Geoadditive Hazard Regression bei intervallzensierten Überlebenszeiten

Thomas Kneib
Teilprojekte **A5**, **B2**, **C1**

1. Kindersterblichkeit in Nigeria
2. Intervallzensierte Überlebenszeiten
3. Strukturierte Hazard Regression
4. Inferenz basierend auf gemischten Modellen
5. Software
6. Kindersterblichkeit in Nigeria II
7. Simulationsergebnisse

SFB 386

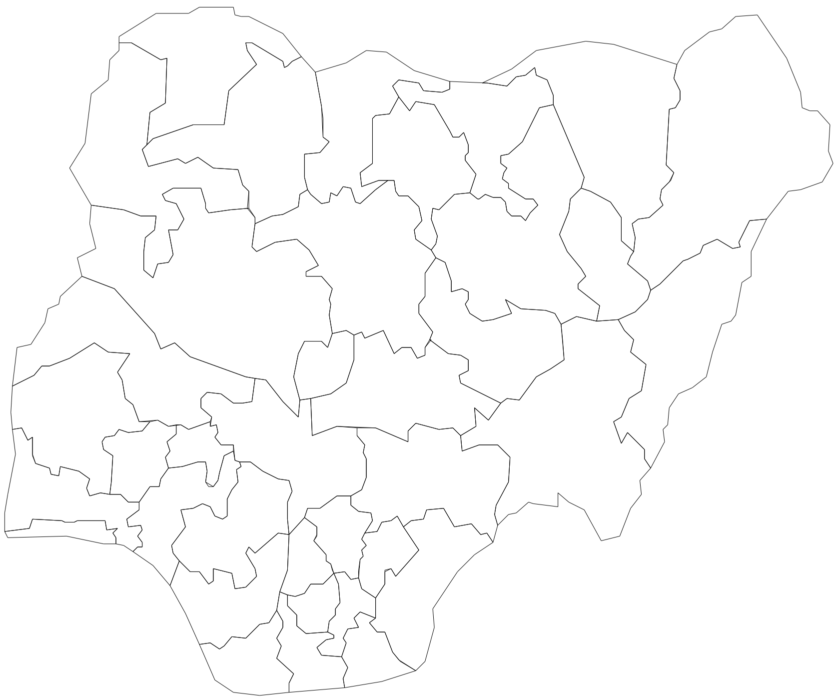
1.7.2005

Institut für Statistik münchen

Kindersterblichkeit in Nigeria

- Daten aus dem 2003 Demographic and Health Survey (DHS) in Nigeria.
- Retrospektive Befragung zum Gesundheitszustand von Frauen im gebärfähigen Alter und deren Kindern.
- Überlebenszeiten zu $n = 5323$ Kindern.
- Zahlreiche Kovariablen inklusive räumlicher Information.
- Analyse basierend auf dem Cox-Modell:

$$\lambda(t; u) = \lambda_0(t) \exp(u'\gamma).$$

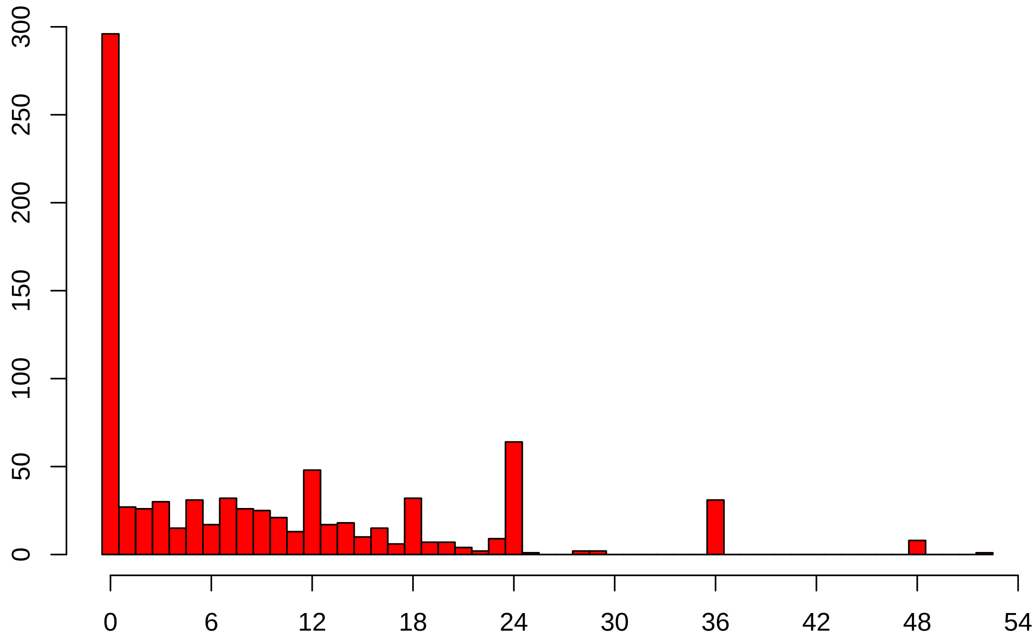


- Einschränkungen des klassischen Cox-Modells:
  - Beschränkt auf rechtszensierte Beobachtungen.
  - Nachträgliche Schätzung der Baseline Hazardrate.
  - Annahme proportionaler Hazardraten.
  - Parametrische Form des Prädiktors.
  - Keine räumlichen Korrelationen.
- Erweiterungen in Bezug auf einzelne Probleme existieren, aber es fehlt eine Kombination aller Probleme.

Intervallzensierte Überlebenszeiten

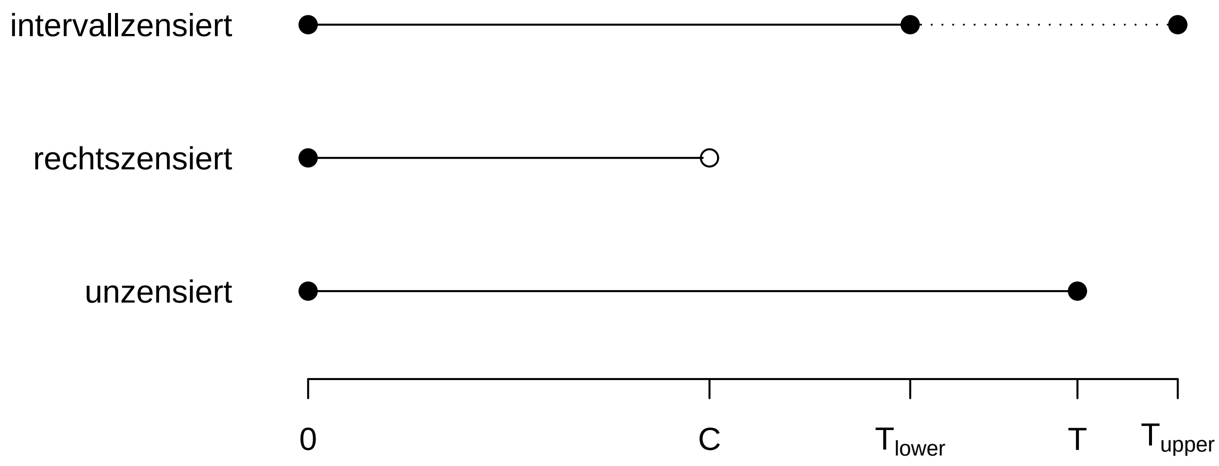
- Theoretisch wird die Überlebenszeit in Tagen erhoben.

- Retrospektives Design $\Rightarrow$ die meisten unzensierten Überlebenszeiten sind gerundet.



- Zensierte Überlebenszeiten sind dagegen exakt erhältlich.

$\Rightarrow$ Betrachte Überlebenszeiten als intervallzensiert.



- Likelihood-Beiträge:

$$P(T \in [T_{lower}, T_{upper}]) = S(T_{lower}) - S(T_{upper})$$

- Können in Abhängigkeit der Hazardrate geschrieben werden:

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right).$$

- Ableitungen der Log-Likelihood werden wesentlich komplizierter.
- Numerische Integrationsverfahren müssen verwendet werden.
- Stückweise konstante, zeitvariierende Kovariablen und Linkstrunkierung können einfach berücksichtigt werden.

