendwo "in der Mitte" ($\rightsquigarrow$ Modellwahl)

[1] d.h. hoher Bias, man kann das Muster einfach nicht erfassen und liegt daher systematisch daneben
[2] d.h. hohe Varianz, das Erscheinungsbild wird von Stichprobe zu Stichprobe stark variieren

Wie sinnvoll sind Polynome höherer Ordnungen generell?

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

Aber:

- die Schätzung von Polynomen höherer Ordnungen ist äußerst instabil
- insb. haben Ausreißer einen sehr starken Einfluss auf die Schätzung
- insgesamt besteht ein **hohes Risiko des Overfitting**[3]
- für jede Variable den optimalen Polynomgrad zu wählen ist umständlich

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

[3] solche *zu* flexiblen Modelle liegen nahe an den Daten, liefern aber insb. schlechte Vorhersagen

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

$$Y_i = s(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

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

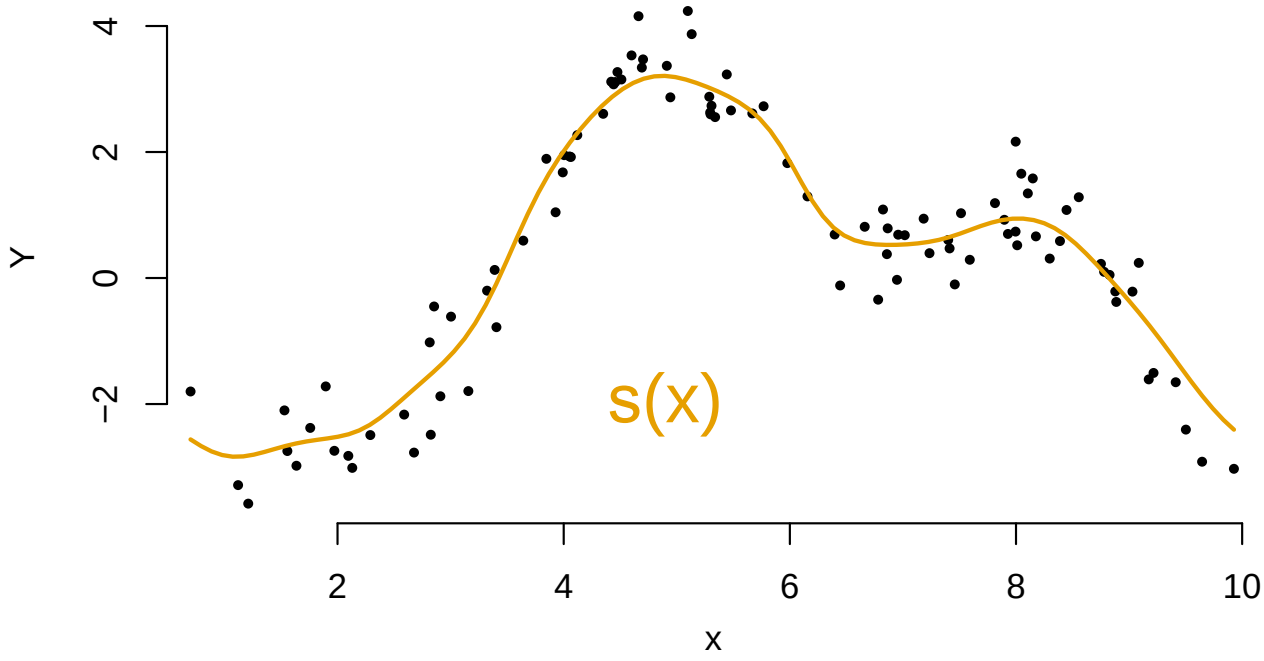
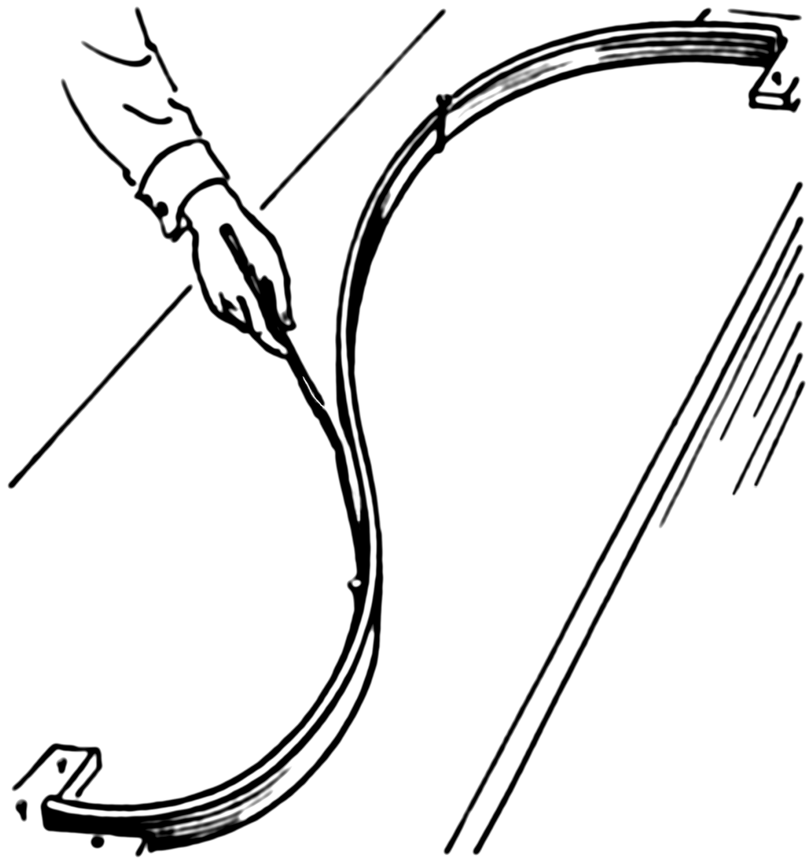


Abbildung: Idealisierte Darstellung von $s(x)$.

