

# Semiparametrische Regression

Thomas Kneib

Institut für Mathematik  
Carl von Ossietzky Universität Oldenburg

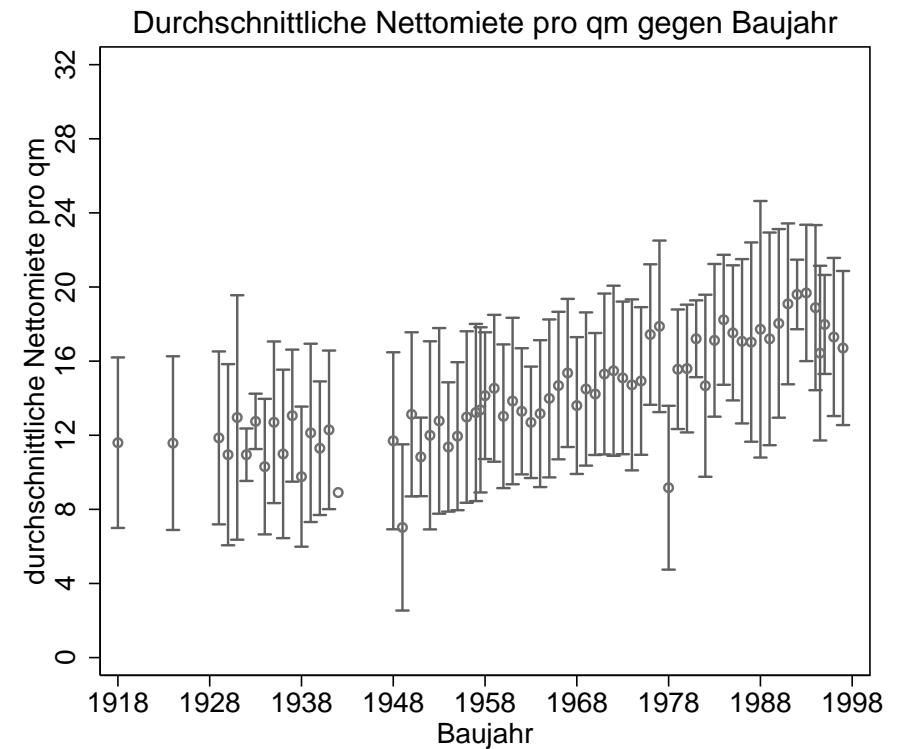
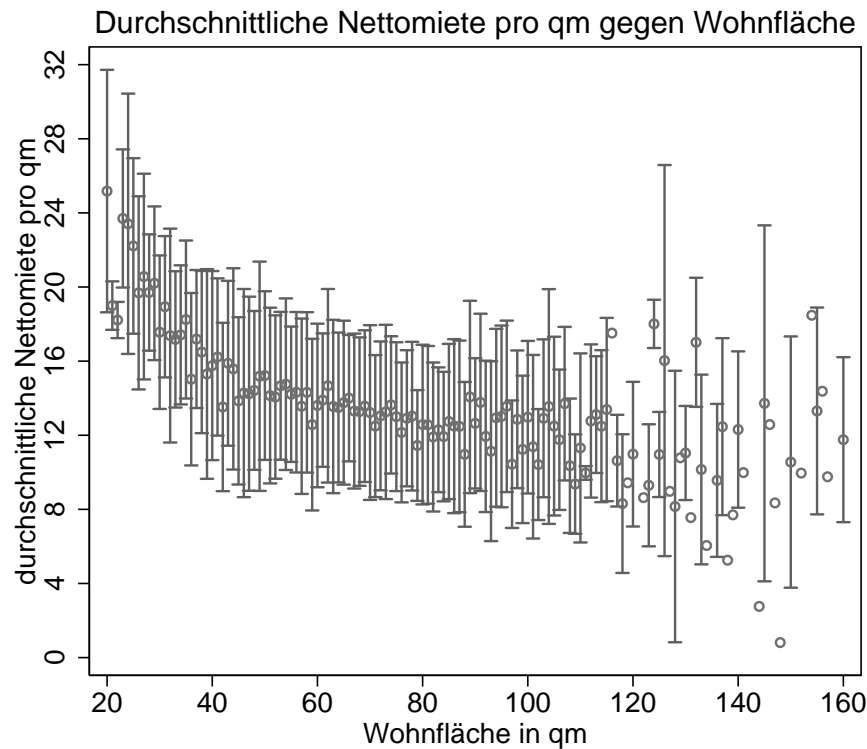
# Überblick

- Mietspiegel München: Geadditive Regression.
- Kartierung von Krankheitsrisiken: Räumliche Glättung.
- Unterernährung in Indien: Quantilregression.

# Mietspiegel München

- Erstellt zur Bestimmung der **ortsüblichen Vergleichsmiete**.
- Ältere Mietspiegel beruhten auf Gruppierung und Mittelwertbildung.
- Zunehmende Verwendung von Regressionsmodellen mit Nettomiete pro Quadratmeter als Zielvariable (volle Ausnutzung der vorhandenen Information).
- Typisches Vorgehen: **Variablenselektion und Modellwahl** zur Bestimmung der relevanten Kovariablen und der geeigneten Modellierungsform.

- Probleme:
  - Einbezug **nichtlinearer Effekte** beispielsweise der Wohnfläche oder des Baujahrs.



- Einbezug **räumlicher Information** über die Experteneinschätzung zur Lage hinaus.



- Ist es wirklich sinnvoll nur einen Teil der Kovariablen in der Prognose neuer Mieten zu verwenden?
- Mögliches Regressionsmodell für die Nettomiete pro qm  $nm$ :

$$nm = f_1(\text{wohnflaeche}) + f_2(\text{baujahr}) + f_3(\text{bezirksviertel}) + \mathbf{x}'\boldsymbol{\beta} + \varepsilon.$$

wobei  $\mathbf{x}'\boldsymbol{\beta}$  die Effekte eines potenziell **hochdimensionalen** Vektors von Kovariablen beinhaltet.

# Semiparametrische Regression

- In semiparametrischen Regressionsmodellen wird die parametrische Modellgleichung

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + x_{ip} \beta_p + \varepsilon_i$$

ersetzt durch

$$y_i = \beta_0 + f_1(z_i) + \dots + f_p(z_i) + \varepsilon_i$$

wobei  $f_1, \dots, f_p$  Funktionen verschiedenen Typs basierend auf generischen Kovariablen  $z$  bezeichnen.

- Die Funktionen werden typischerweise regularisiert geschätzt, indem ein Strafterm zu dem Schätzkriterium hinzugenommen wird.

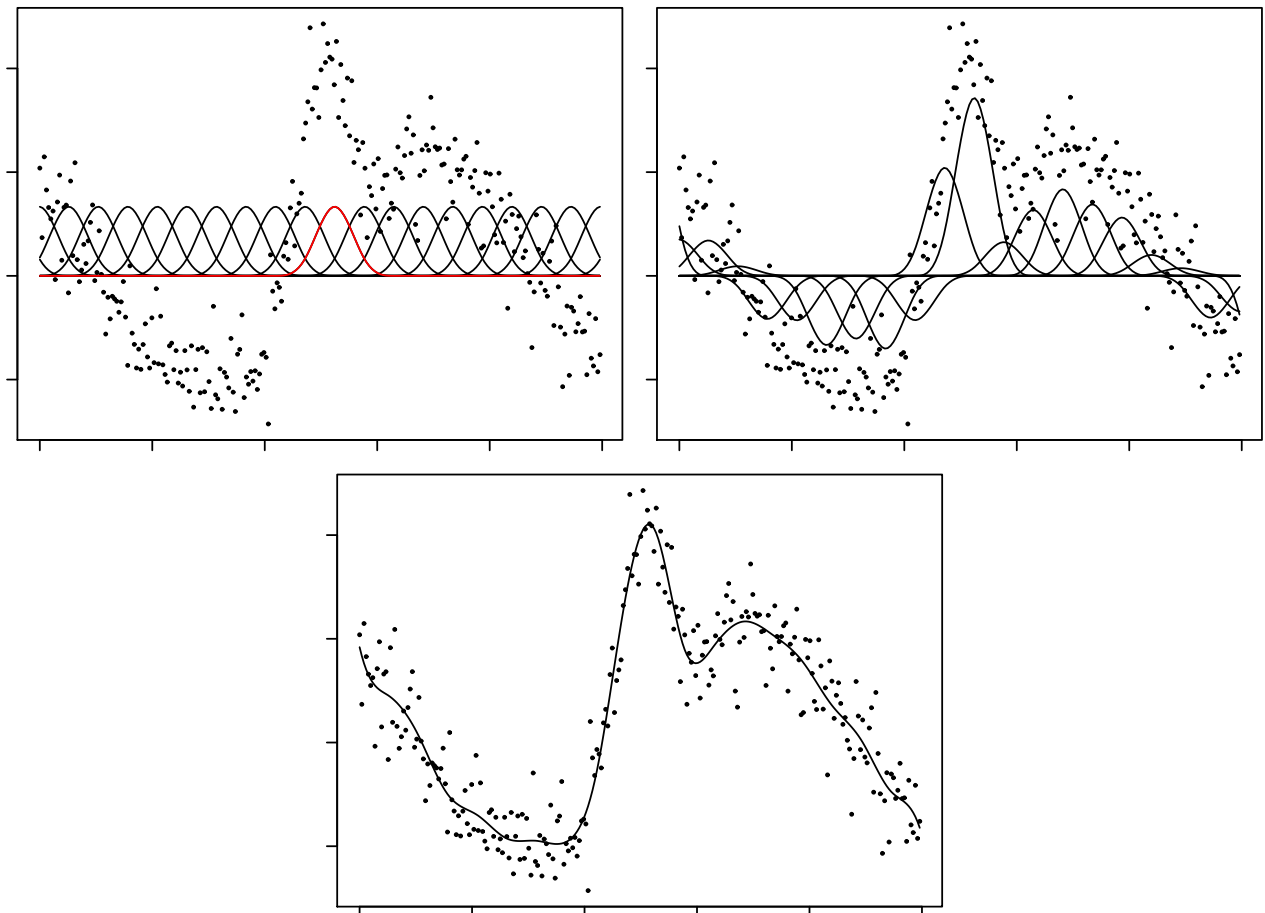
- Beispiele:
  - Lineare Effekte:  $f_j(\mathbf{z}) = \mathbf{x}'\boldsymbol{\beta}$ .
  - Nichtlineare, glatte Effekte metrischer Kovariablen:  $f_j(\mathbf{z}) = f(x)$
  - Interaktionsoberflächen  $f_j(\mathbf{z}) = f(x_1, x_2)$ .
  - Räumliche Effekte:  $f_j(\mathbf{z}) = f_{\text{spat}}(s)$ .
  - Zufällige Effekte:  $f_j(\mathbf{z}) = b_c$  mit Cluster-Index  $c$ .