Strukturierte Hazard Regression

- Flexibles, semiparametrisches Modell für die Hazardrate:

$$\lambda(t;\cdot) = \exp\left[g_0(t) + \sum_{j=1}^{q} g_j(t) z_j + \sum_{k=1}^{p} f_k(x_k) + f_{spat}(s) + u(t)'\gamma\right]$$

wobei

$g_0(t) = \log(\lambda_0(t))$	logarithmierte Baseline-Hazardrate,
g_j	zeitvariierende Effekte der Kovariablen z_j,
f_k	nonparametrische Funktionen der stetigen Kovariablen x_k,
f_{spat}	räumliche Funktion,
$u(t)$	zeitvariierende Kovariablen.

- P-Splines:
  - Approximiere $f_k(x_k)$ bzw. $g_j(t)$ durch große Zahl von B-Spline-Basisfunktionen.
  - Bestrafe Differenzen benachbarter Regressionskoeffizienten.
- Markov Zufallsfelder:
  - Geeignet für Regionendaten.
  - Definiere Nachbarschaften für die Regionen.
  - Erwartungswert von $f_{spat}(s)$ ist der Mittelwert der Funktionsauswertungen benachbarter Regionen.
- Stationäre Gaussfelder (Kriging):
  - Geeignet für exakte Lokationen.
  - Der räumliche Effekt folgt einem stationären stochastischen Prozess.
  - Die Korrelation zweier Punkte wird durch eine intrinsische Korrelationsfunktion beschrieben.

- Erweiterungen:
  - Interaktionsoberflächen (2D P-splines).
  - Variierende Koeffizienten (stetige und räumliche Effektmodifizierer).
  - Frailties (i.i.d. random effects).
- Alle Effekte können in einem einheitlichen Rahmen geschätzt werden.

Inferenz basierend auf gemischten Modellen

- Jeder Term im Prädiktor wird beschrieben durch einen Vektor von Regressionskoeffizienten mit uneigentlicher multivariater Normalverteilungs-Priori:

$$p(\beta_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right)$$

- K_j ist eine Strafmatrix, τ_j^2 ein Glättungsparameter.

$\Rightarrow$ Reparametrisierung des Modells in ein gemischtes Modell mit normierbaren Verteilungen.

- Empirische Bayes-Schätzer erhält man durch Iteration der folgenden Schritte:
  - Penalisierte Maximum Likelihood-Schätzung der Regressionskoeffizienten.
  - Restricted Maximum Likelihood- / Marginale Likelihood-Schätzung der Varianzparameter.

Software

- Implementiert im Softwarepaket BayesX.

- BayesX erlaubt
  - beliebige Kombinationen von Links-, Rechts-, und Intervallzensierung sowie Linkstrunkierung,
  - die gleichzeitige Schätzung der Baseline-Hazardrate und der Kovariableneffekte,
  - den Einbezug zeitvariierender Effekte und zeitvariierender Kovariablen,
  - nonparametrische Effekte, räumliche Effekte, Interaktionsoberflächen und Frailties.