Smoothing Splines

Wie zeichnet man eine glatte Kurve?



Smoothing Splines (im Handwerk)

Die vorige Slide zeigt eine sogenannte Straklatte (“smoothing spline”). Diese wurden früher[4] genutzt, um harmonische Linien zu zeichnen, z.B. im Schiffsbau.

Kriterien an die Latte:

1. soll einigermaßen flexible geschwungene Form zulassen
2. soll dabei nicht zu wackelig sein

Für 1. braucht man biegsame, für 2. eher starre Latte. Die gewünschte Form entscheidet über die ideale Biegsamkeit der Latte.

[4] d.h. vor dem Computerzeitalter

Smoothing Splines (Übersetzung in die Statistik)

Übertragung der Idee auf die Regressionsanalyse: lege eine Art Smoothing Spline “mitten durch die Punktewolke”, so dass

1. die Daten gut erklärt werden (d.h. Funktion liegt nahe an den Punkten)
2. die geschätzte Funktion dabei möglichst glatt ist (d.h. nicht zu unruhig)

Ein uns bekanntes Maß für 1. ist die Fehlerquadratsumme,

$$\sum_{i=1}^{n} \big(y_i - s(x_i)\big)^2.$$

Zu 2.: eine Funktion s ist unruhig, wenn sie starke Krümmungen vorweist — und das Ausmaß an Krümmung können wir durch die zweite Ableitung messen.

Die Funktion s, welche das Zielkriterium

$$\sum_{i=1}^{n} \left(y_i - s(x_i)\right)^2 + \lambda \int_a^b s''(x)^2 dx$$

minimiert[5], nennen wir einen **Smoothing Spline** (auf $[a, b]$).

- Fehlerquadratsumme: misst, wie nahe die Funktion an den Daten liegt
- Gesamtkrümmung[6]: ein Maß dafür, wie unruhig die Funktion ist

Zusätzlich ist λ ein (festzulegender) **Glättungsparameter** (s. nächste Slide).

[5] wie diese Funktion gefunden wird besprechen wir hier nicht, da zu technisch

[6] Integral über die (quadrierte) Krümmung in jedem Punkt

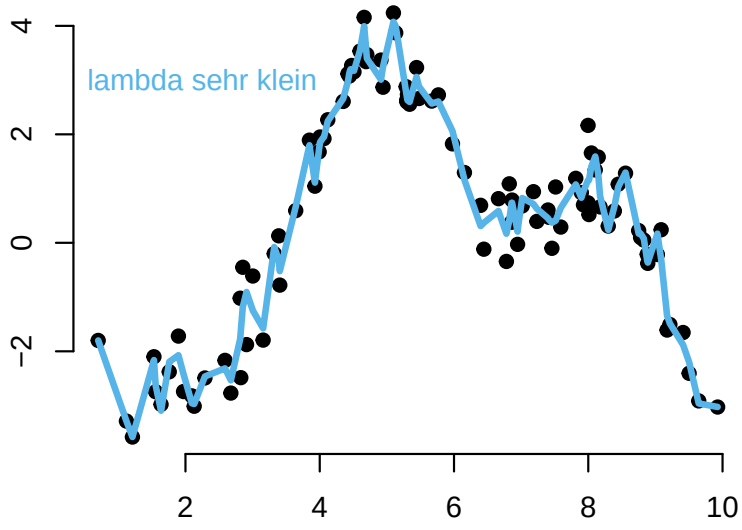
$$\sum_{i=1}^{n} \left(y_i - s(x_i)\right)^2 + \lambda \int_a^b s''(x)^2 dx$$

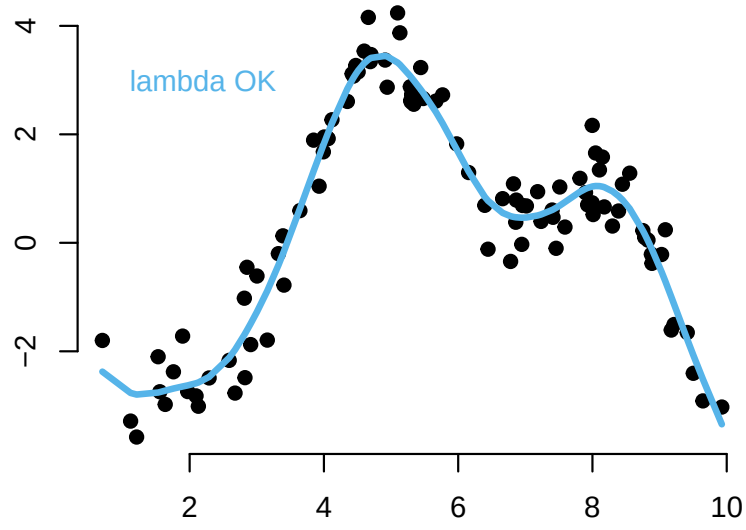
Glättungsparameter λ bestimmt, wie wichtig uns “Biegsamkeit” der Funktion ist:

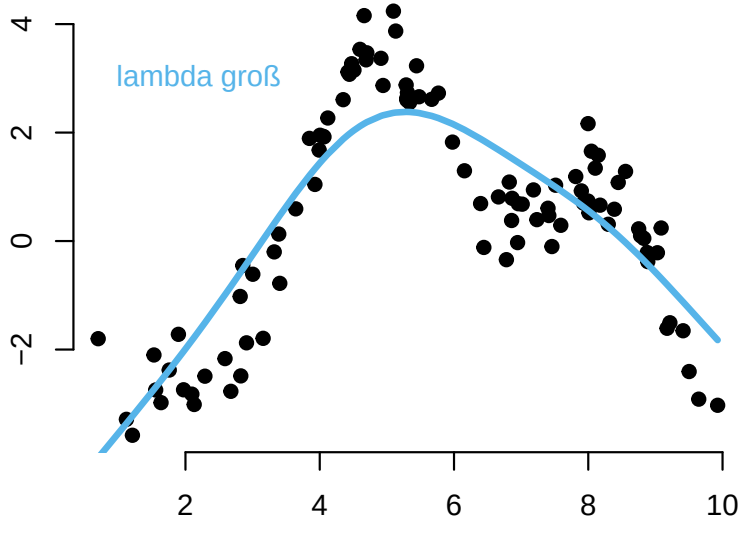
$$\begin{aligned} \lambda \to 0 \quad &\Rightarrow \quad s \text{ wird zur Interpolation der Datenpunkte} \\ \lambda \to \infty \quad &\Rightarrow \quad s \text{ wird zur Geraden} \end{aligned}$$

