Tag der Mathematik 2009

Survival of the Fittest
Wie statistische Modelle an Daten angepasst werden

Thomas Kneib

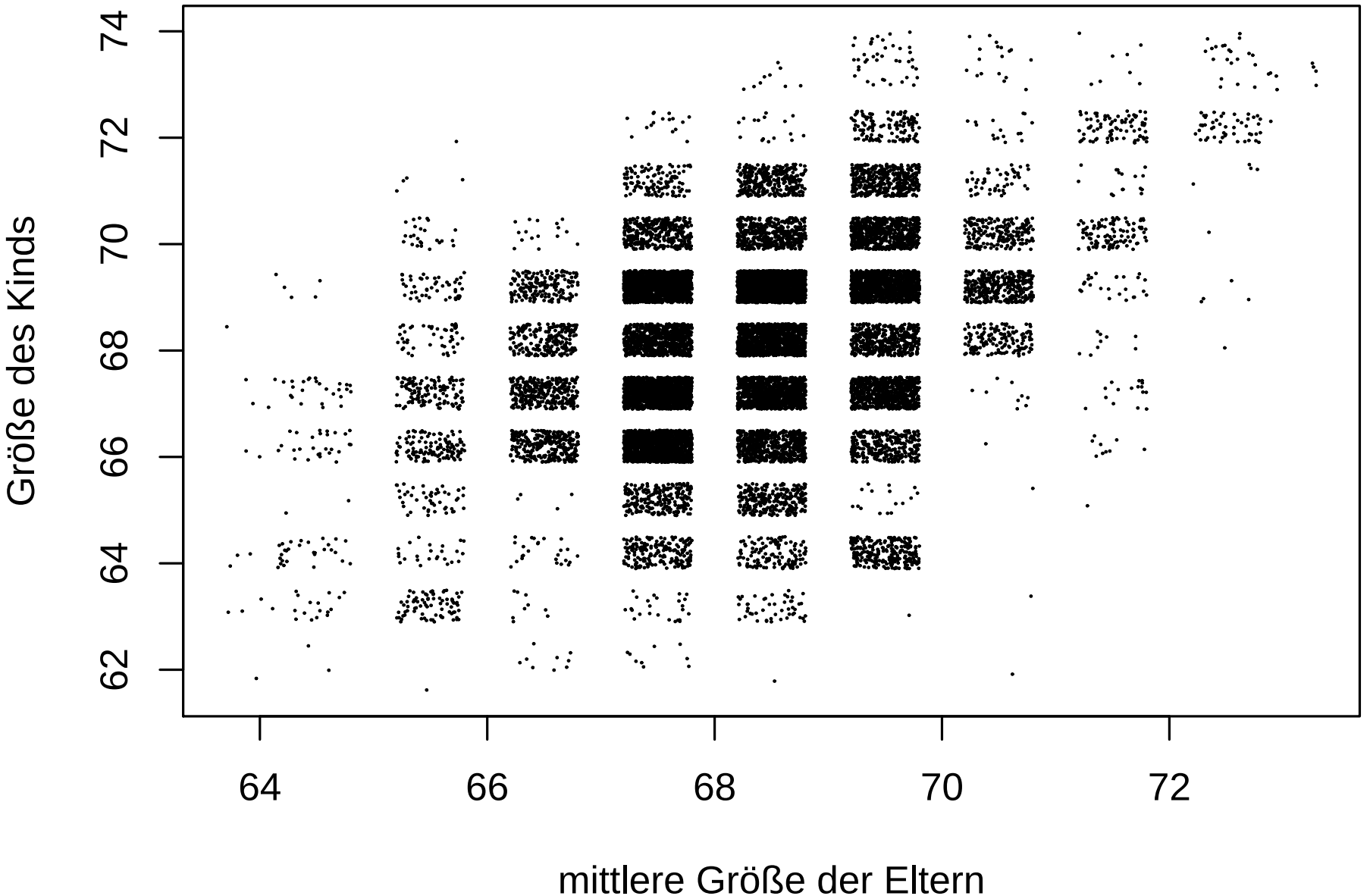
Fakultät für Mathematik und Naturwissenschaften
Carl von Ossietzky Universität Oldenburg

CARL VON OSSIETZKY universität OLDENBURG
FAKULTÄT V
MATHEMATIK UND NATURWISSENSCHAFTEN

Oldenburg, 28.10.2009

Statistische Modellierung

- Francis Galton (1822 – 1911): Wie hängen Körpergröße der Eltern und Körpergröße der Kinder zusammen?
- Zu erklärende Variable: Körpergröße des Kindes in Inch.
- Erklärende Variable: Mittlere Körpergröße der Eltern in Inch. Um Männer und Frauen vergleichbar zu machen, multiplizierte Galton die Körpergröße der Frauen mit 1.08.
- Daten zu 18.918 Kindern.


Größe des Kinds
62
64
66
68
70
72
74
64
66
68
70
72
mittlere Größe der Eltern

- Prinzipien statistischer Modellierung:
  - Ziel: Vereinfachende Beschreibung (möglicherweise komplexer) Zusammenhänge.
  - Gleichzeitig Berücksichtigung der Unsicherheit aufgrund unvollständiger Information (z.B. aufgrund der Ziehung einer Stichprobe).
  - Schätzung des Modells basierend auf Daten bzw. Anpassung des Modells an die Daten.
  - Prognose zukünftiger Beobachtungen aus einem geschätzten Modell.

- Anpassung ist einer der Schlüsselbegriffe in Charles Darwins (1809 – 1882) Evolutionstheorie :

  "Survival of the Fittest" = "Überleben des am Besten Angepassten"

  (eigentlich geprägt durch den Sozialphilosophen Herbert Spencer).

- Wird häufig in Zusammenhang gebracht und verwechselt mit:

  "Only the strong survive" = "Überleben des Stärkeren".

- Francis Galton war ein Halbcousin von Darwin.

- Galton wählte die folgende Form eines statistischen Modells:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, 18.918,$$

  d.h. ein lineares Modell mit

  - Zielvariable y_i (Größe des Kinds),
  - Einflussgröße x_i (mittlere Größe der Eltern),
  - Modellabweichung ε_i (Fehler).

- In der Statistik wird ε_i (und damit auch y_i) als zufällige Größe aufgefasst.

  $\Longrightarrow$ Stochastische Modellierung.

- Aufgabe der Statistik: Trennung von Signal (Gerade) und Rauschen (Fehler) sowie Abschätzung des dabei gemachten Fehlers.