- Erhältlich unter

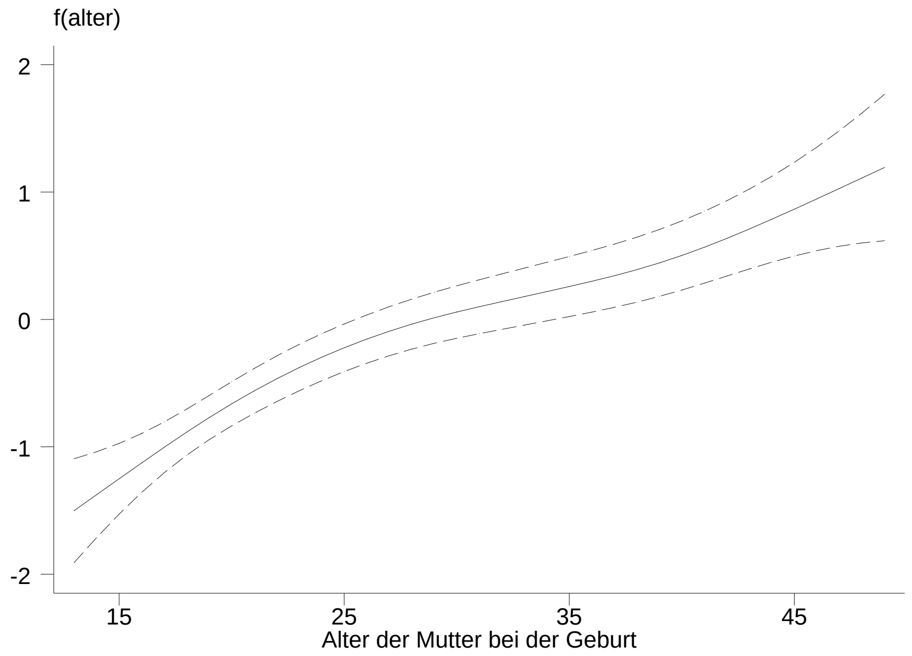
`http://www.stat.uni-muenchen.de/~bayesx`

Kindersterblichkeit in Nigeria II

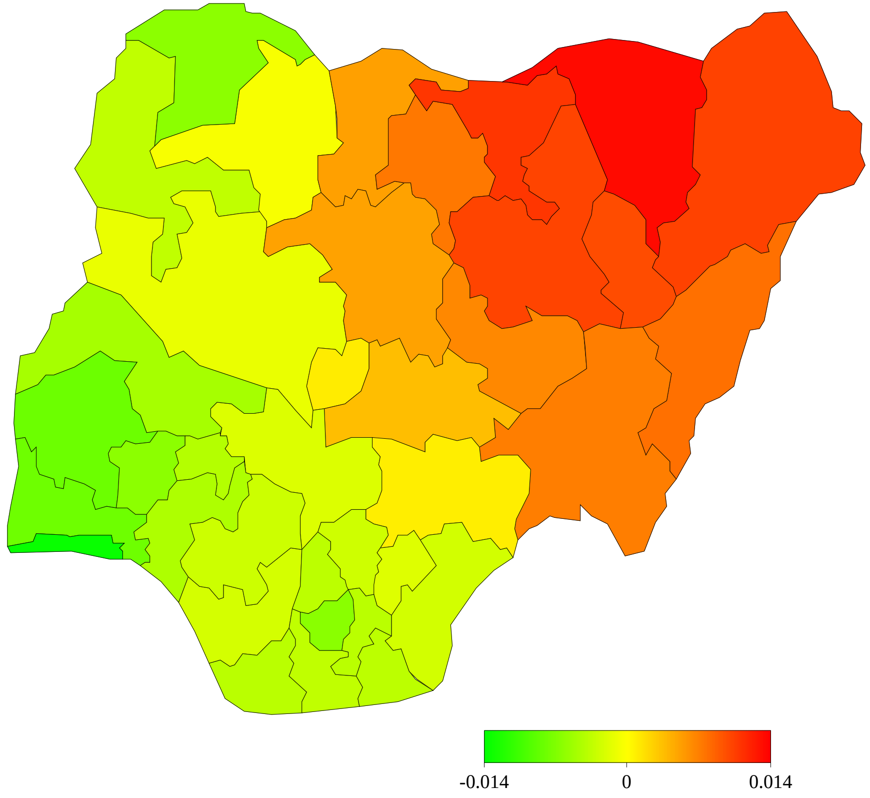
- Modell:

$$\lambda(t) = \exp(\eta(t))$$

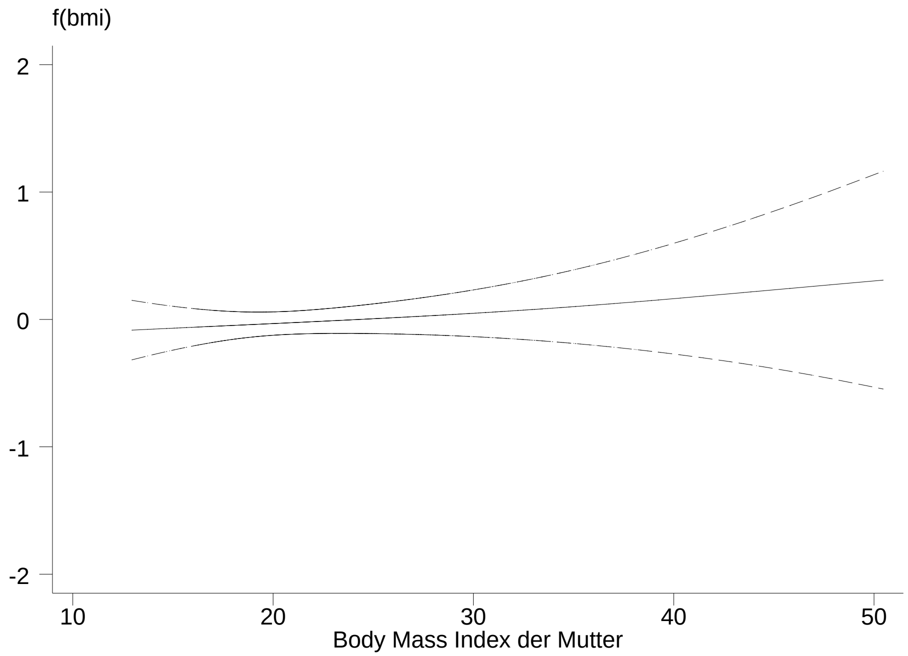
$$\begin{aligned}\eta(t) &= g_0(t) \\ &\quad +f_{spat}(distrikt) \\ &\quad +f_1(bmi) + f_2(alter) + f_3(geburtsordnung) + f_4(haushaltsgroesse) \\ &\quad +\beta_1 stillen(t) \\ &\quad +\beta_2 geschlecht + \ldots\end{aligned}$$



Alter der Mutter bei der Geburt.