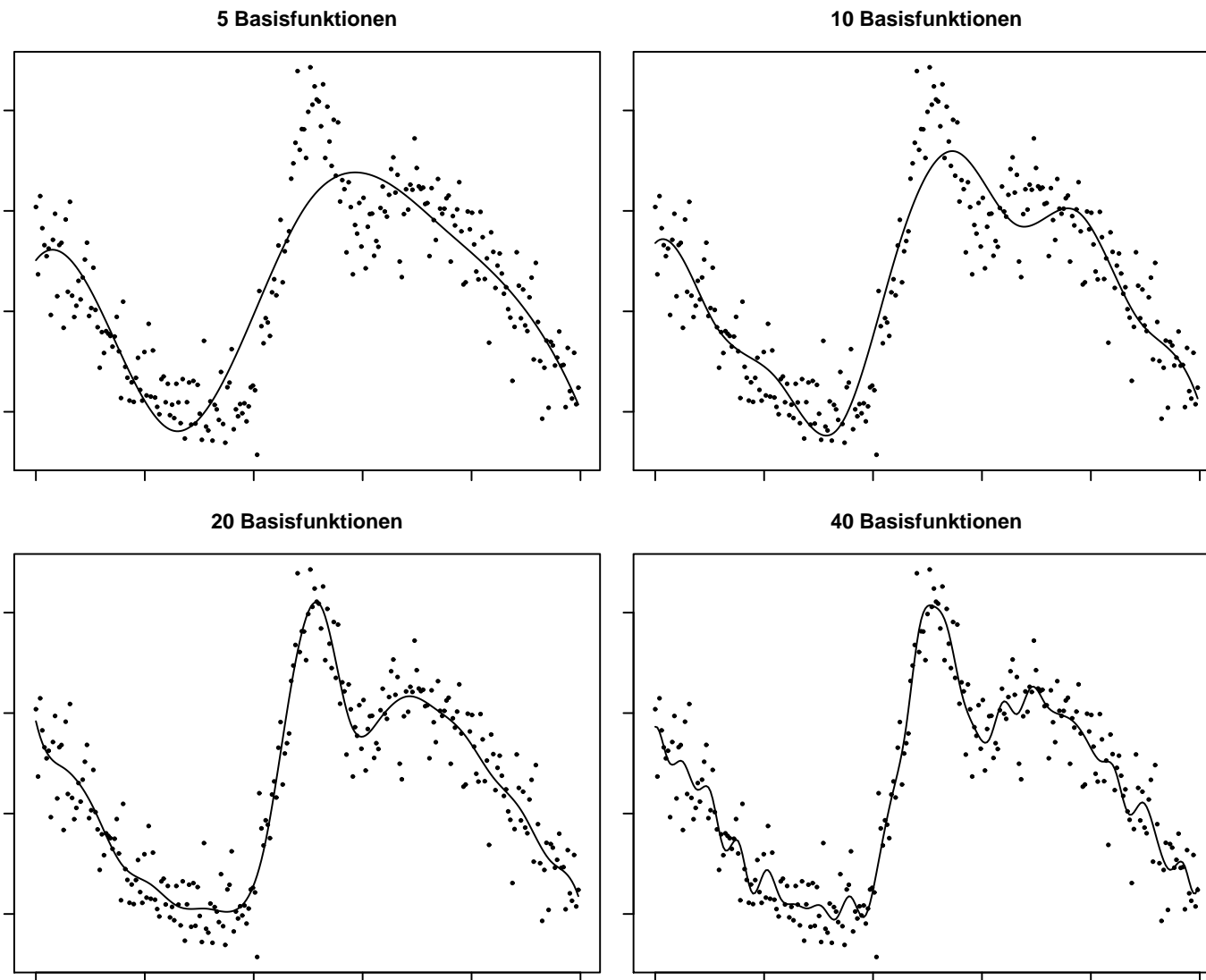
# Penalisierte Splines

- Approximiere eine nichtparametrisch zu schätzende Funktion  $f(x)$  durch eine Linearkombination von **B-Spline Basisfunktionen**  $B_j(x)$

$$f(x) = \sum_j \beta_j B_j(x)$$



- B-Spline Schätzungen für variierende Anzahlen von Basisfunktionen:



- Unregularisierte Schätzungen hängen stark von der Anzahl der Basisfunktionen ab.  
⇒ Ergänze das Schätzkriterium um einen **Regularisierung-Term** der raue Funktions-schätzungen bestraft.
- Beliebter Ansatz: Bestrafung der quadrierten zweiten Ableitung

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

- Einfache Approximation für B-Splines: **Differenzen-Strafterme**, z.B. für erste Differenzen

$$\begin{aligned} \text{pen}(\boldsymbol{\beta}) &= \lambda \sum_j (\beta_j - \beta_{j-1})^2 \\ &= \lambda \boldsymbol{\beta}' \mathbf{K} \boldsymbol{\beta} \end{aligned}$$

- Penalisiertes Kleinste Quadrate-Kriterium:

$$\begin{aligned}\text{PKQ}(\boldsymbol{\beta}) &= \sum_{i=1}^n \left( y_i - \sum_j \beta_j B_j(x) \right)^2 + \lambda \sum_j (\beta_j - \beta_{j-1})^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}' \mathbf{K} \boldsymbol{\beta}.\end{aligned}$$

- Der **Glättungsparameter**  $\lambda$  bestimmt den Einfluss der Regularisierung auf die Schätzung:
  - $\lambda \rightarrow \infty \Rightarrow$  Glatte Schätzung (konstant bzw. linear).
  - $\lambda \rightarrow 0 \Rightarrow$  Raue Schätzung.

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

## Räumliche Effekte für Regionendaten

- Ziel: Schätzung eines separaten Regressionsparameters  $\beta_s$  für jede Region.
- Instabile Schätzung bei im Verhältnis zum Stichprobenumfang großer Regionenzahl.  
⇒ Regularisierte Schätzung, um **räumlich glatte Effekte** zu erhalten.
- Räumlich glatt: Effekte benachbarter Regionen sollten ähnlich sein.
- Strafterm basierend auf **Differenzen benachbarter Regionen**:

$$\text{pen}(\beta) = \lambda \sum_s \sum_{r \in N(s)} (\beta_s - \beta_r)^2$$

wobei  $N(s)$  die Menge der Nachbarn der Region  $s$  bezeichnet.

- In stochastischer Formulierung ergibt sich ein **Markov Zufallsfeld**

$$\beta_s | \beta_r, r \in N(s) \sim N \left( \frac{1}{|N(s)|} \sum_{r \in N(s)} \beta_r, \frac{\tau^2}{|N(s)|} \right).$$

- Wieder erhält man einen quadratischen Strafterm

$$\text{pen}(\beta) = \lambda \beta' \mathbf{K} \beta$$

und damit ein penalisiertes Kleinste-Quadrate-Kriterium.