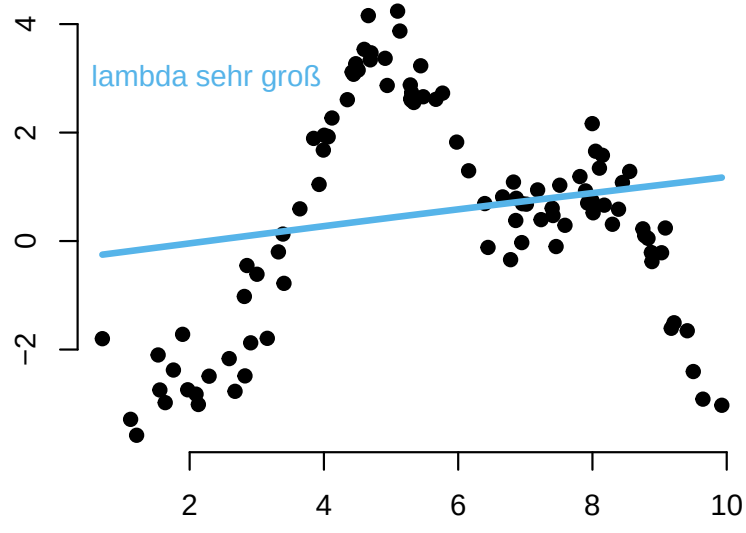
Insbesondere haben wir es wieder mit dem **Bias-Varianz-Trade-off** zu tun:

- niedriges λ führt zu niedrigem Bias aber hoher Varianz
- hohes λ führt zu niedriger Varianz aber hohem Bias

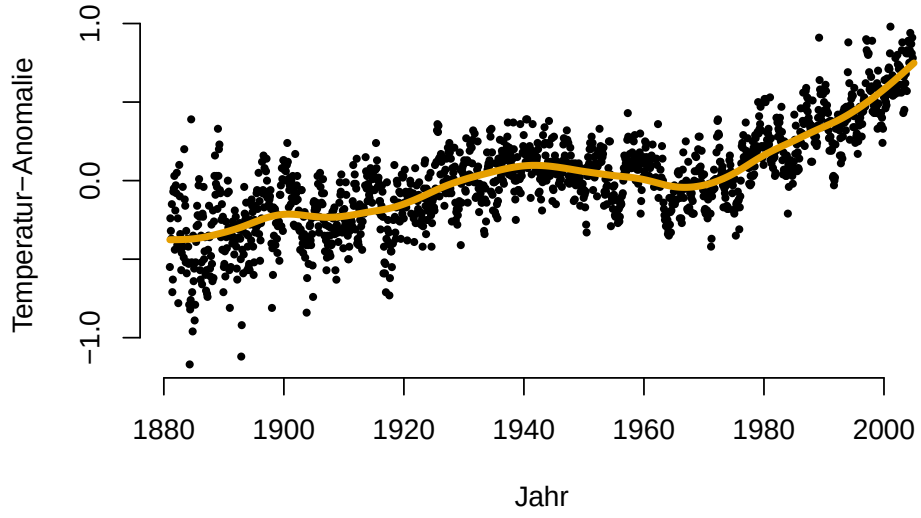

lambda sehr klein
4
2
0
−2
2
4
6
8
10


lambda OK
4
2
0
−2
2
4
6
8
10


lambda groß
4
2
0
−2
2
4
6
8
10


lambda sehr groß
4
2
0
−2
2
4
6
8
10

Smoothing Splines in R




```
# Daten einlesen und Streudiagramm erstellen:
klima <- read.csv("http://www.rolandlangrock.com/Daten/klima.csv",header=TRUE)
attach(klima)
plot(Jahr, Anomalie, pch=19, bty="n", ylab="Temperatur-Anomalie")

# Paket "npreg" installieren und laden:
install.packages("npreg")
library(npreg)

# Smoothing Spline anpassen und einzeichnen:
mod <- ss(Jahr, Anomalie, lambda=10^(-6))
lines(mod, col="#E69F00",lwd=4)
```

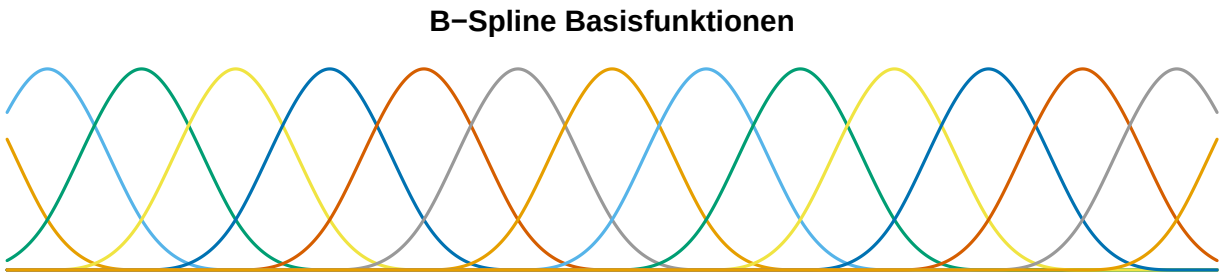
P–Splines

Wir betrachten nun mit **P–Splines** noch einen weiteren Ansatz, welcher eng verwandt mit den eben betrachteten Smoothing Splines ist.