- Das lineare Modell ist parametrisch und wird charakterisiert durch
  - den Achsenabschnitt β_0 und
  - die Steigung β_1.
- Anpassung an die gegebenen Daten: Wie kann die Übereinstimmung einer Geraden mit den Daten gemessen werden?
- Quantifiziere die Abweichung zwischen Modell und Daten.
- Gaußsche Methode der kleinsten Fehlerquadrate: Minimiere die quadrierten Abweichungen $y_i - \beta_0 - x_i\beta_1$, d.h.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \rightarrow \min_{\beta_0,\beta_1} .$$

- Lösung des Minimierungsproblems:
  - Ableiten der Quadratsumme nach β_0 und β_1.
  - Nullsetzen der resultierenden Ausdrücke.
  - Auflösen nach β_0 und β_1.
  - Überprüfen, dass es sich tatsächlich um ein Minimum handelt
- Man erhält eine Lösung in geschlossener Form.
- Im Beispiel ergeben sich

$$\hat{\beta}_0 = 27.777 \qquad \hat{\beta}_1 = 0.591$$

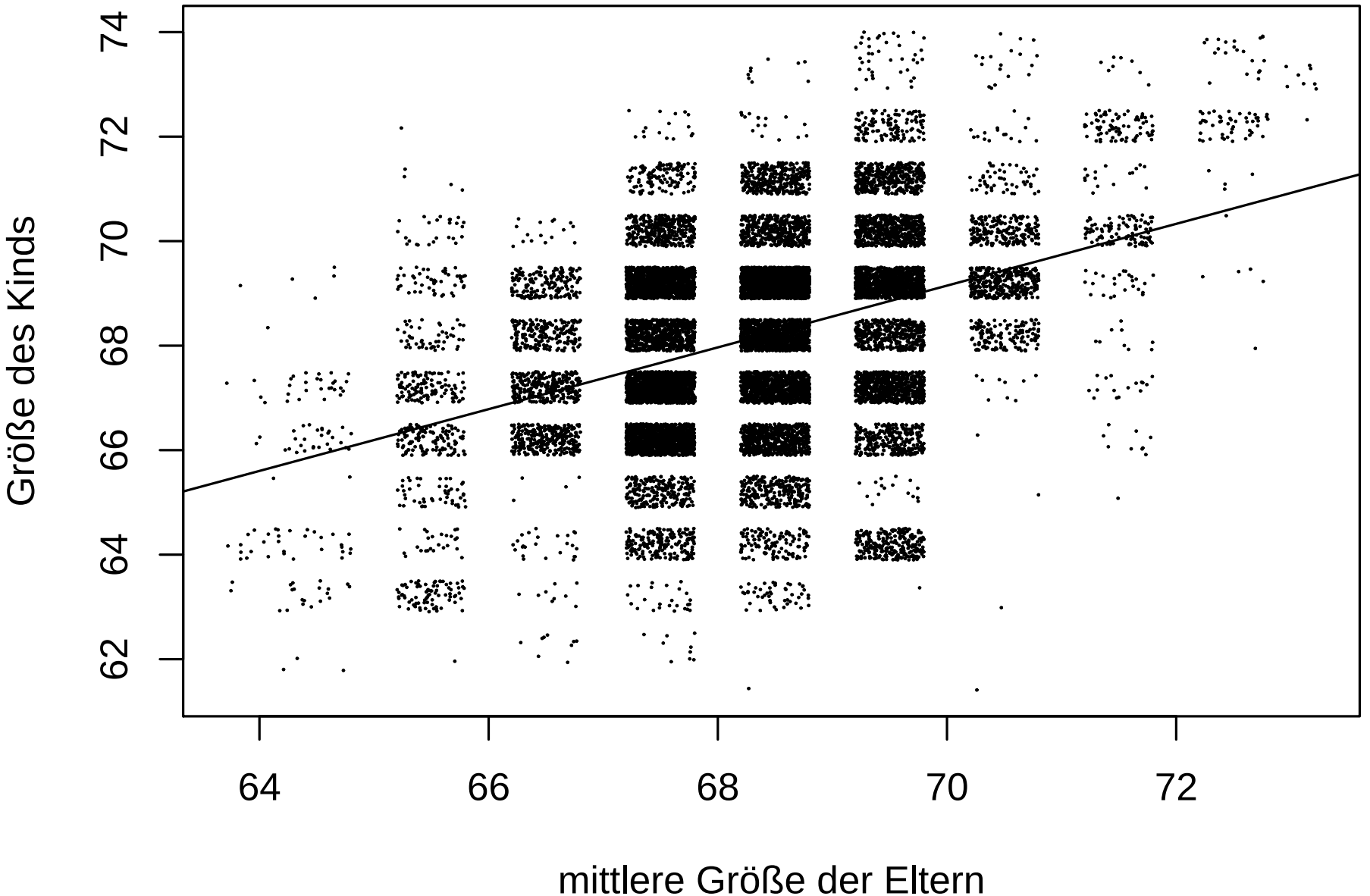
als Punktschätzer und

$$[29.156, 26.397] \qquad [0.571, 0.611]$$

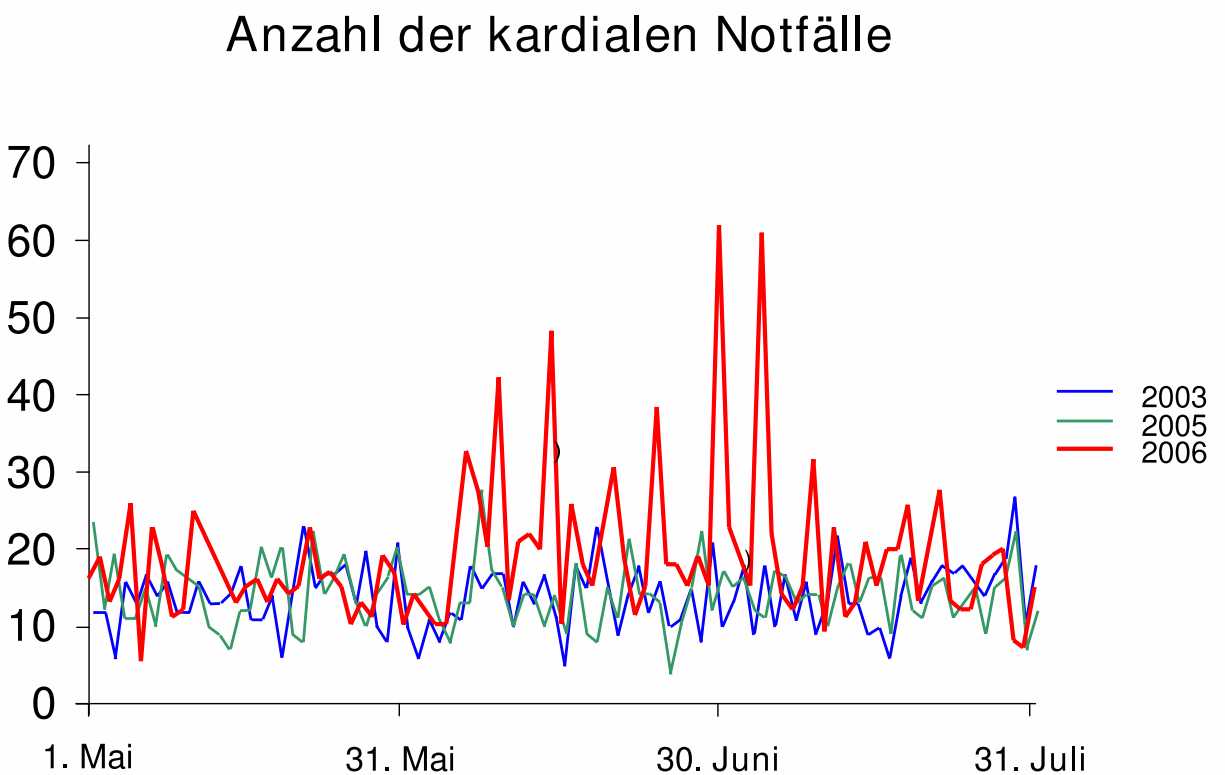
als entsprechende Unsicherheitsbereiche.