## Regularisierung in hochdimensionalen Modellen

- Regularisierung in Regressionsmodellen mit **vielen Kovariablen**: Bevorzuge sparsame Modelle in denen eine große Zahl von Koeffizienten nahe oder gleich Null ist.
- Beispiel Ridge-Regression:

$$\text{PKQ}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^p \beta_j^2 \rightarrow \min_{\boldsymbol{\beta}}.$$

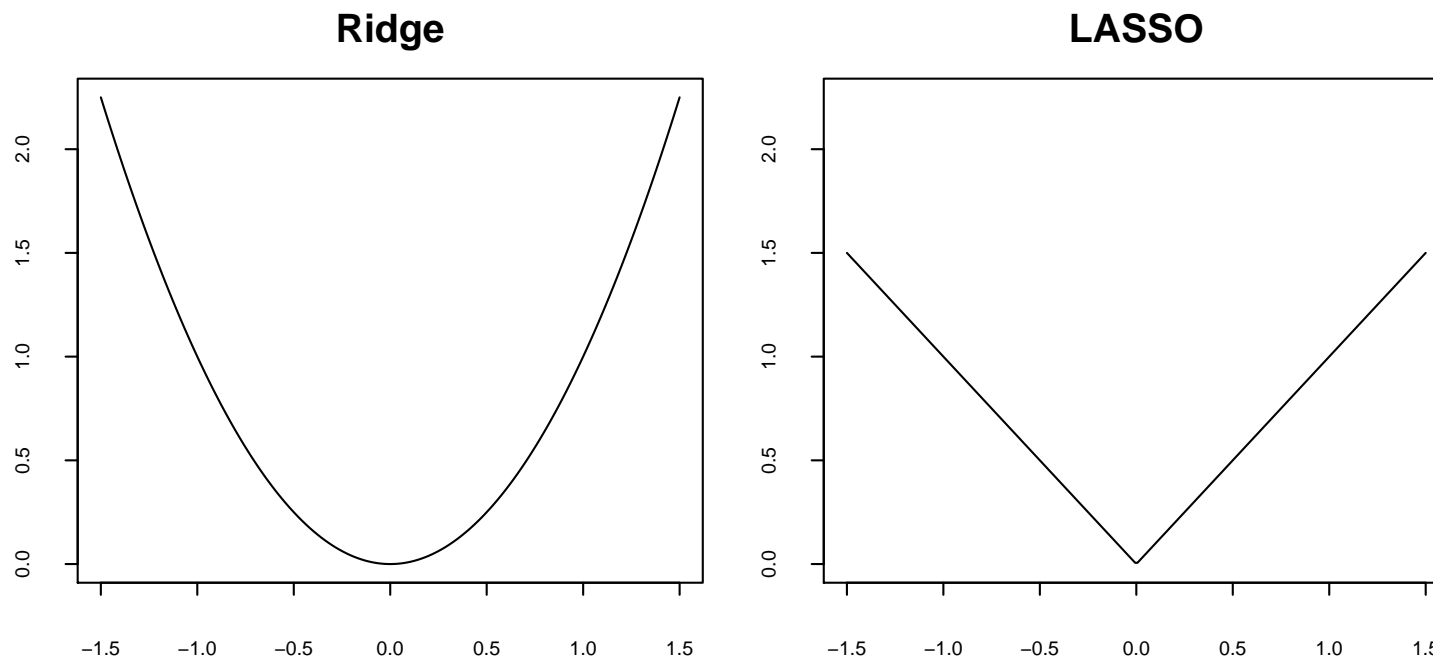
- Penalisierter KQ-Schätzer

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}.$$

- Der PKQ-Schätzer ist **verzerrt**, besitzt aber im Vergleich zum KQ-Schätzer eine **geringere Varianz**.
- Geeignete Glättungsparameter sollten zu einem **reduzierten MSE** führen.

- Nachteil der Ridge-Regression: Die resultierenden Modelle sind nicht sparsam genug.  
⇒ Betrachte Strafterme mit Peak in der Null.
- Beispiel LASSO: Ersetze den quadratischen Strafterm durch den **Absolutbetrag**:

$$\text{PKQ}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^p |\beta_j| \rightarrow \min_{\boldsymbol{\beta}} .$$





# Inferenz

- Allgemeine Struktur semiparametrischer Regressionsmodelle:

$$y_i = \beta_0 + f_1(\mathbf{z}_i) + \dots + f_p(\mathbf{z}_i) + \varepsilon_i$$

bzw.

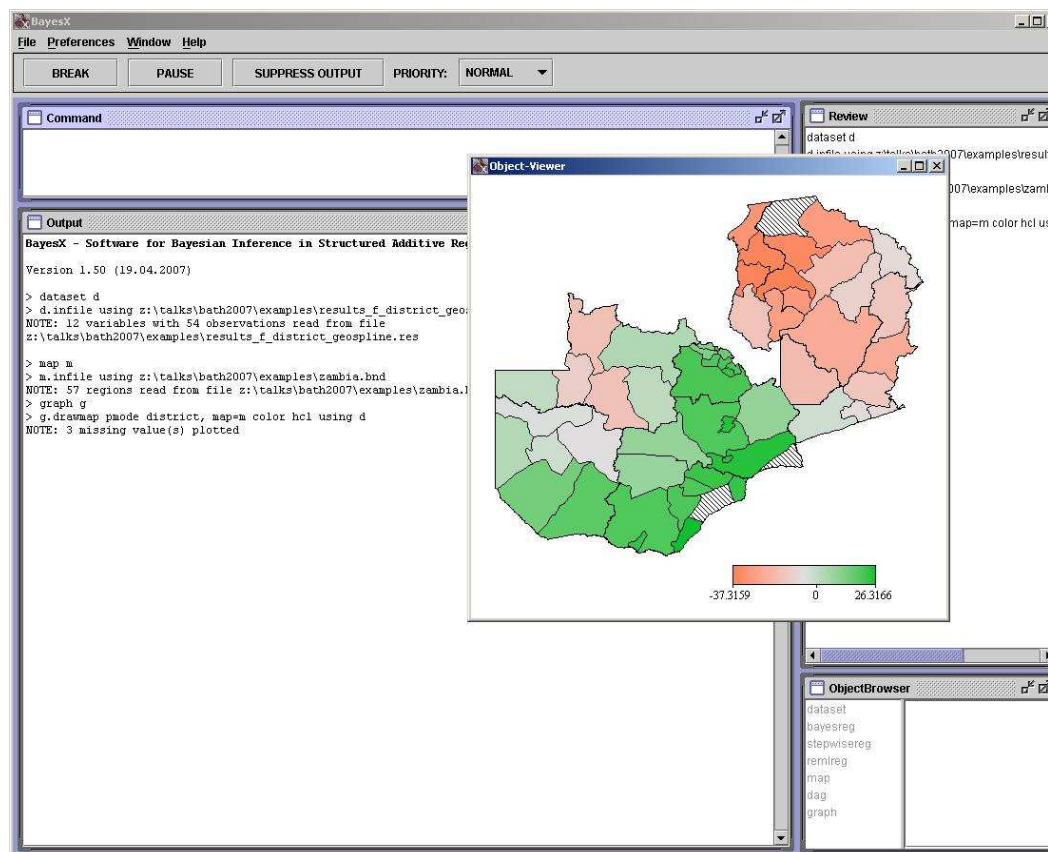
$$\mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \dots + \mathbf{X}_p\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}.$$

- **Penalisierte Kleinste Quadrate-Schätzung** der Regressionskoeffizienten.

- **Datengesteuerte Wahl der Glättungsparameter** zum Beispiel über
  - Bayesianische Ansätze und Markov Chain Monte Carlo Simulations-Verfahren,
  - die Darstellung semiparametrischer Regressionsmodelle als Modelle mit zufälligen Effekten,
  - Boosting-Verfahren.
- Die Methodik ist nicht nur für normalverteilte Zielgrößen sondern **allgemeiner anwendbar** (penalisierte Maximum Likelihood-Schätzung, penalisierte Optimalitätskriterien).

# BayesX

- Die beschriebene Methodik ist implementiert im freien Software-Paket BayesX.