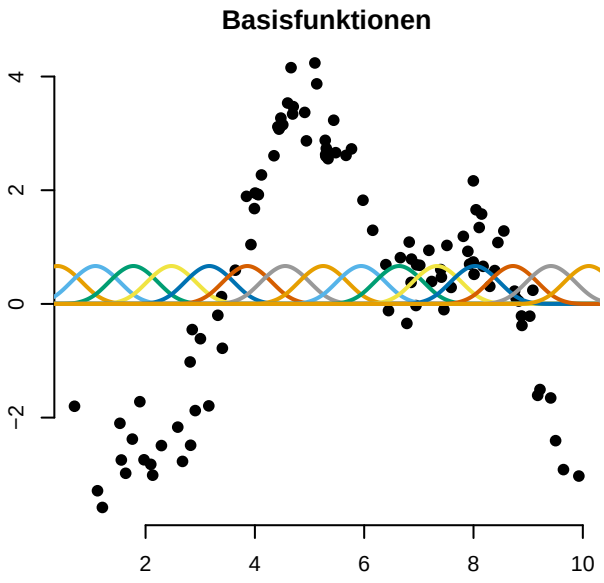
Grundidee: konstruiere $s(x)$ als gewichtete Summe **fester Basisfunktionen**:

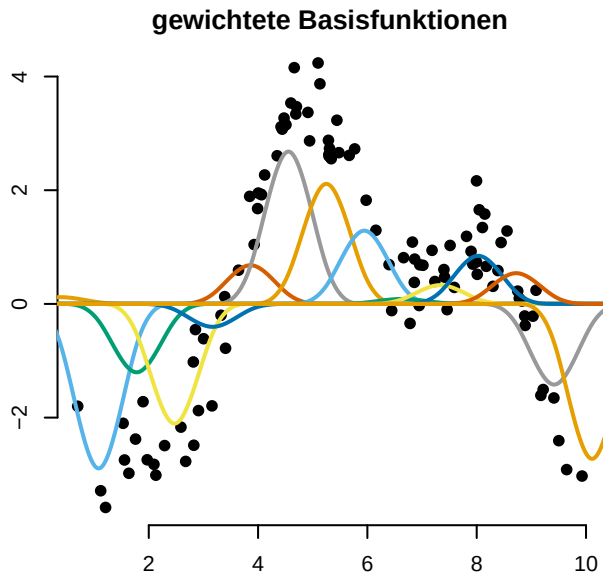
$$\begin{aligned} Y &= s(x) + \epsilon \\ &= \gamma_1 B_1(x) + \gamma_2 B_2(x) + \ldots + \gamma_K B_K(x) + \epsilon. \end{aligned}$$

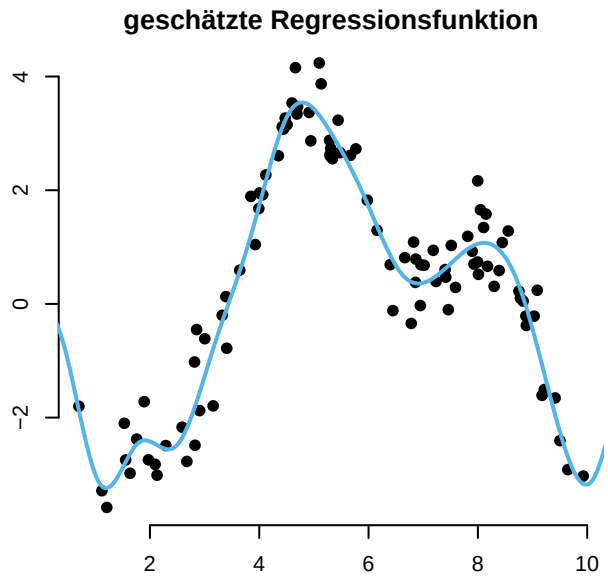
Schätzung der γ_j[7] ergibt einen Schätzer für s. Als Basisfunktionen wählen wir hier sogenannte B–Splines, da diese wünschenswerte Eigenschaften haben.



[7] durch Minimierung der Fehlerquadratsumme


Basisfunktionen
4
2
0
−2
2
4
6
8
10


gewichtete Basisfunktionen
4
2
0
−2
2
4
6
8
10


geschätzte Regressionsfunktion
4
2
0
−2
2
4
6
8
10

Kontrolle der Flexibilität der konstruierten Funktion

Je mehr Basisfunktionen, desto mehr Flexibilität, so dass die Regressionsfunktion immer genauer den Daten folgen wird ($\rightsquigarrow$ Varianz hoch, Bias niedrig).

Zu wenige Basisfunktionen: Regressionsfunktion kann Muster möglicherweise nicht genau abbilden ($\rightsquigarrow$ Bias hoch, Varianz niedrig).

Wie finden wir den richtigen Mittelweg im Sinne des **Bias-Varianz-Trade-offs**?

1. versuche geeignetes K zu wählen

   oder

2. nehme großes K, so dass zunächst einmal fast beliebig viel Flexibilität vorhanden ist — penalisiere dann aber Krümmung der geschätzten Funktion

Wir konzentrieren uns auf 2., da dieser Ansatz günstigere Eigenschaften hat und sich 1. schlichtweg nicht bewährt hat.