- Interpretation:
  - Die geschätzte Steigung $\hat{\beta}_1$ gibt an, um wieviele Einheiten sich die erwartete Körpergröße des Kindes ändert, wenn sich die durchschnittliche Größe der Eltern um eine Einheit ändert.
  - $\hat{\beta}_1 < 1$ bedeutet, dass Kinder großer Eltern tendenziell kleiner sind als ihre Eltern während Kinder kleiner Eltern tendenziell größer sind als ihre Eltern.
  - Dieses Phänomen bezeichnete Galton als "Regression (Rückkehr) zum Mittelwert" ("Regression towards mediocrity in hereditary stature").
- Prognose für neue Beobachtungen:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

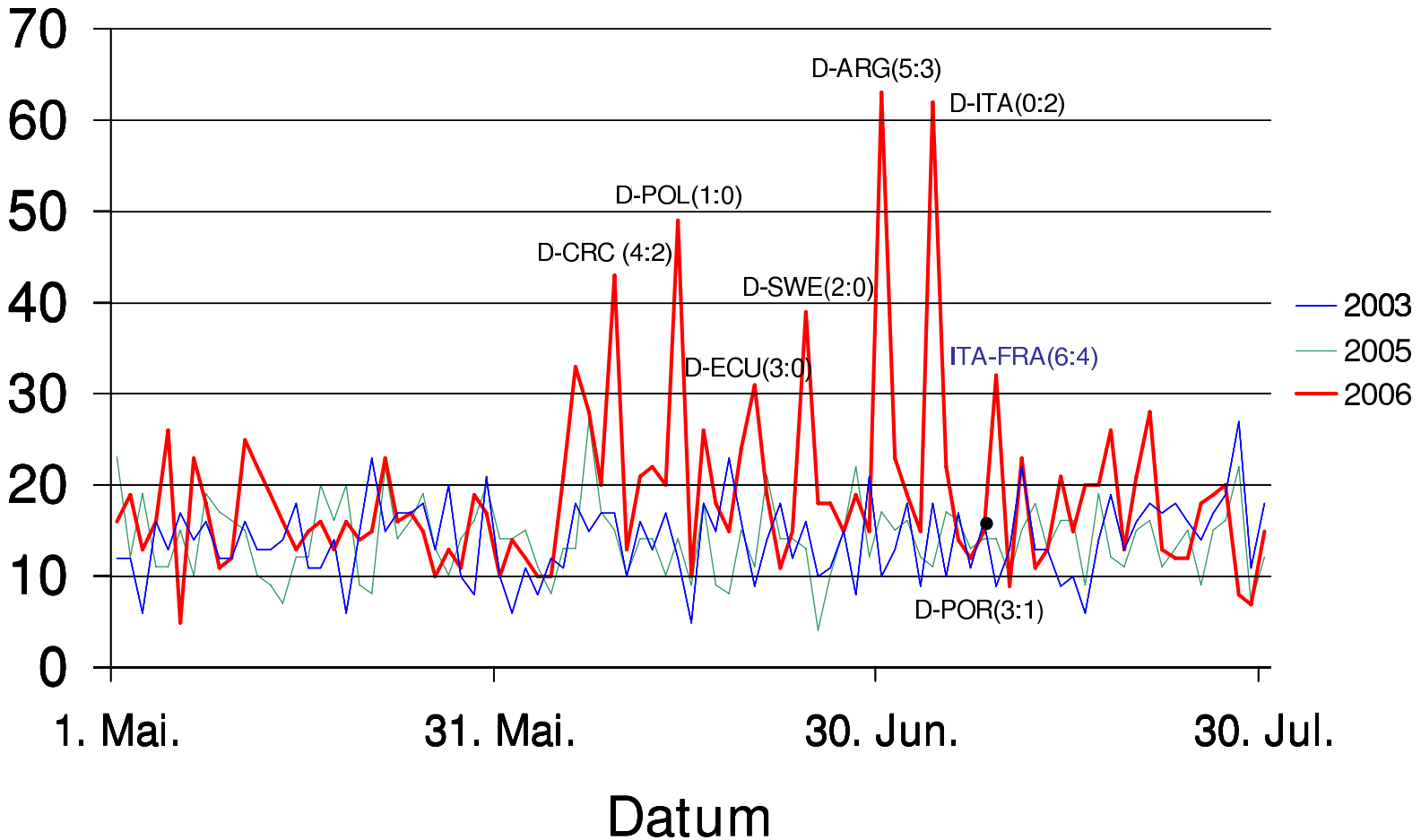

Größe des Kinds
62
64
66
68
70
72
74
64
66
68
70
72
mittlere Größe der Eltern

Kardiovaskuläre Notfälle in München



- Was unterscheidet 2006 von früheren Jahren?

Zahl der kardiovaskulären Notfälle



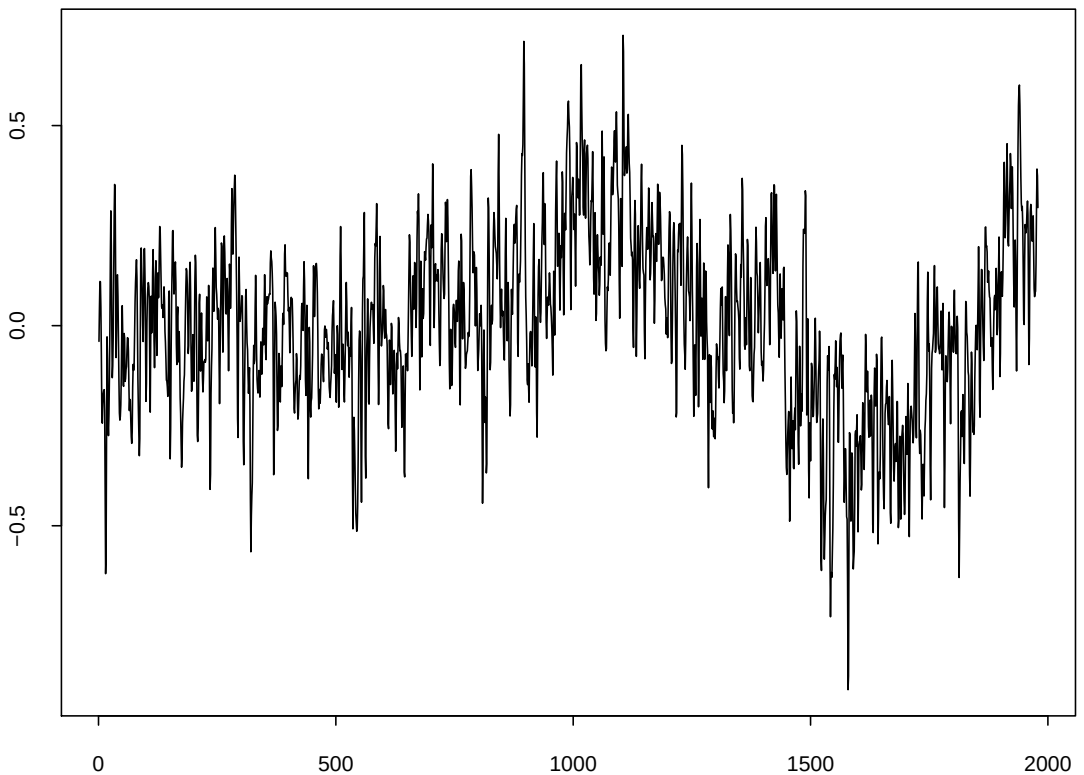
- Modellierung von Häufigkeiten $\Rightarrow$ Kleinste-Quadrate-Schätzung nicht geeignet.
- Verwende ein stochastisches Modell basierend auf der Poisson-Verteilung.
- Unterscheidung zwischen normalen Tagen (keine Spiele der deutschen Mannschaft) und Tagen mit Spielen der deutschen Mannschaft.
- Relatives Risiko:

$$\frac{\text{Risiko an Tagen mit Spielen der deutschen Mannschaft}}{\text{Risiko an normalen Tagen}}$$

- Das geschätzte relative Risiko ist 2.66 (Unsicherheitsbereich 2.33 – 3.04).
- Erhöhtes Risiko insbesondere für Männer mit Herzerkrankung (relatives Risiko 4.22).
- Kein erhöhtes Risiko für Frauen ohne Herzerkrankung nachweisbar.

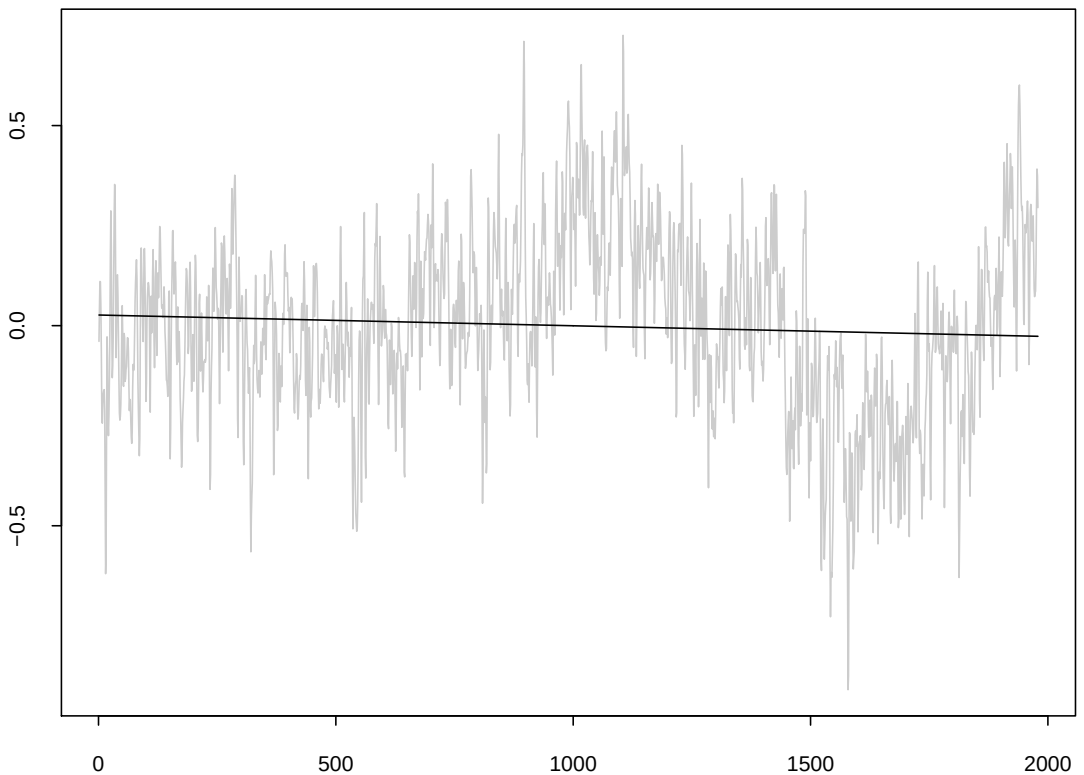
Nichtlineare Modellierung

- In zahlreichen Anwendungen sind lineare Modelle nicht flexibel genug.
- Beispiel: Temperatur-Rekonstruktion auf der nördlichen Erdhalbkugel für die letzten 2000 Jahre.



Nichtlineare Modellierung

- In zahlreichen Anwendungen sind lineare Modelle nicht flexibel genug.
- Beispiel: Temperatur-Rekonstruktion auf der nördlichen Erdhalbkugel für die letzten 2000 Jahre.



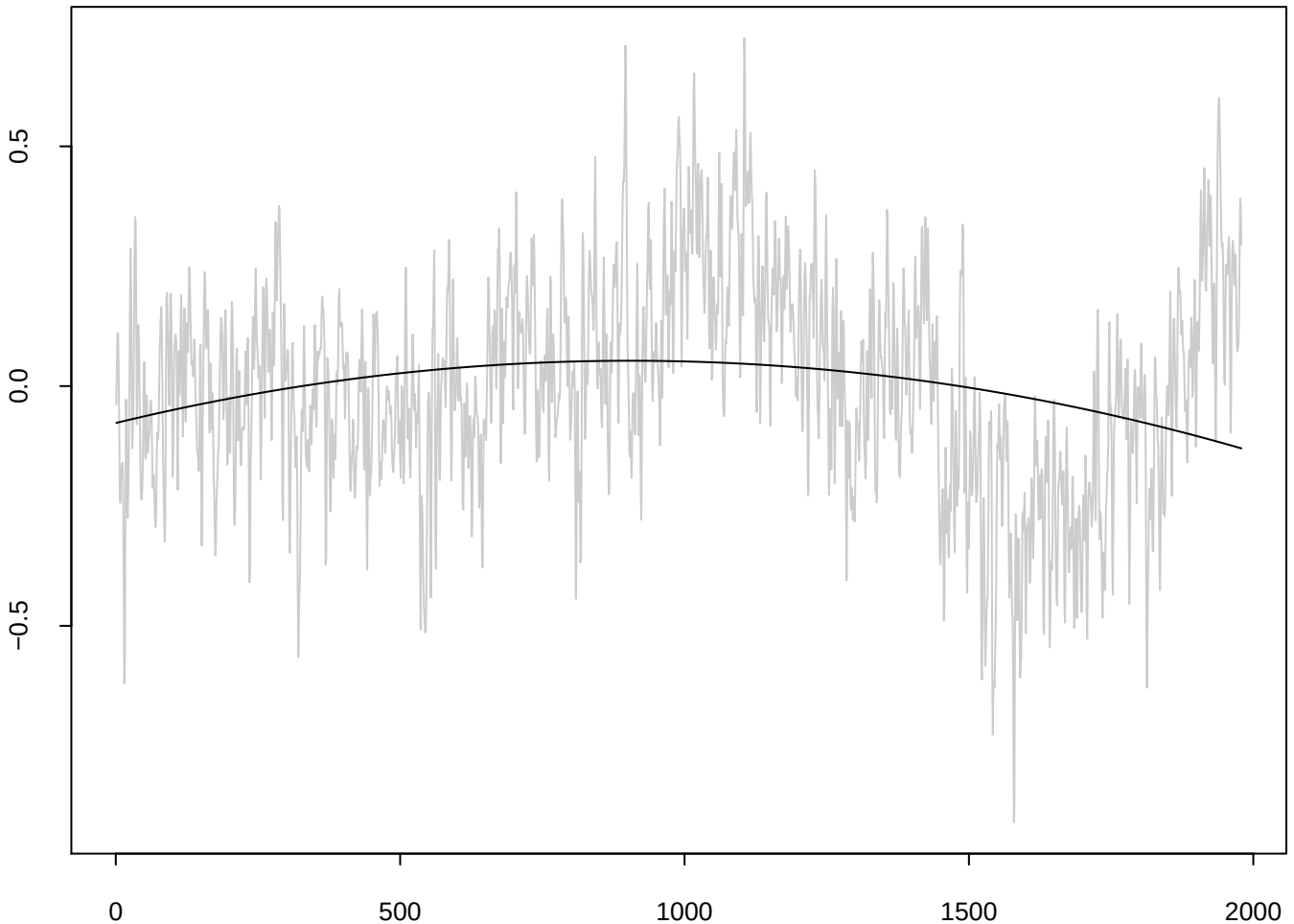
- Erweitere das lineare Modell zu einem polynomialen Modell:

$$y_i = \beta_0 + x_i\beta_1 + \ldots + x_i^p\beta_p + \varepsilon_i.$$

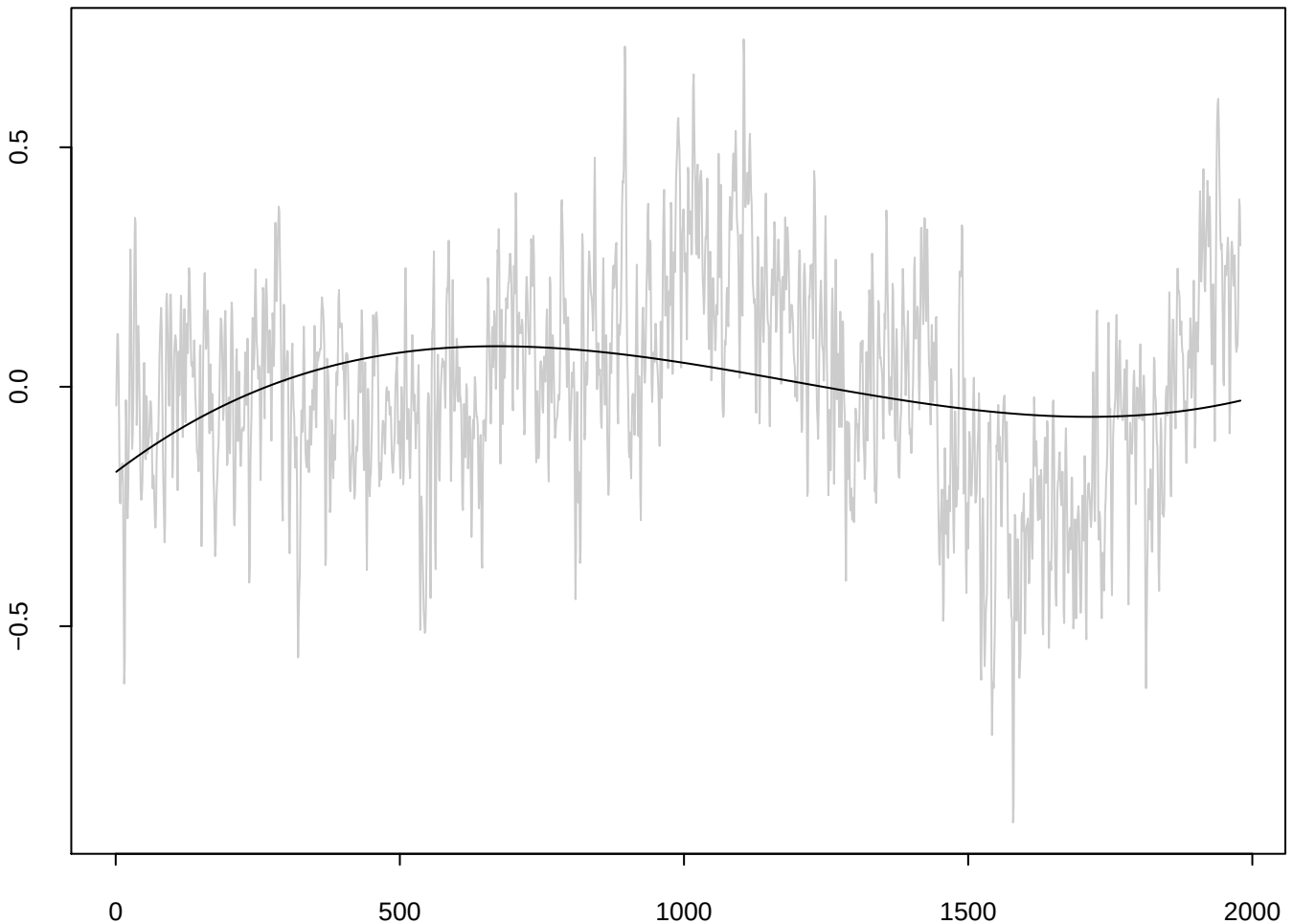
- Kleinste-Quadrate-Schätzung der Parameter:

$$\sum_{i=1}^{n}(y_i - \beta_0 - x_i\beta_1 - \ldots - x_i^p\beta_p)^2 \rightarrow \min_{\beta_0,\beta_1,\ldots,\beta_p}.$$