- Erhältlich unter

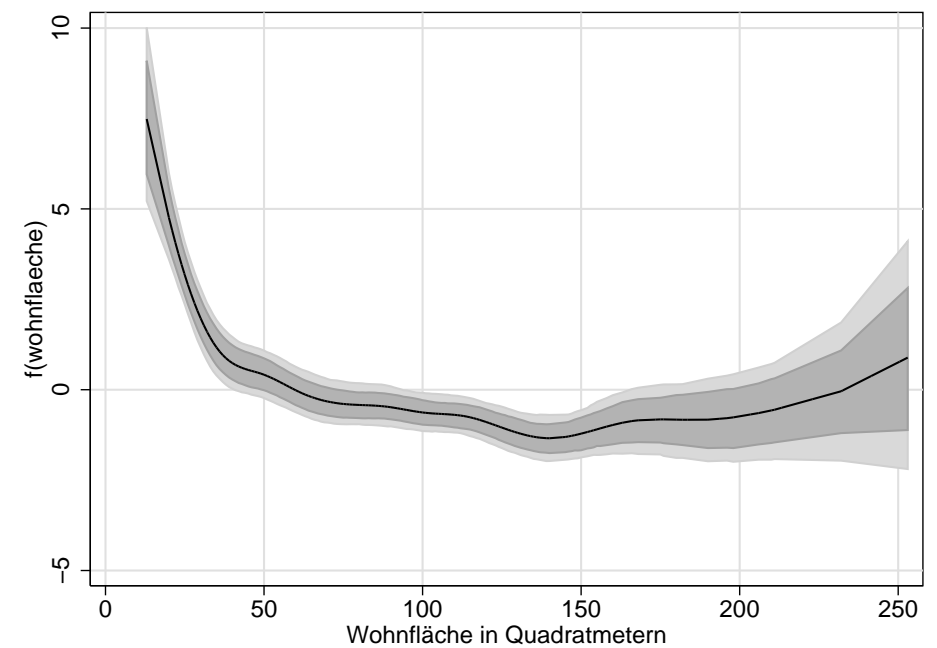
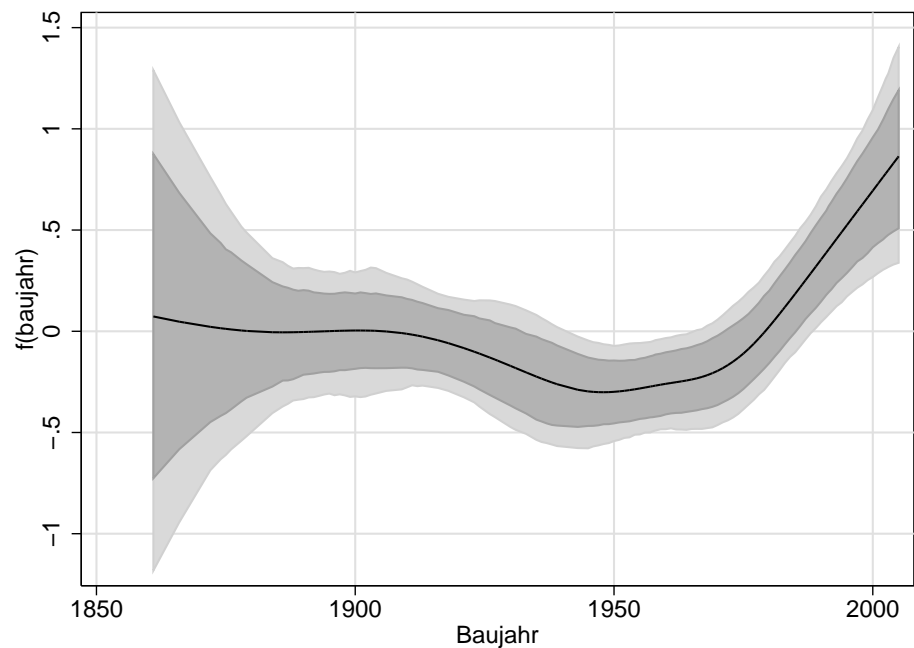
<http://www.stat.uni-muenchen.de/~bayesx>

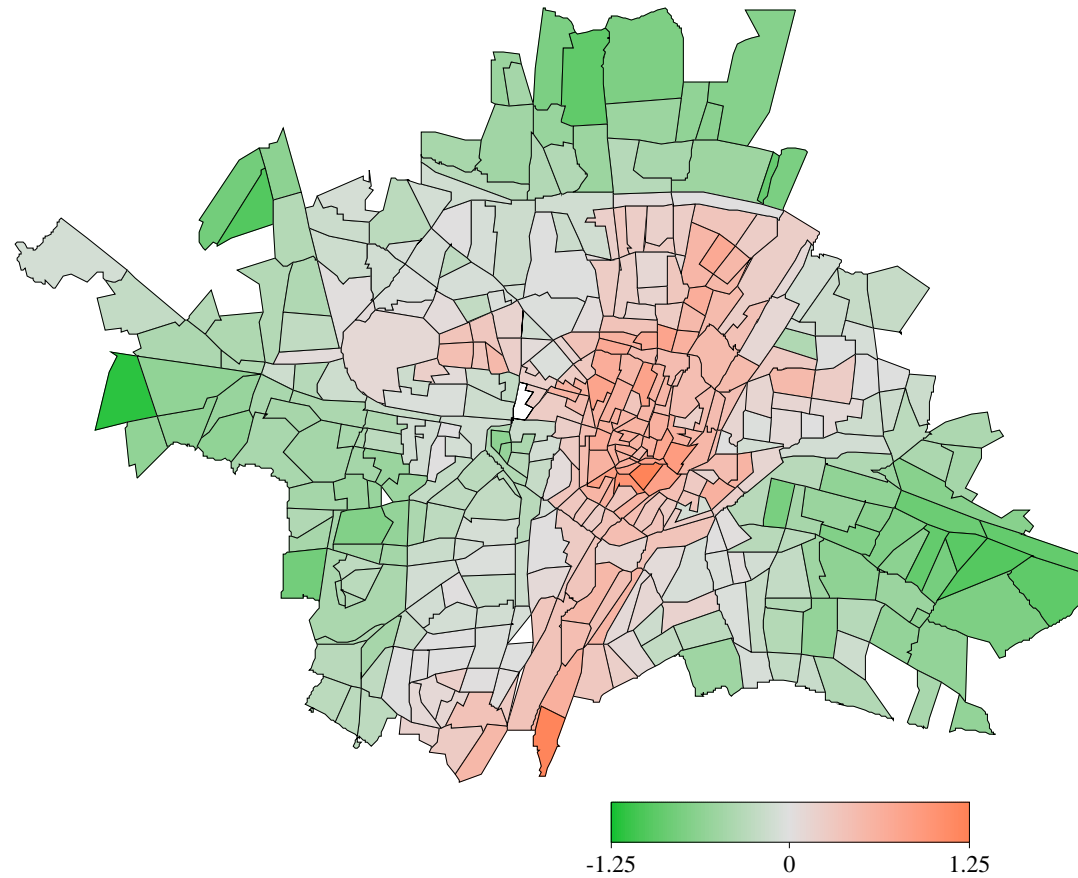
# Münchener Mietspiegel: Ergebnisse

- Modellgleichung

$$nm = f_1(\text{wohnflaeche}) + f_2(\text{baujahr}) + f_3(\text{bezirksviertel}) + \mathbf{x}'\boldsymbol{\beta} + \varepsilon.$$

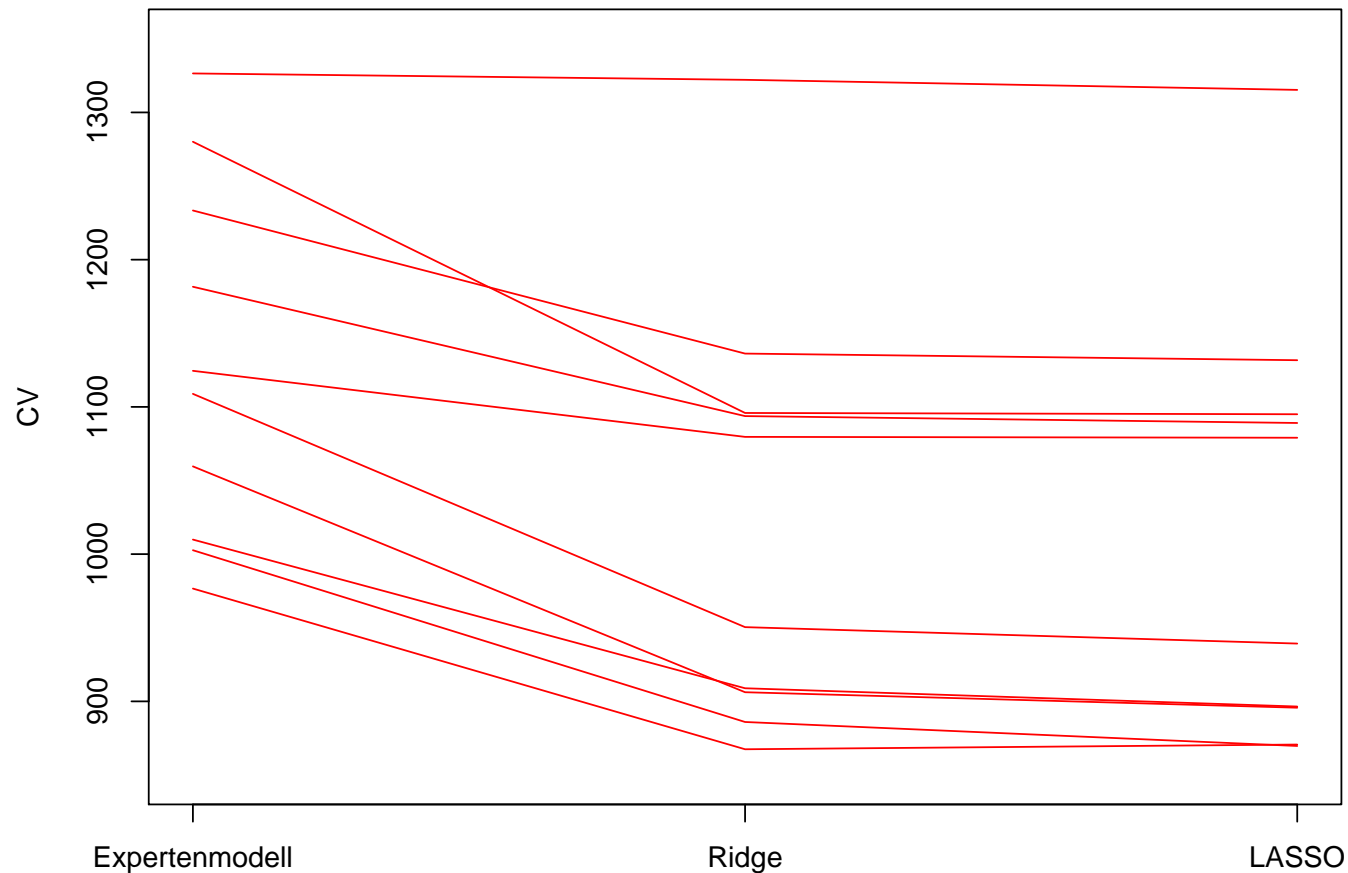
- Im Folgenden Ergebnisse bei LASSO-Regularisierung des Vektors  $\boldsymbol{\beta}$ .