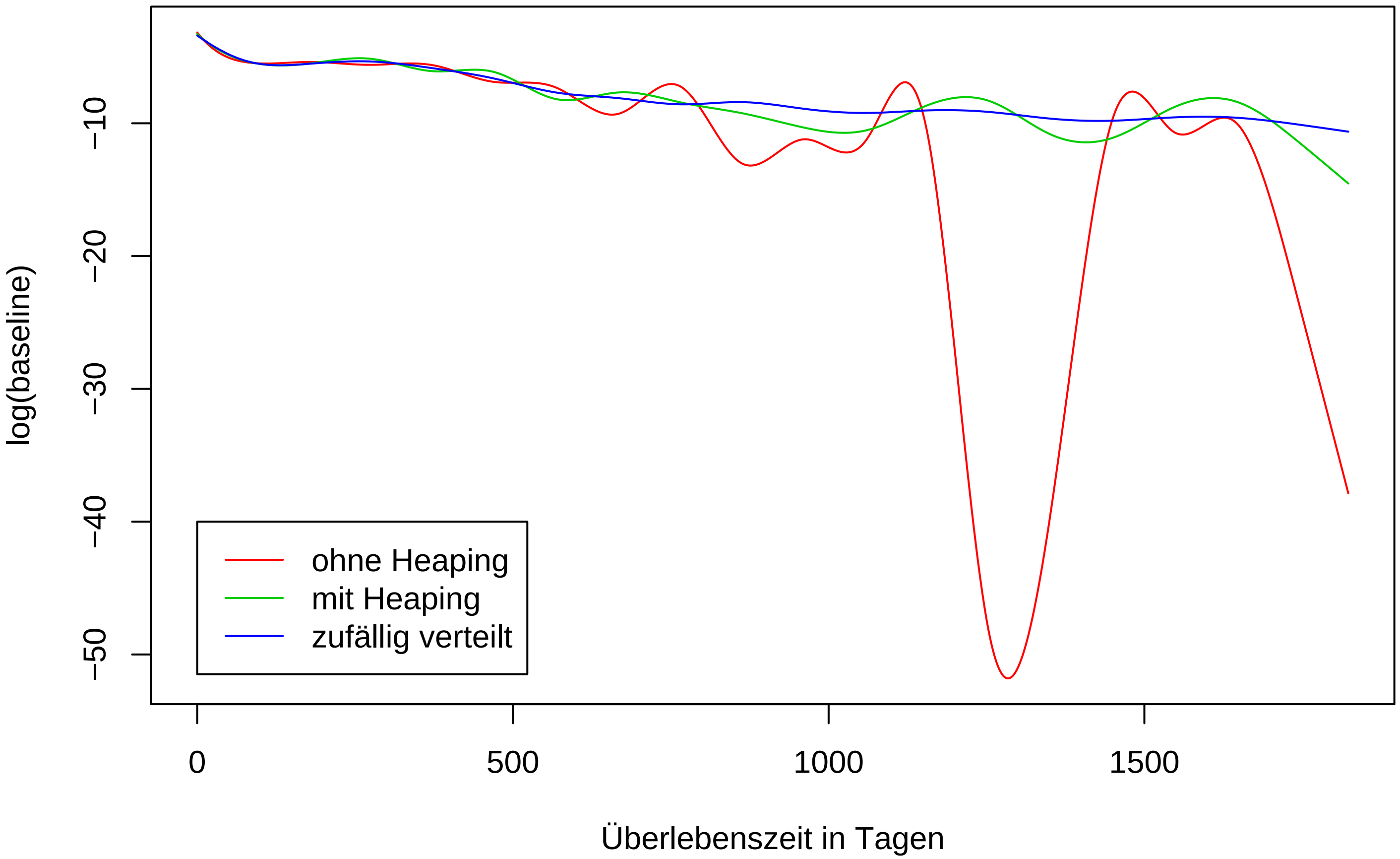


Räumlicher Effekt.



Body Mass Index der Mutter.

Kovariable	Schätzer	p-Wert
Intercept	-8.27	$<$0.0001
Stillen (zeitvariierend)	-4.27	$<$0.0001
Geburt im Krankenhaus	-0.54	0.0001
Lange Geburt	0.28	0.0047
Christen	-0.52	0.0001
Moslems	0.79	
Andere	-0.27	0.2879



log(baseline)
−10
−20
−30
−40
−50
0
500
1000
1500
Überlebenszeit in Tagen
ohne Heaping
mit Heaping
zufällig verteilt

Simulationsergebnisse

- Vergleich von
  - Intervallzensierung,
  - Zeitdiskreten Modellen,
  - Zufälliger Verteilung der Beobachtungen über die Intervalle.
- Design:
  - $\lambda(t) = \lambda_0(t) \exp\left[f_1(x) + f_{spat}(s)\right].$
  - Intervallzensierte und rechtszensierte Beobachtungen.
  - Rauhe und glatte Baseline-Hazardraten.

- Ergebnisse für die Baseline-Hazardrate:
  - Bei rauher Baseline-Hazardrate bessere Schätzung durch den Intervallzensierungsansatz.
  - Bei glatter Baseline können andere Verfahren besser abschneiden.
- Ergebnisse für die Kovariableneffekte:
  - Bei kleiner Anzahl breiter Intervalle liefert der Intervallzensierungsansatz die besten Ergebnisse.
  - Bei vielen schmalen Intervallen fast keine Unterschiede mehr.

Diskussion

- Bayesianisches Modell zur Analyse komplexer Hazard Regressionsmodelle bei intervallzensierten Daten ohne Verwendung von MCMC Simulationstechniken.

  $\Rightarrow$ Keine Fragen zu Konvergenz und Mixing.

- Eng verwandt mit penalisierter Likelihood Schätzung in einem frequentistischen Kontext.

- Erweiterungen:
  - Mehrzustands-Modelle.
  - Konkurrierende Risiken.
  - Einbezug von Intervallzensierung in diese komplexeren Modellen.

Referenzen

- Kneib, T. and Fahrmeir, L. (2004): A mixed model approach for structured hazard regression. SFB 386 Discussion Paper 400, University of Munich.

- Kneib, T. (2005): Geoadditive hazard regression for interval censored survival times. SFB 386 Discussion Paper 447, University of Munich.

- Erhältlich unter

  `http://www.stat.uni-muenchen.de/~kneib`