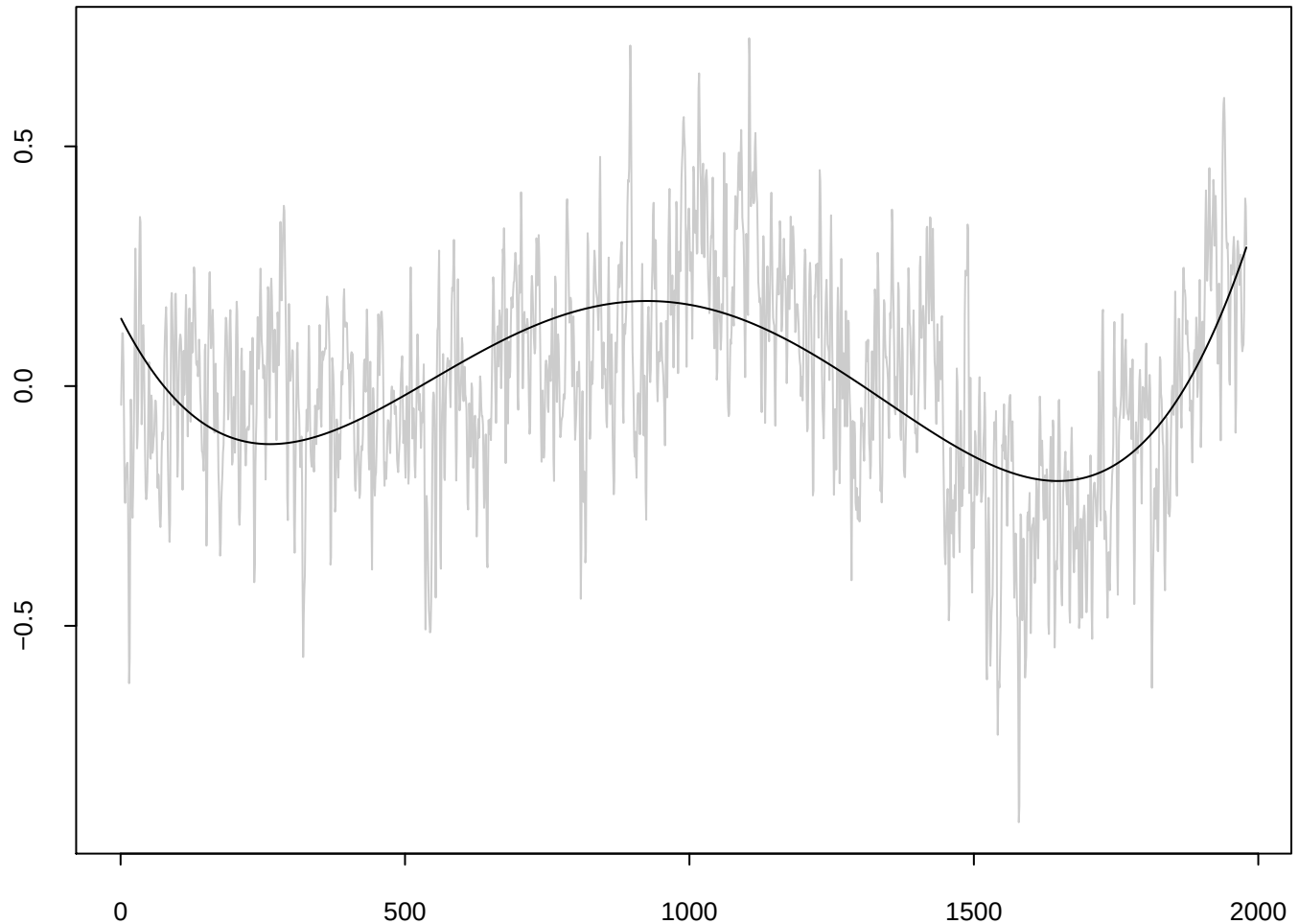
- Quadratisch:



- Kubisch:



- Quartisch:



- Auch Polynome sind nicht flexibel genug, um den nichtlinearen Zusammenhang zu beschreiben.
- Betrachte nichtparametrische Modelle der Form

$$y_i = f(x_i) + \varepsilon_i$$

wobei f datengesteuert aus den Daten geschätzt werden soll.

- Das Kleinste-Quadrate-Kriterium

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 \rightarrow \min_f$$

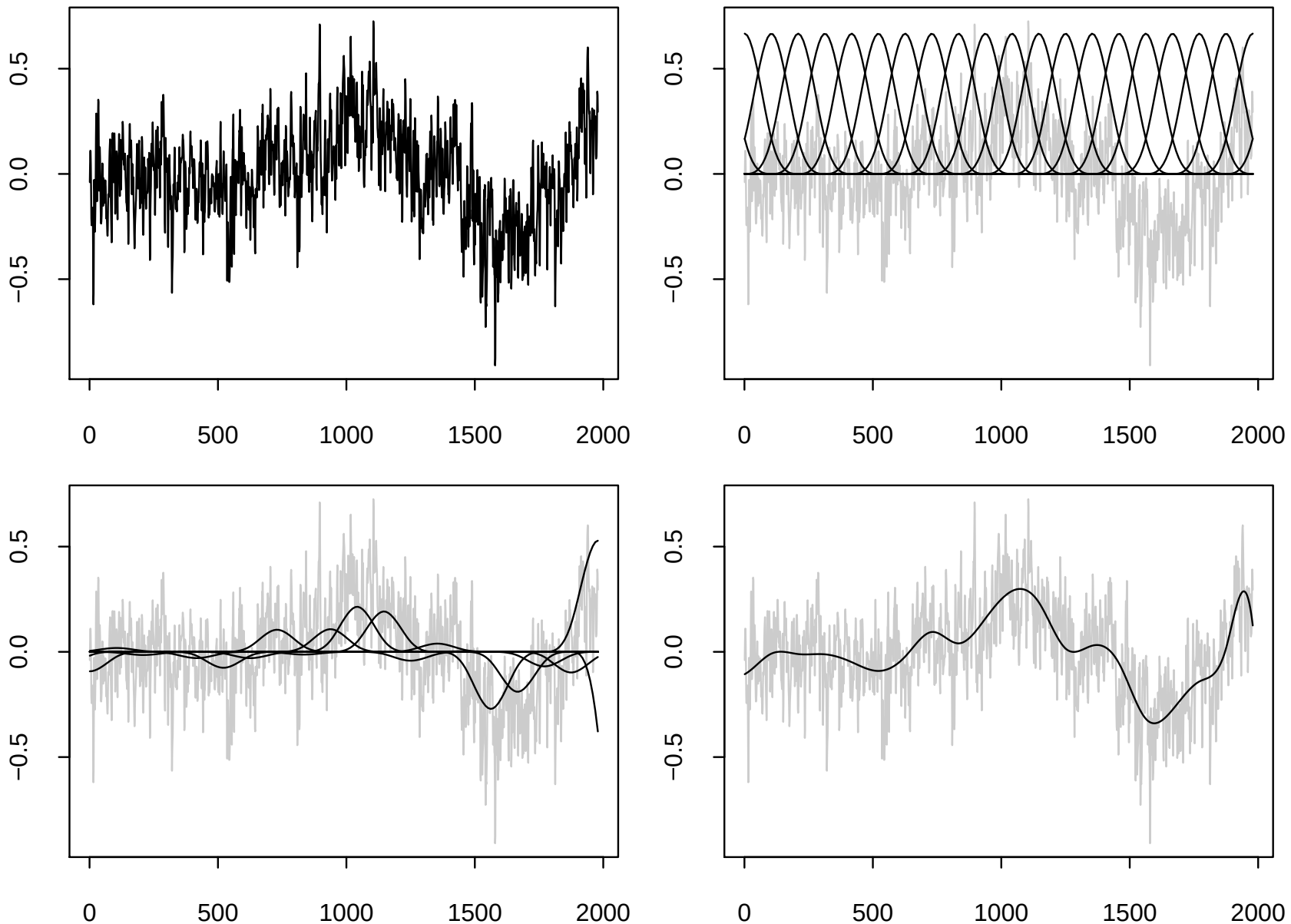
führt (ohne weitere Annahmen an f) zur Interpolation der gegebenen Daten.