- Interpretierbare Ergebnisse, aber was gewinnt man für die Prognose?

- Vergleich eines Expertenmodells (Subvektor + Transformation der Kovariablen), der Ridge-Regression und der LASSO-Regression über 10-fache Kreuzvalidierung.



⇒ Deutlich verbesserte Vorhersageeigenschaften durch Regularisierung!

## Räumliche Kartierung von Krankheitsrisiken

- Analyse der **geografischen Variation** des Erkrankungs- oder Mortalitätsrisikos bezüglich ein Krankheit.
- Ziel: **Identifikation unbekannter Risikofaktoren** / Beurteilung von **Kovariableneffekten**.
- Interessierende Zielvariable: Anzahl der Todesfälle  $y_s$  aufgrund einer Krankheit in Regionen  $s = 1, \dots, S$  (z.B. Bundesländern, Landkreisen, Gemeinden, etc.).

- Verteilungsannahme:

$$y_s \sim P(e_s \lambda_s)$$

wobei

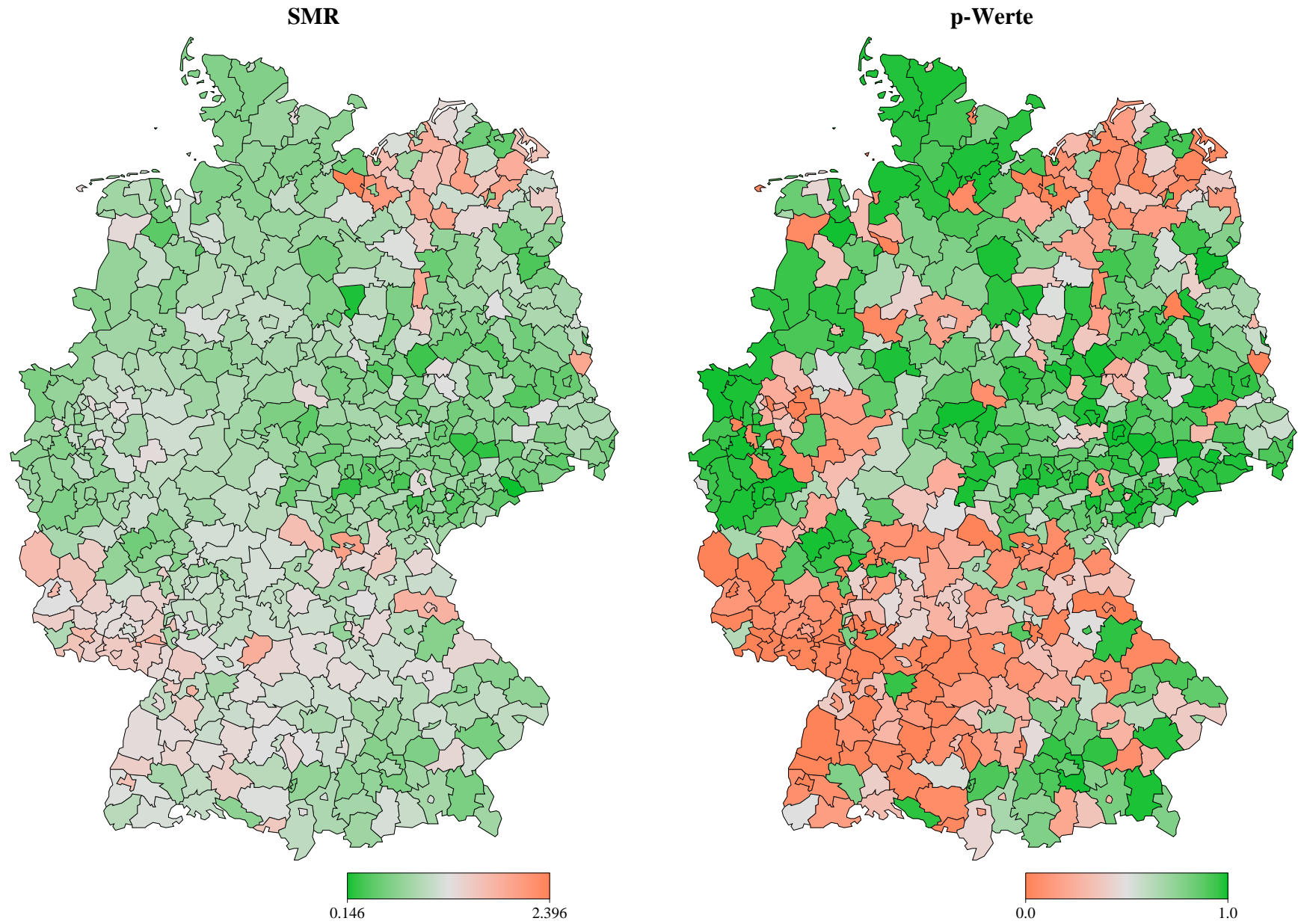
- $e_s$  der erwarteten Anzahl von Todesfällen entspricht (z.B. berechnet basierend auf Bevölkerungsgröße und Altersverteilung) und
- $\lambda_s$  das **relative Risiko** bezeichnet.

- Beispiel: Mortalität bezüglich Mundhöhlenkrebs in Deutschland zwischen 1985 und 1990.
- Beobachtete und erwartete Fallzahlen liegen auf Kreisebene vor.
- Zur deskriptiven Analyse nimmt man zunächst an, dass die Fallzahlen  $y_s$  unabhängig sind, also keine räumlichen Korrelationen vorliegen.
- Dann erhält man als Maximum-Likelihood-Schätzer für  $\lambda_s$  die **Standard-Mortalitätsraten**

$$\hat{\lambda}_s = \frac{y_s}{e_s}.$$

- Alternativ können **p-Werte** zum Test auf  $\lambda_s > 1$  bestimmt werden.





- Probleme der deskriptiven Ansätze:
  - **Räumliche Korrelationen** werden nicht berücksichtigt. Die Standardfehler der geschätzten Regressionskoeffizienten stimmen nicht (in der Regel unterschätzt).
  - Die Reliabilität der Aussagen hängt wesentlich von der Zahl erwarteter Todesfälle ab. Beispiel: Die Standardabweichung der Standardmortalitätsraten ist gegeben durch