Nichtparametrische Regression durch **penalisierte B–Splines** (bzw. kurz: **P–Splines**) beinhaltet die Minimierung des folgenden Zielkriteriums:

$$\sum_{i=1}^{n}\Big(y_i - \big(\gamma_1 B_1(x_i) + \ldots + \gamma_K B_K(x_i)\big)\Big)^2 + \lambda \sum_{k=3}^{K} (\Delta^2 \gamma_k)^2$$

Hierbei ist λ wieder ein (festzulegender) **Glättungsparameter**, und Δ^2 bezeichnet die Differenzen 2. Ordnung, d.h.

$$\Delta^2 \gamma_k = \Delta(\gamma_k - \gamma_{k-1}) = (\gamma_k - \gamma_{k-1}) - (\gamma_{k-1} - \gamma_{k-2})$$

Man beachte, dass $\Delta^2 \gamma_k = 0$ genau dann, wenn γ_{k-2}, γ_{k-1} und γ_k auf einer Gerade liegen, die Krümmung in diesem Bereich also gleich Null ist.[8]

[8] umgekehrt ist $\Delta^2 \gamma_k$ genau dann hoch, wenn γ_{k-2}, γ_{k-1} und γ_k eine “Zacke” aufspannen — es handelt sich hier also um eine Approximation der Gesamtkrümmung ähnlich wie bei Smoothing Splines

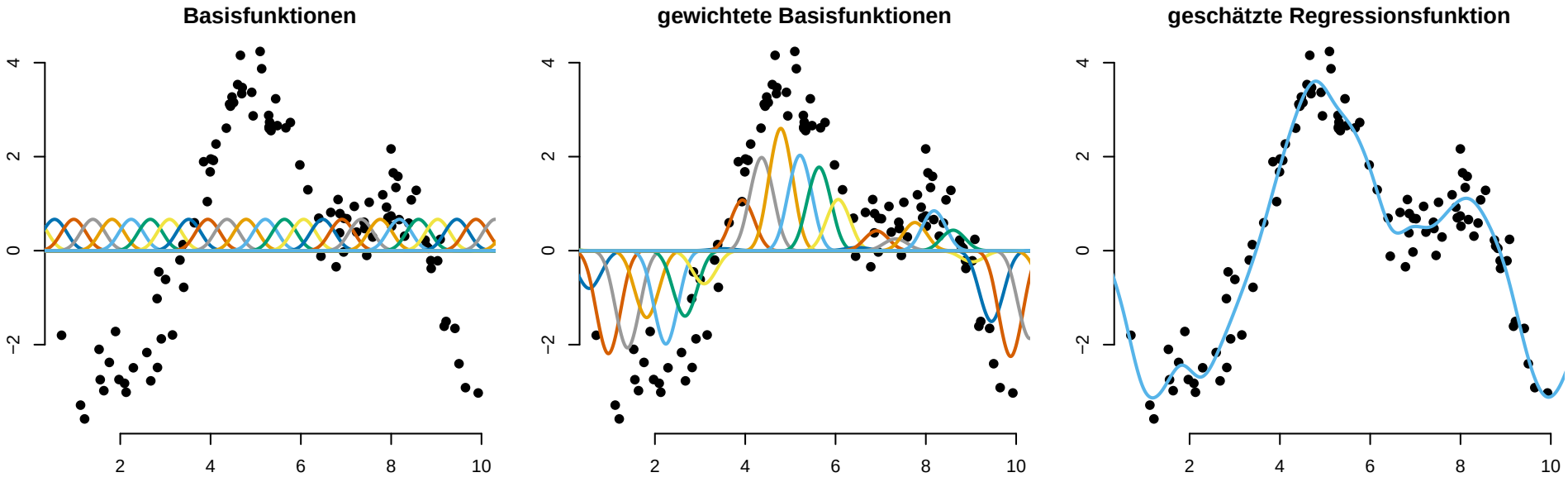


Abbildung: P-Spline-Schätzung mit $\lambda = 0.1$.

Niedriger Strafterm für Krümmung $\rightsquigarrow$ **Overfitting**

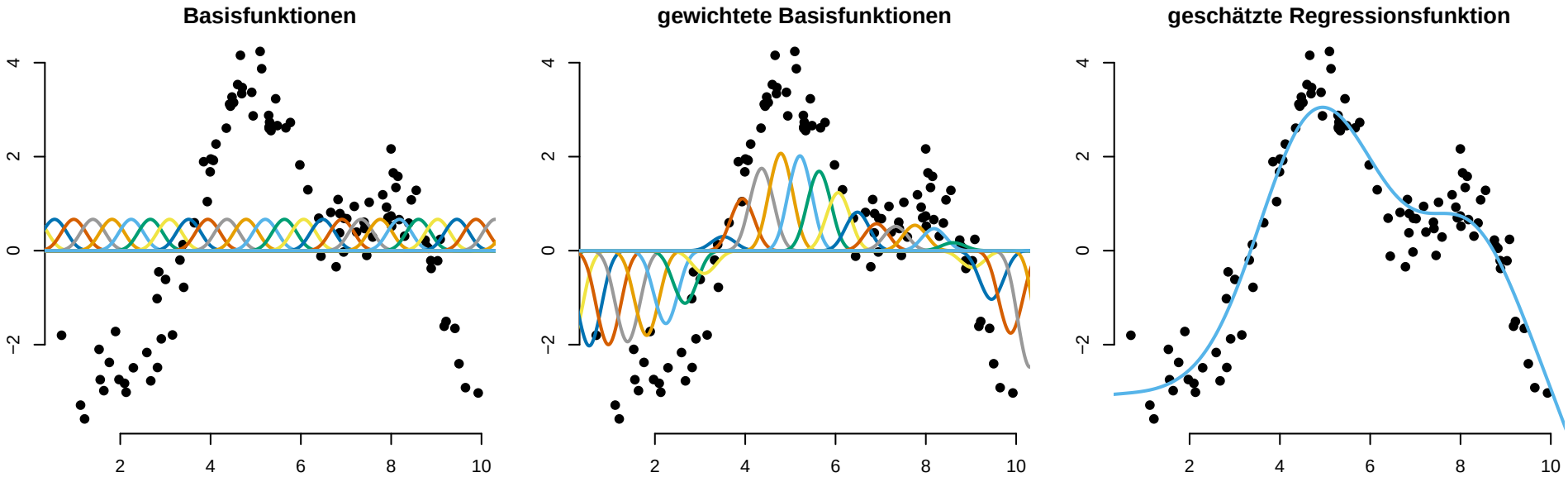


Abbildung: P-Spline-Schätzung mit $\lambda = 20$.

Moderater Strafterm für Krümmung $\rightsquigarrow$ sieht gut aus!