$\Rightarrow$ Wähle f optimal aus einer geeigneten Funktionenklasse.

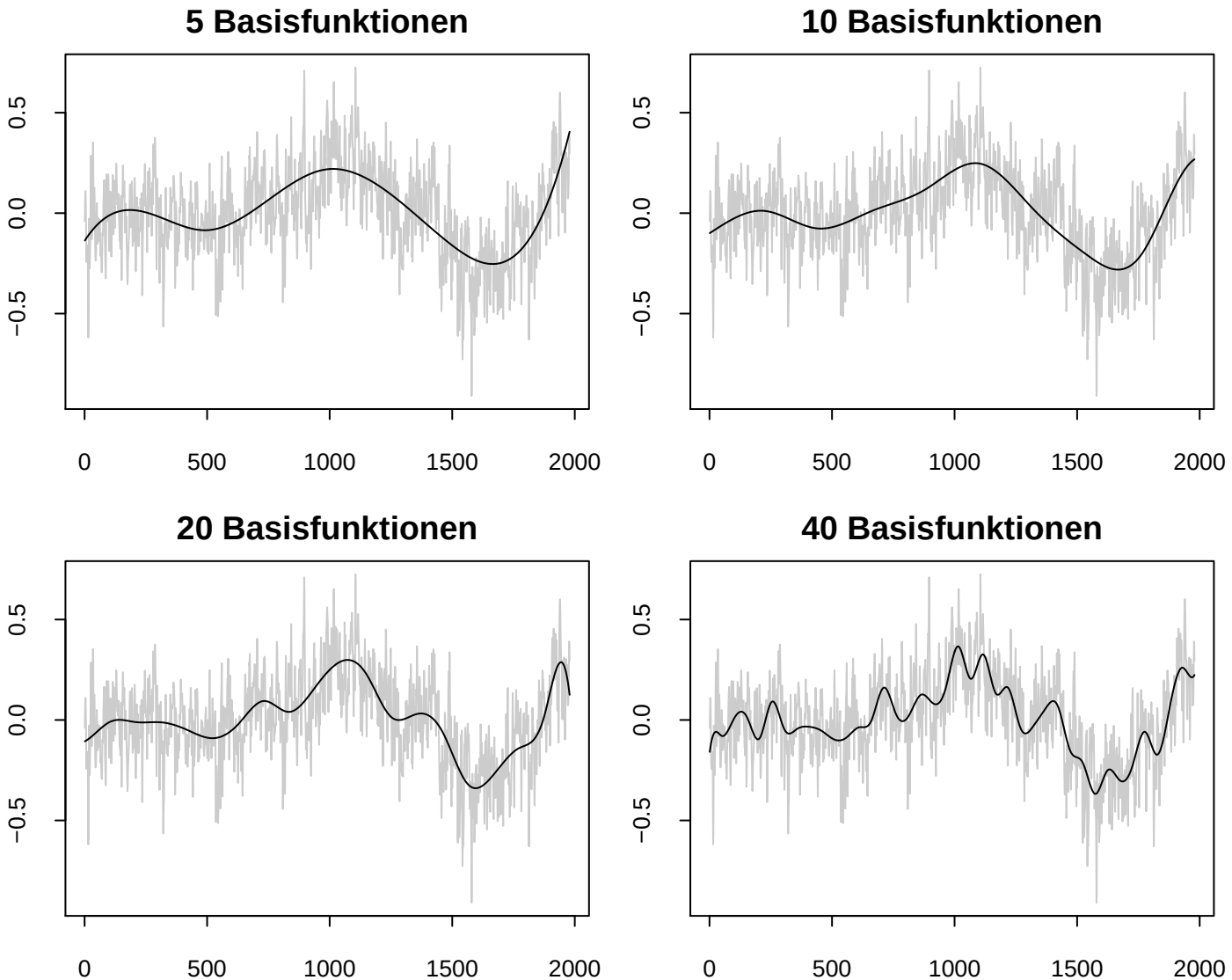
- Approximiere $f(x)$ durch Entwicklung in Basisfunktionen $B_k(x)$:

$$f(x) = \sum_{k=1}^{K} \beta_k B_k(x).$$

- Führt die nichtparametrische Schätzung von f in eine semiparametrische Schätzung mit vielen Parametern über.
- Wir verwenden Polynom-Splines basierend auf einer B-Spline-Basis.


0.5
0.0
−0.5
0
500
1000
1500
2000

- B-Spline-Schätzungen für variierende Anzahlen von Basisfunktionen:



- Die Schätzungen hängen stark von der Anzahl der Basisfunktionen ab.

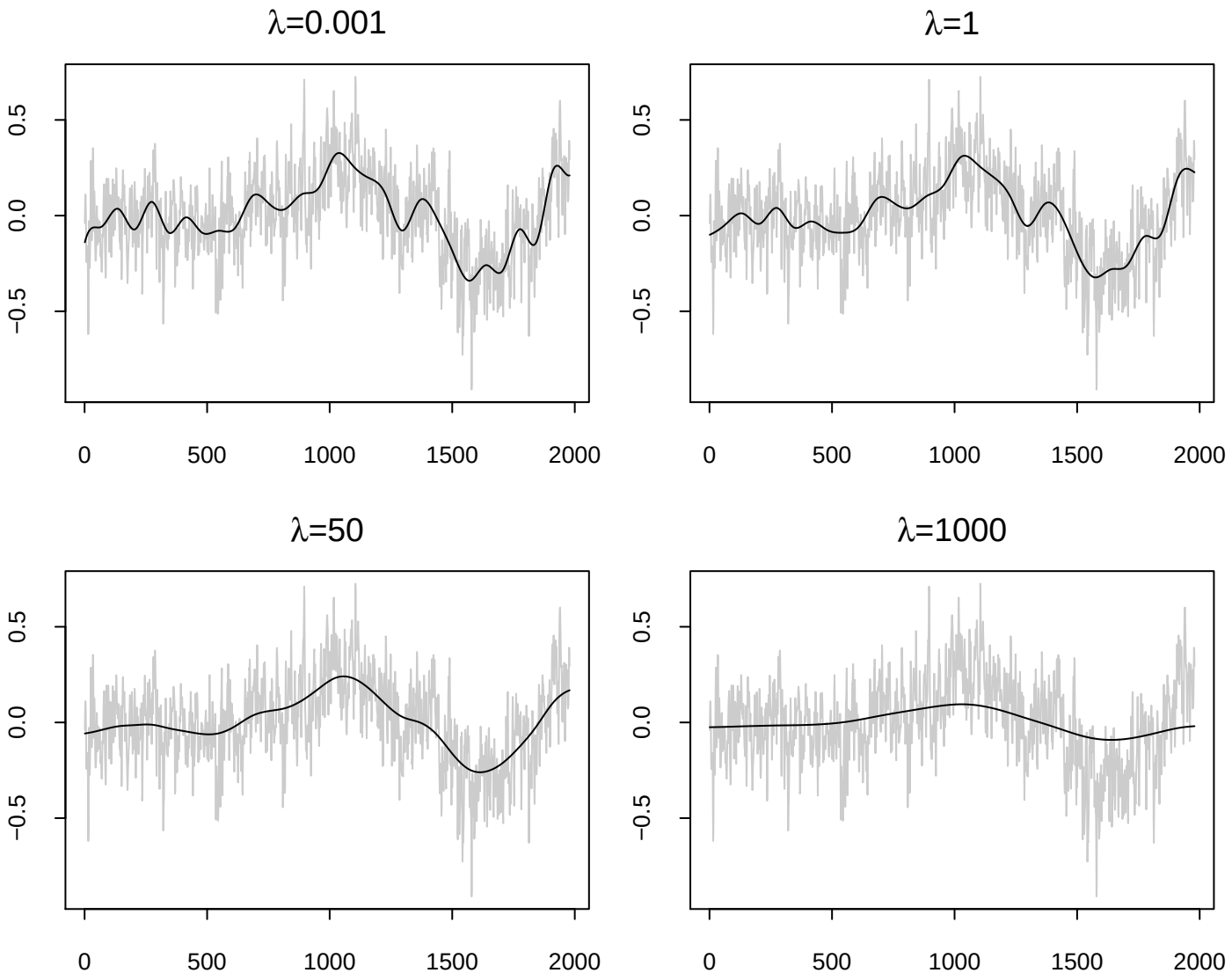
  $\Rightarrow$ Ergänze das Kleinste-Quadrate-Kriterium um einen Regularisierungs-Term der raue Funktionsschätzungen bestraft.