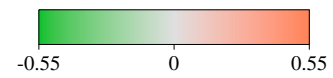
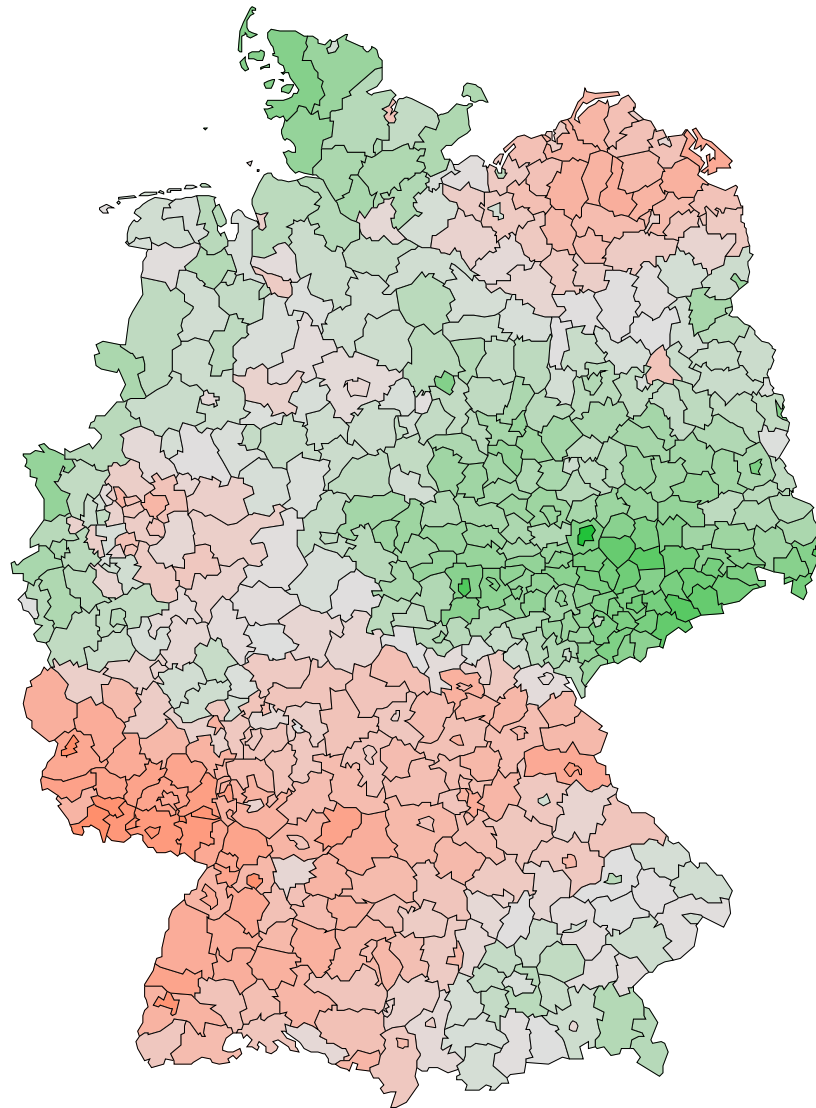
$$\text{sd}(\hat{\lambda}_{ML}) = \frac{\sqrt{y_s}}{e_s}.$$

- Keine Berücksichtigung von **Kovariablen**.

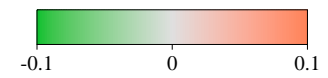
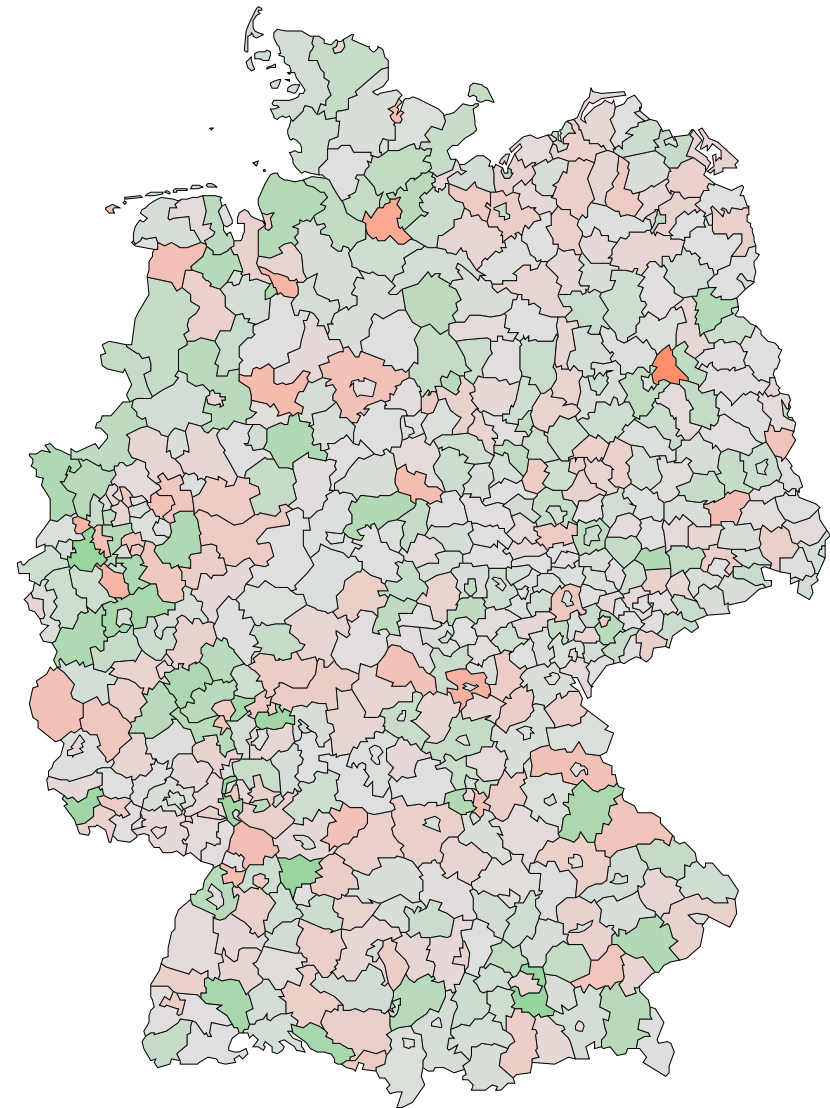
⇒ **Regressionsmodelle mit räumlichen Effekten.**

- Für eine detailliertere Analyse trennen wir das Risiko in einen räumlich korrelierten und einen räumlich unkorrelierten Effekt.

**Strukturiert**



**Unstrukturiert**



## Quantilregression: Unterernährung in Indien

- Übliche Regressionsmodelle beschreiben den Erwartungswert einer Zielvariablen.
- In manchen Anwendungen sind aber Randbereiche der Verteilung (“extreme Ereignisse”) von Interesse.
- Beispiel: Unterernährung von Kindern in Entwicklungs- und Schwellenländern.
- Der Ernährungszustand wird durch Z-Scores beurteilt:

$$Z_i = \frac{I_i - \mu}{\sigma}$$

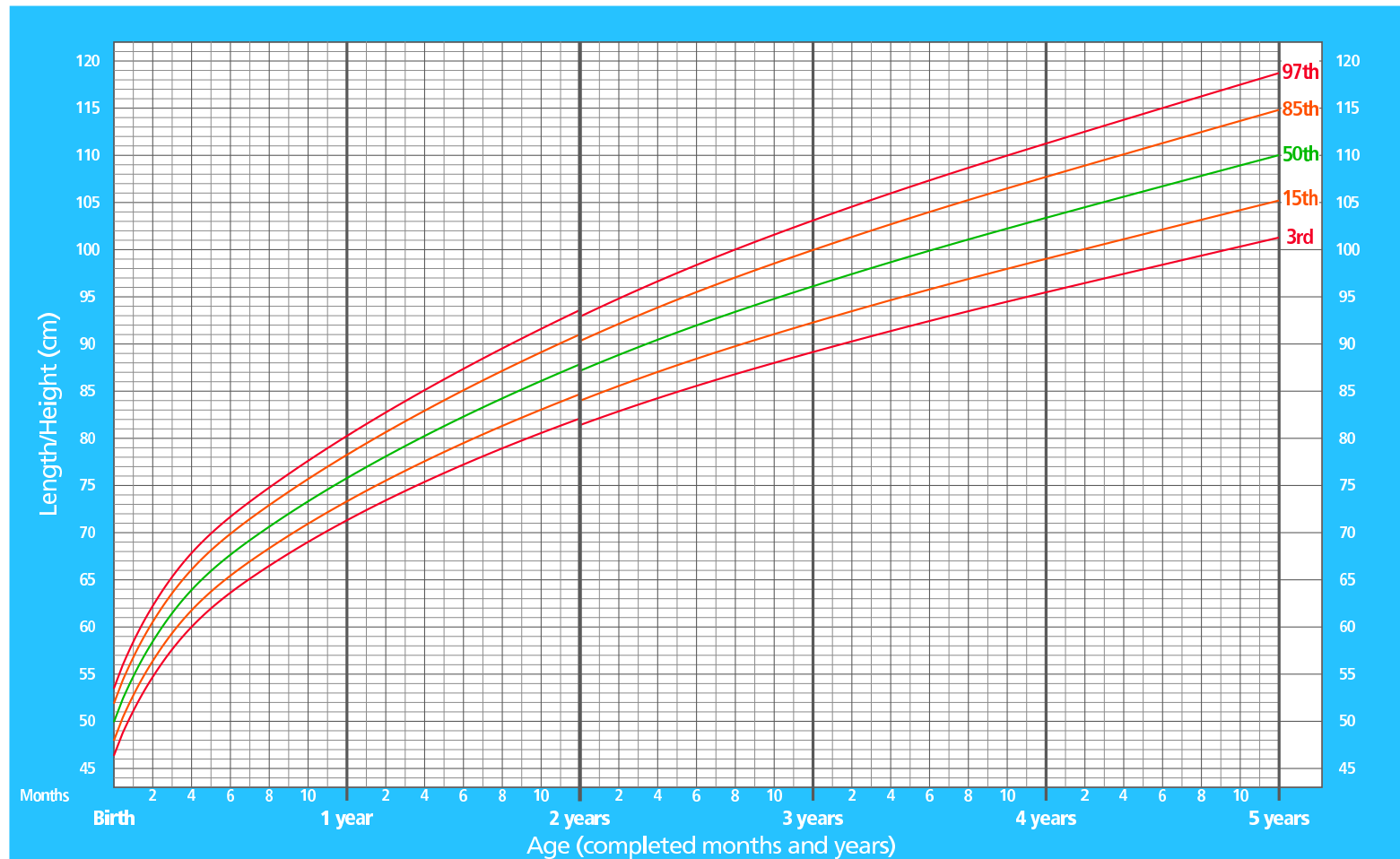
wobei  $I$  einen Indikator wie Gewicht oder Größe bezeichnet und  $\mu$  und  $\sigma$  Median und Standardabweichung aus einer Referenzpopulation.