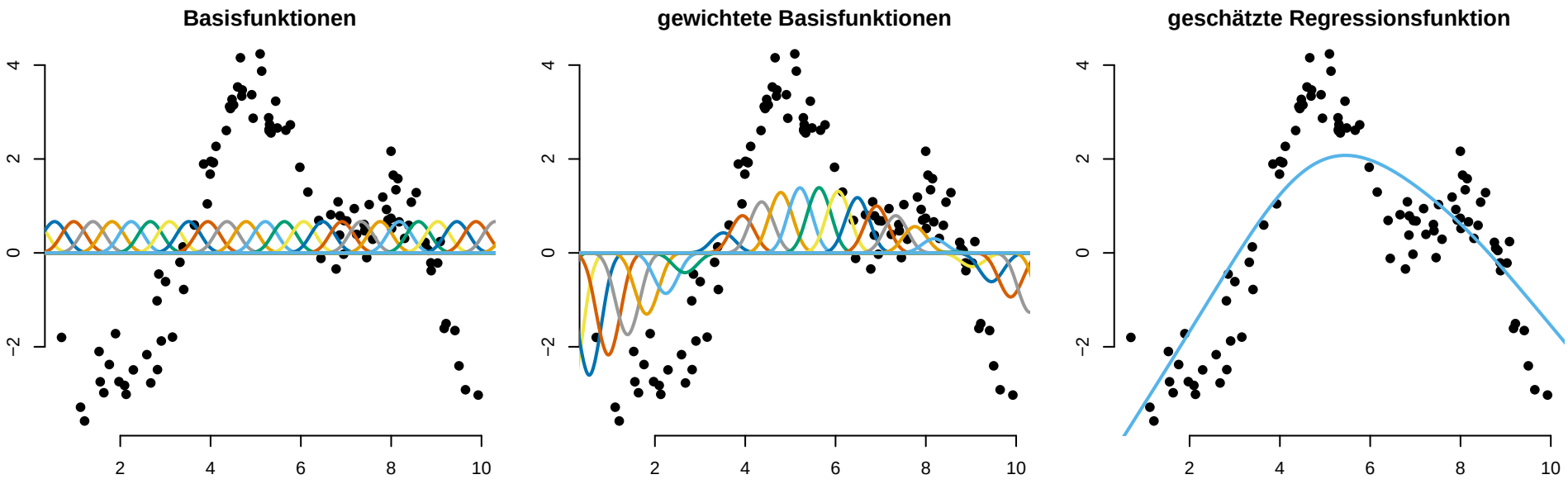


Abbildung: P-Spline-Schätzung mit $\lambda = 200$.

Etwas zu hoher Strafterm für Krümmung $\rightsquigarrow$ **Underfitting**, d.h. zu starke Betonung auf Glattheit (Unterschätzung von Gipfeln & Überschätzung von Tälern).

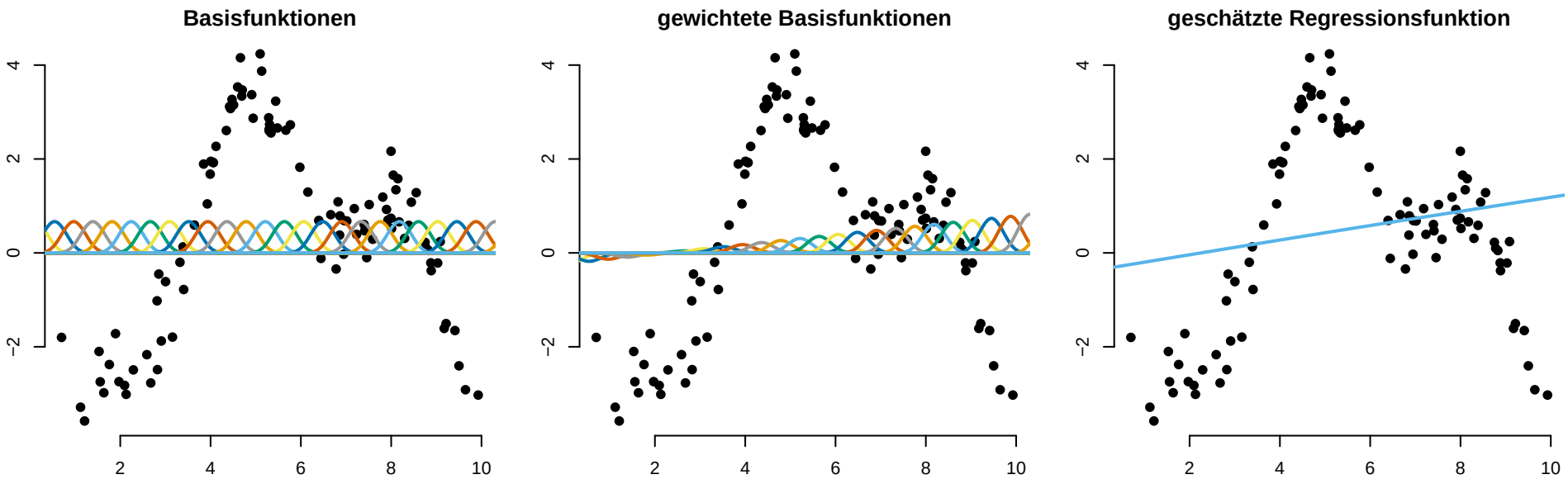
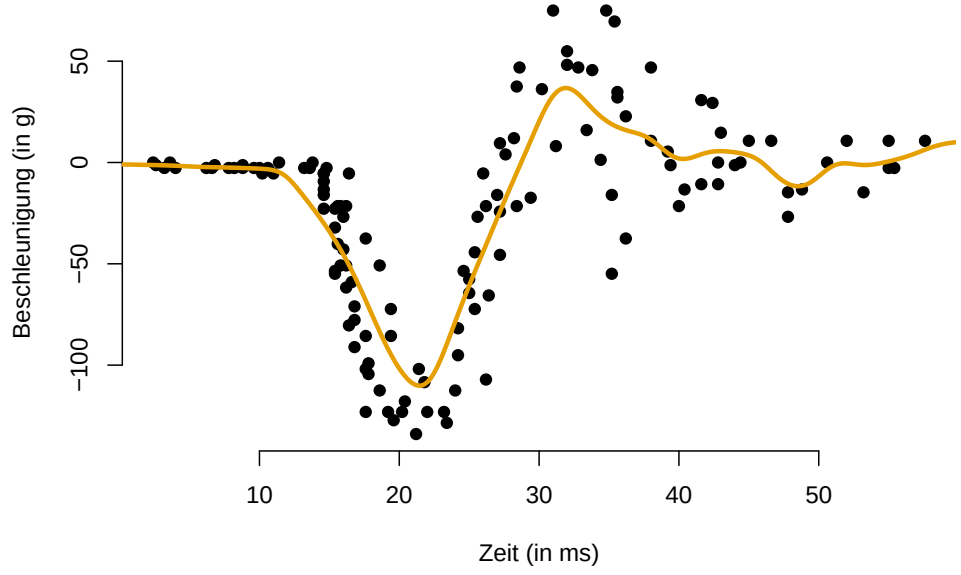