- Häufiger Ansatz: Bestrafung der quadrierten zweiten Ableitung

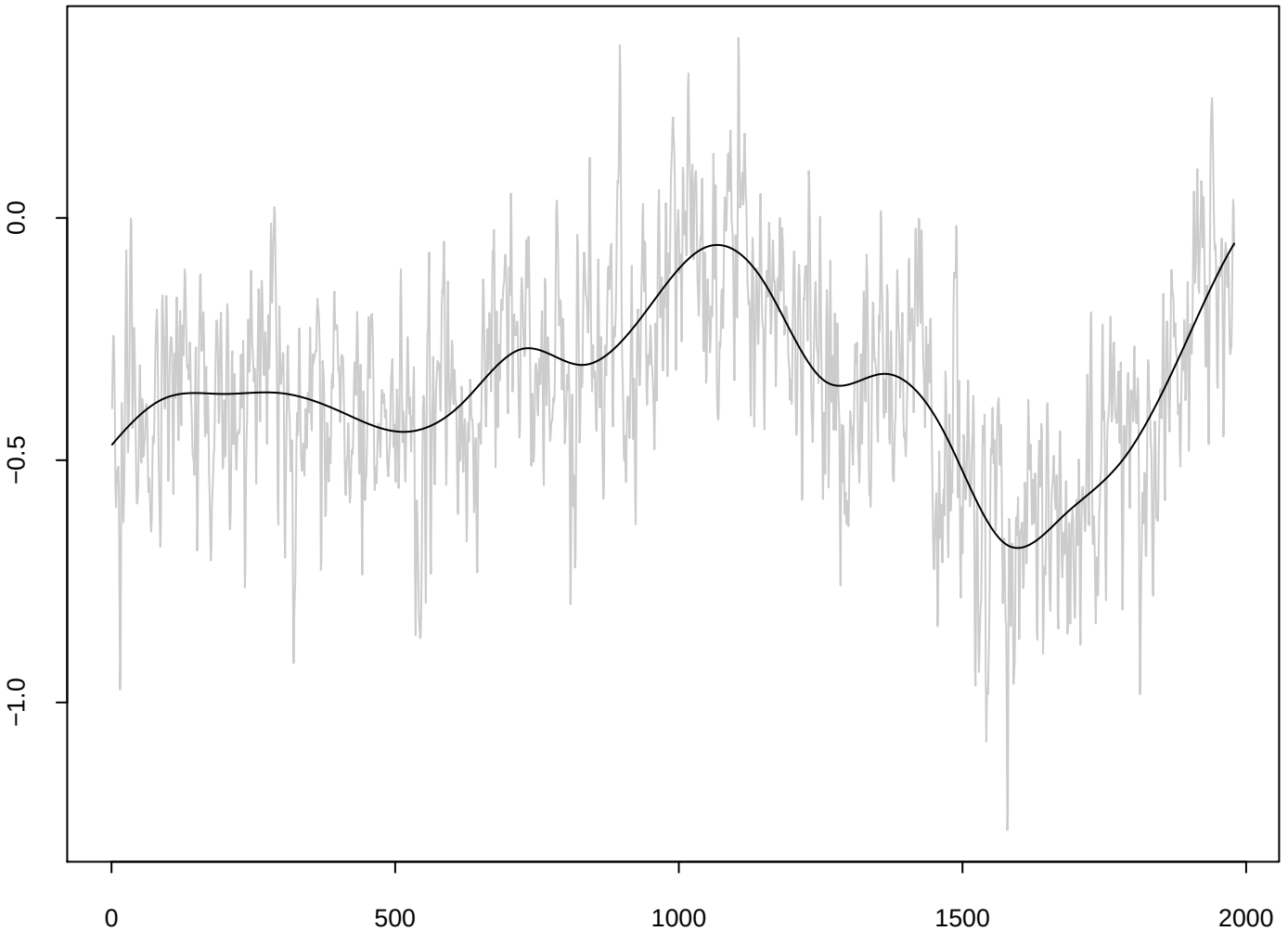
$$\text{pen}(f) = \lambda \int (f''(x))^2 dx.$$

- Der Glättungsparameter λ bestimmt den Einfluss der Regularisierung auf die Schätzung

  $\Rightarrow$ Die Schätzung des Glättungsparameters ist die eigentliche Schwierigkeit.


λ=0.001
λ=1
λ=50
λ=1000
0.5
0.0
−0.5
0
500
1000
1500
2000

- Optimale Schätzung für die Temperatur-Rekonstruktionen:



Fazit

- Die Arbeit von Statistikern ist geprägt durch eine Kombination aus Kenntnissen in
  - Mathematik (Abstraktion, Modellierung),
  - Informatik (Programmieren, wissenschaftliches Rechnen)
  - Anwendungsgebieten (Lebens-, Natur- und Wirtschaftswissenschaften, etc.).
- Aufgabe eines modernen Statistikers: Den Rohstoff Daten zu Wissen und Erkenntnis veredeln.
- Statistik – die wissenschaftliche Disziplin der verantwortungsvollen Datenanalyse.
- New York Times (5.8.2009): "the sexy job in the next 10 years will be statisticians" (Hal Varian, Chef-Ökonom von Google).
- A place called home:

`http://www.staff.uni-oldenburg.de/thomas.kneib`