- Chronische Unterernährung (Stunting) wird über die Körpergröße relativ zum Alter gemessen.

- Kinder werden basierend auf Referenzcharts der WHO als unterernährt eingestuft:

## Length/height-for-age BOYS

Birth to 5 years (percentiles)



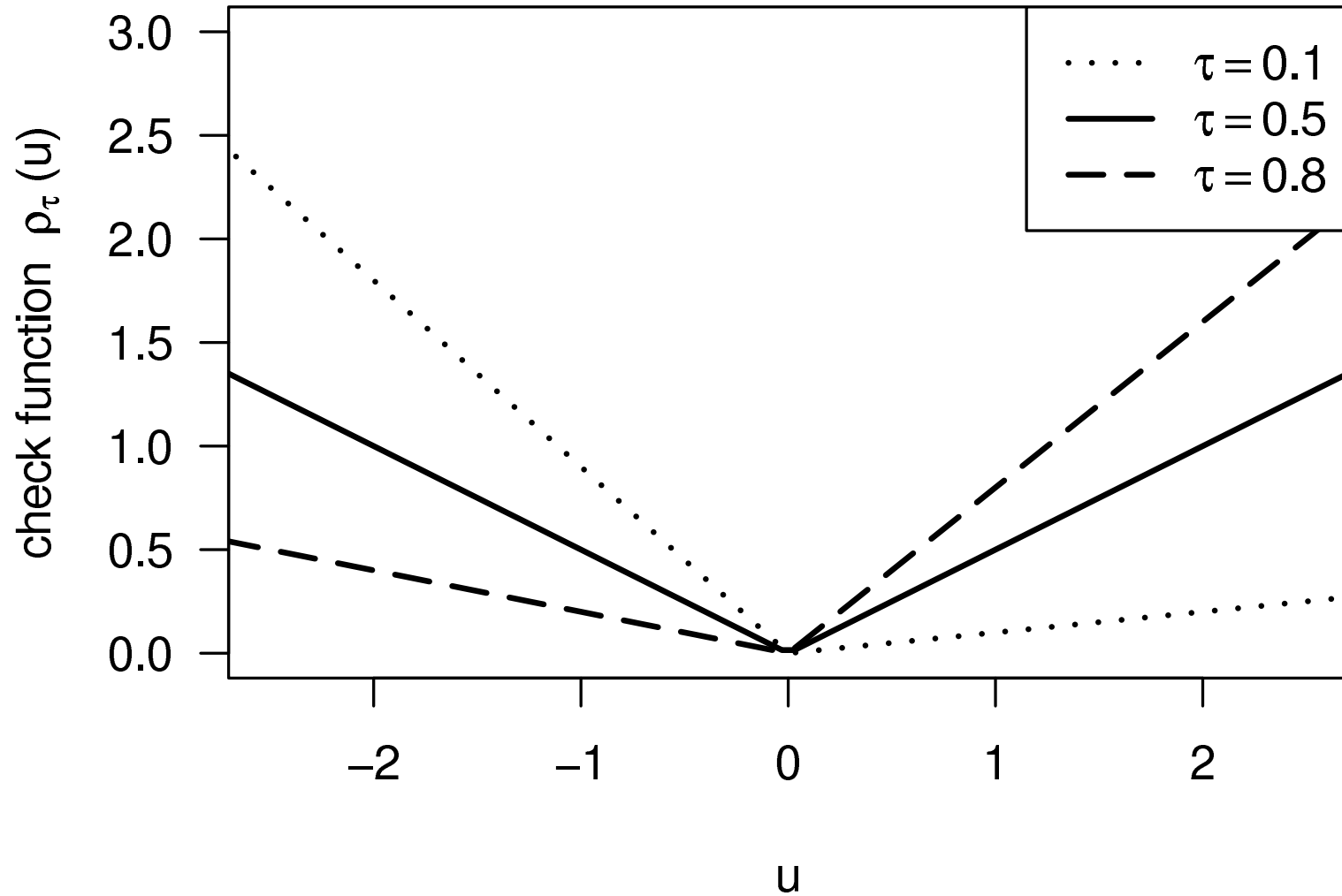
WHO Child Growth Standards

- Im Folgenden Analyse des 2005/06 **India Demographic and Health Survey** (<http://www.measuredhs.com>).
- Repräsentative Erhebung zu Familienplanung, Gesundheit, Ernährungszustand, Demographie, etc.
- Modelle der Quantilregression erklären bedingte Quantile statt des bedingten Erwartungswerts
- Formulierung über eine Verlustfunktion:

$$\hat{\beta}_\tau = \operatorname{argmin}_{\beta_\tau} \sum_{i=1}^n \rho_\tau(Z_i - \mathbf{x}'_i \beta_\tau)$$

wobei

$$\rho_\tau(u) = \begin{cases} u \tau & u \geq 0 \\ u(\tau - 1) & u < 0 \end{cases}$$



- Äquivalente Formulierung als **Regressionsproblem**:

$$Z_i = \mathbf{x}_i' \boldsymbol{\beta}_\tau + \varepsilon_{\tau i}$$

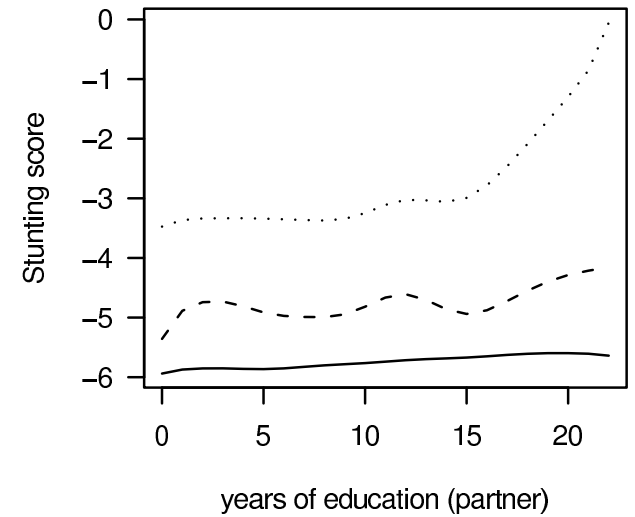
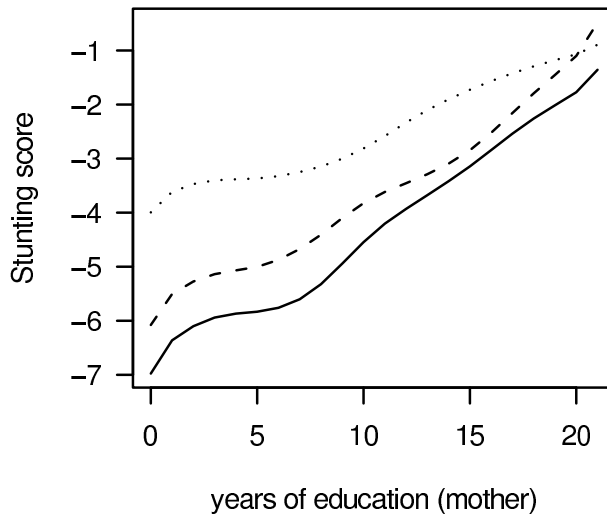
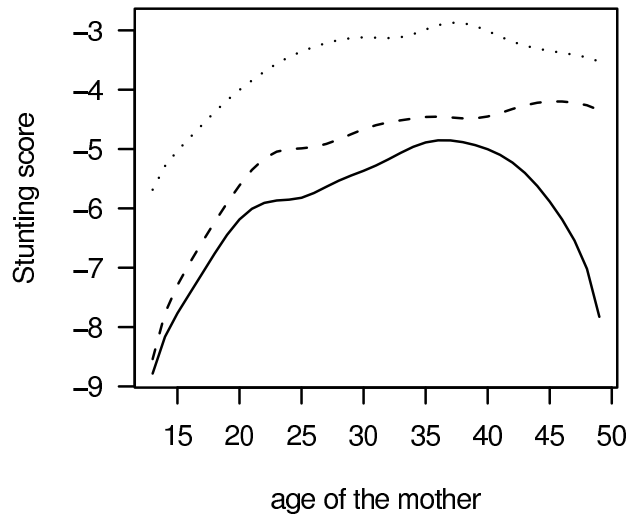
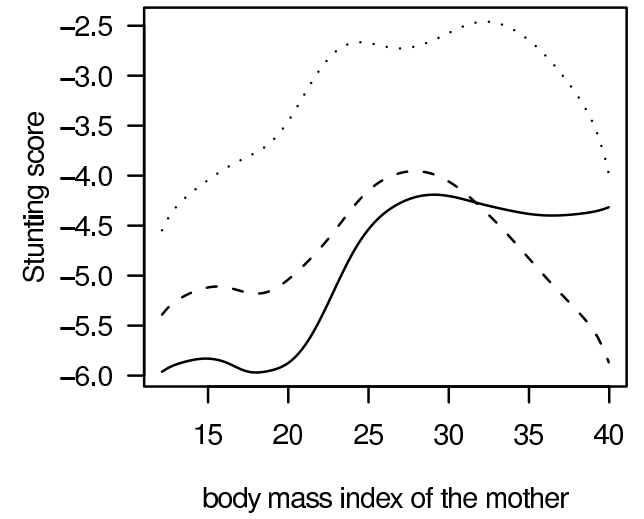
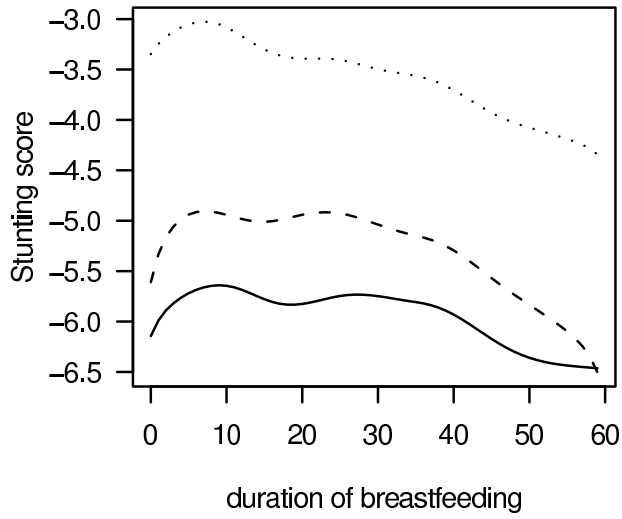
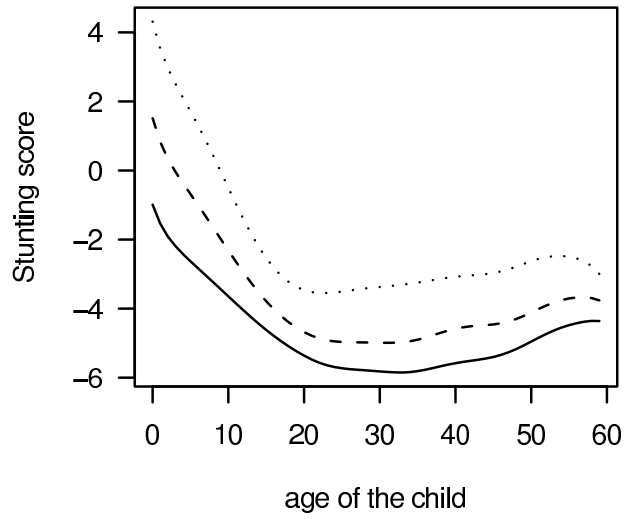
wobei  $\varepsilon_{\tau i}$  unabhängige Fehlerterme mit