Abbildung: P-Spline-Schätzung mit $\lambda = 100000$.

Sehr hoher Strafterm für Krümmung $\rightsquigarrow$ man erhält als Schätzer die am besten passende Funktion **ohne jegliche Krümmung**, also eine Gerade.

P–Splines in R




```
# Daten einlesen und Streudiagramm erstellen:
Time <- MASS::mcycle$times
Acceleration <- MASS::mcycle$accel
plot(Time, Acceleration, pch=19, xlab="Zeit (in ms)", ylab="Beschleunigung (in g)", bty="n")

# Paket "mgcv" installieren und laden:
install.packages("mgcv")
library(mgcv)

# P-Spline Schätzung durchführen und einzeichnen:
mod <- gam(Acceleration~s(Time), bs="ps", sp=0.005) # sp ist effektiv gleich unser lambda
z <- seq(0, 70, length=1000)
lines(z, predict(mod,data.frame(Time=z)), col="#E69F00", lwd=3)
```

Generalisierte additive Modelle (GAMs)

“Curse of dimensionality”

Kann man ein Modell der Form

$$Y = s(x_1, \ldots, x_p) + \epsilon$$

auch mit $p > 2$ sinnvoll nichtparametrisch schätzen?

Antwort: ja, aber der Schätzer ist sehr “datenhungrig”: Um den 3D–Raum ähnlich gut abzudecken wie den 1D–Raum, braucht man z.B. 100 Mal so viele Daten[9].

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

[9] und nicht etwa nur 3 Mal so viele

Additive Zerlegung der Regressionsfunktion

In der Praxis wird meist eine **additive Struktur** angenommen, d.h.

$$s(x_1, \ldots, x_p) = s_1(x_1) + \ldots + s_p(x_p).$$

Hiermit wird das Problem des “curse of dimensionality” umgangen:

- effektiv haben wir so ein univariates Schätzproblem pro Variable
- im univariaten Fall liegen mehr Datenpunkte pro Umgebung, die lokale Schätzung funktioniert hier also!

$\rightsquigarrow$ restriktive Annahme, da Interaktionseffekte nicht mehr abgebildet werden

$\rightsquigarrow$ wegen des “curse of dimensionality” hat man oft aber keine andere Wahl

Generalisierte additive Modelle (GAMs) beinhalten drei Komponenten:

1. Verteilungsannahme für die Zielvariable[10]

2. additiver Prädiktor mit nichtlinearen Effekten der erklär. Variablen:

$$\eta = \beta_0 + s_1(x_1) + \ldots + s_p(x_p)$$

   (einzelne s_j können aber auch lineare/quad. Funktionen sein)

3. Transformation des linearen Prädiktors, damit $E(Y)$ den für die gegebene Verteilungsannahme richtigen Wertebereich hat[11]

- ⇝ flexibler als parametrische Modelle (wg. nichtlinearer Effekte)
- ⇝ weniger flexibel als nichtparametrische Modelle in denen $s(x_1, \ldots, x_p)$ ohne Annahme der additiven Zerlegung geschätzt wird
- ⇝ in der anwendungsorientierten Forschung inzwischen wahnsinnig populär

[10] z.B. normal, Poisson, Bernoulli

[11] z.B. exp bei Poisson, inverse logistische Funktion bei Bernoulli

GAM mit normalverteilter Zielgröße

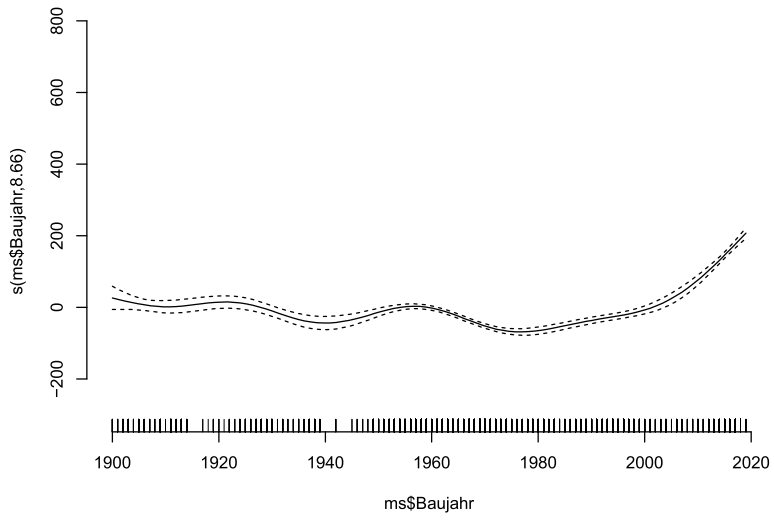
Mögliches GAM für eine normalverteilte Zielvariable Y_i:

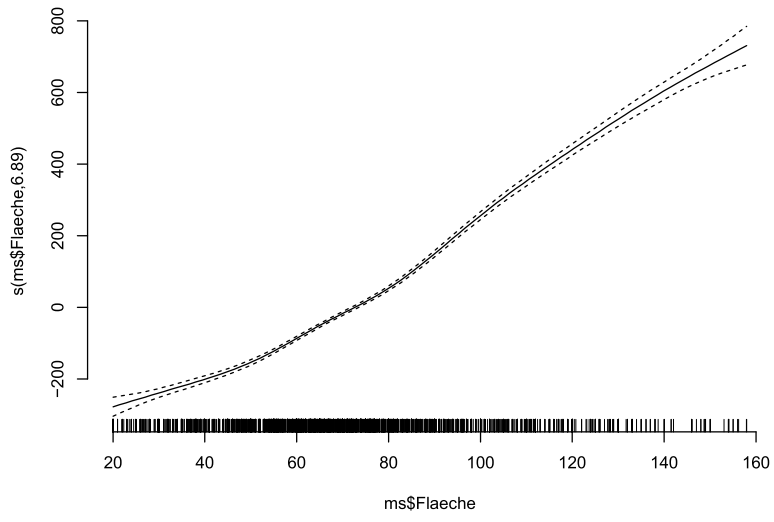
$$Y_i \sim N(\mu_i, \sigma^2),$$

$$\mu_i = E(Y_i) = \beta_0 + s_1(x_{i1}) + \ldots + s_p(x_{ip})$$

Im Vgl. zum linearen Modell werden also einfach die linearen Effekte $\beta_j x_{ij}$ durch (nichtparametrisch zu schätzende) nichtlineare Effekte $s_j(x_{ij})$ ersetzt.

GAM im Mietspiegel–Beispiel





Geschätzte nichtlineare Effekte $\hat{s}_1(\text{Baujahr}_i)$ und $\hat{s}_2(\text{Flaeche}_i)$ im GAM

$$\text{Miete}_i \sim N(\mu_i, \sigma^2), \quad \mu_i = E(\text{Miete}_i) = 547.14 + \hat{s}_1(\text{Baujahr}_i) + \hat{s}_2(\text{Flaeche}_i)$$

```
Mietdaten <- read.csv("http://www.rolandlangrock.com//Daten//rents.csv")
attach(Mietdaten)
# library(mgcv)
# P-Spline Schätzung durchführen und einzeichnen:
mod <- gam(rent~s(year)+s(area), bs="ps")
plot(mod)
```

GAM mit Poisson–verteilter Zielgröße

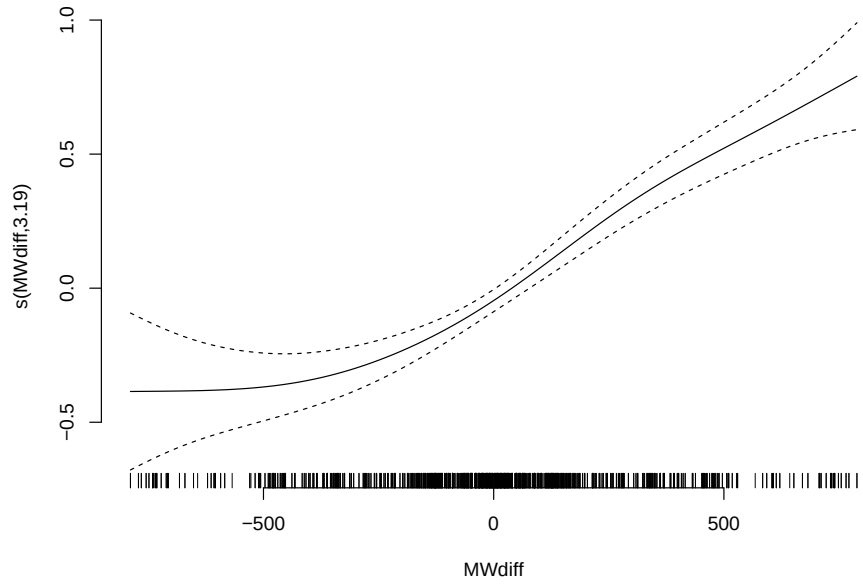
Mögliches GAM für eine Zählvariable Y_i:

$$Y_i \sim Po(\lambda_i),$$

$$\lambda_i = E(Y_i) = e^{\beta_0 + s_1(x_{i1}) + \ldots + s_p(x_{ip})}$$

Auch hier wird im Vgl. zum Poissonregressionsmodell einfach jeder lineare Effekt durch einen nichtlinearen ersetzt.

GAM im Fussballbeispiel



Geschätzter nichtlinearer Effekt $\hat{s}_1(\text{MWdiff}_i)$ im Poisson-GAM

$$\text{Tore}_i \sim Po(\lambda_i), \qquad \lambda_i = E(\text{Tore}_i) = e^{0.255+\hat{s}_1(\text{MWdiff}_i)+0.242\cdot\text{Heim}_i}$$

```
Bundesliga <- read.csv("http://www.rolandlangrock.com//Daten//Bundesliga.csv")
attach(Bundesliga)
# library(mgcv)
# P-Spline Schätzung durchführen und einzeichnen:
mod <- gam(Tore~s(MWdiff)+Heim, bs="ps")
plot(mod)
```

Zusammenfassung: nichtparametrische Regression

- mit den besprochenen nichtparametrischen Ansätzen haben wir unseren Regressions-Werkzeugkasten noch einmal stark erweitert
- **es geht immer darum, den optimalen Bias-Varianz-Trade-off zu finden**
- Einsatzgebiete nichtparametrischer Regressionsmodelle:
  - explorative Datenanalyse — schnell mal gucken, wie der Zshg. aussehen könnte
  - wenn parametrische Modelle zu unflexibel sind
  - wenn ein parametrischer Zshg. aus konzeptuellen Gründen unsinnig erscheint (typisches Beispiel: Längen- und Breitengrad als erklärende Variablen)