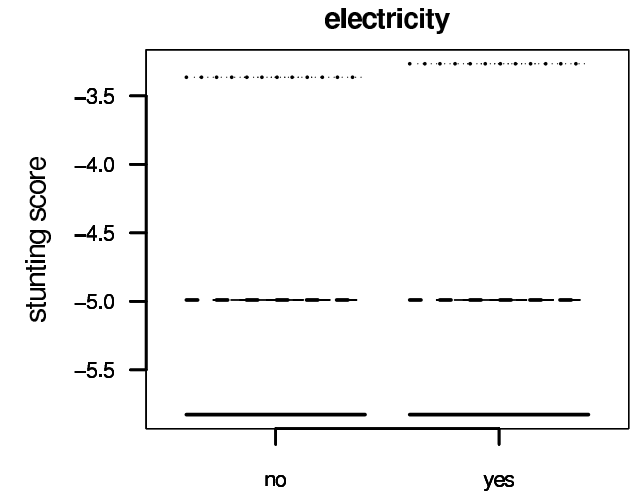
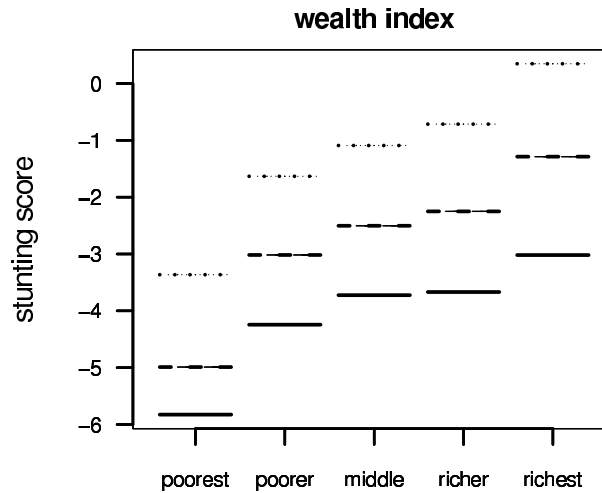
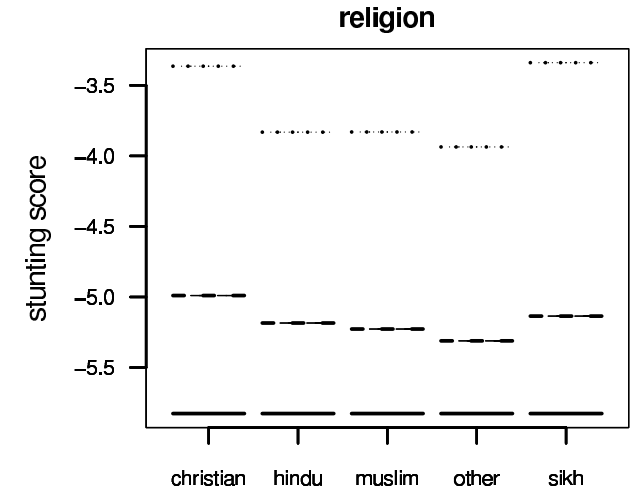
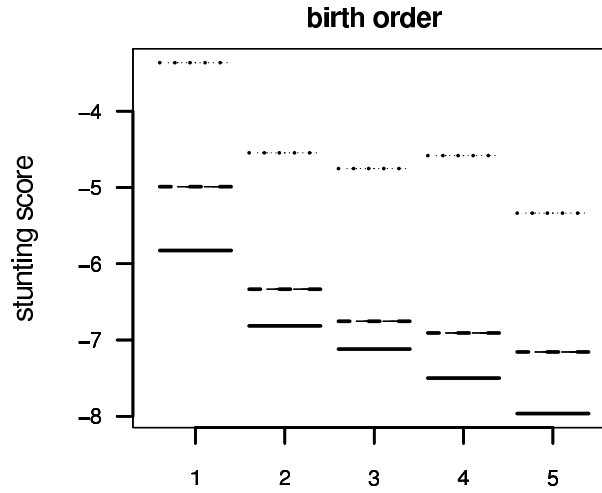
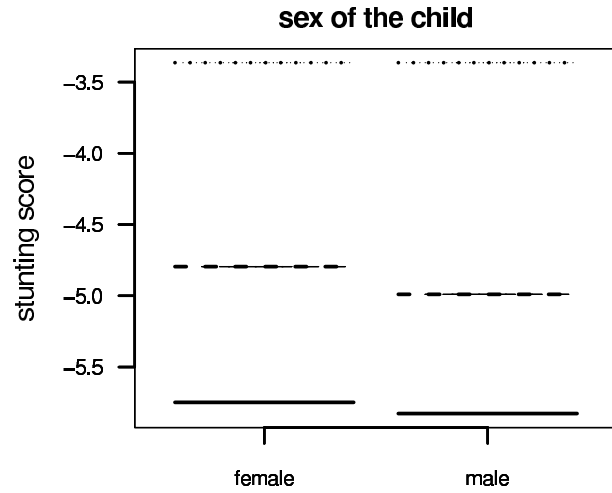
$$F_{\varepsilon_{\tau i}}(\tau) = 0$$

bezeichnet und  $F_{\varepsilon_{\tau i}}(\tau)$  die Verteilungsfunktion des Fehlers ist.

- Entscheidend:
  - **Keine explizite Verteilungsannahme** für die Fehler.
  - Die Fehler müssen **nicht die gleiche Verteilung besitzen**.
- ⇒ Quantilregression erlaubt Heteroskedastizität.
- Auch hier ist die Erweiterung auf semiparametrische Modelle möglich.







# Zusammenfassung

- Semiparametrische Regressionsmodelle erlauben die flexible Modellierung statistischer Zusammenhänge.
- Insbesondere nichtlineare und räumliche Effekte.
- Datengesteuerte Schätzung auch der Glattheitseigenschaften.
- A place called home:

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