,

Analyse von Lebensdauern

Dozent: Christian Heumann [1]

[1]Institut für Statistik
Ludwig-Maximilians-Universität München

Analyse von Lebensdauern
WiSe 2010/11
Kapitel 5: Modifikationen und Erweiterungen von Hazardraten-Modellen

Kapitel 5: Modifikationen und Erweiterungen von Hazardraten-Modellen

1. Zeitdiskrete Survivalmodelle
   - Modelle und Datenlage
   - Inferenz
   - Beispiele
2. Frailty-Modelle
   - Modelle mit zufälligen Effekten
   - Schätzkonzepte
   - Beispiel

Erweiterungen (1)

Einige mögliche Erweiterungen von Hazardratenmodellen:

- Diskrete Lebensdauer $\longrightarrow$ Zeitdiskrete Survivalmodelle
- Erweiterungen des Cox-Modells: Regularisierung bei hochdimensionalen Prädiktoren
- Vorhersage bei Survivalmodellen
  - zeitlich abhängige ROC-Kurven
  - Prognosefehler
- Das Additive-Modell von Aalen $\lambda(t) = Y(t) \cdot \mathbf{x}'\beta(t)$
- Das Additiv-Multiplikative-Modell

$$\lambda(t) = Y(t) \cdot \left\{\mathbf{x}'\beta(t) + \lambda_0(t)\exp(\mathbf{y}'\gamma(t))\right\}.$$

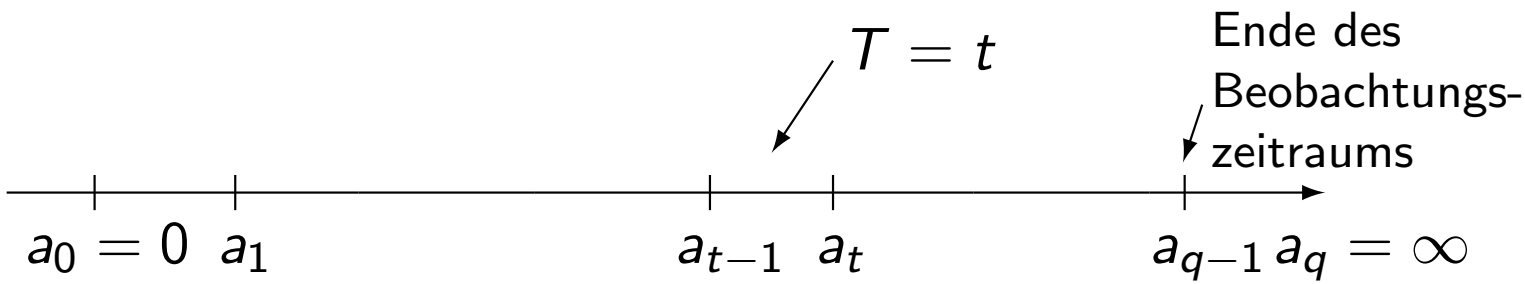
Erweiterungen (2)

- Vergleich von Behandlungsformen nicht nur anhand von Lebensdauern sondern auch von *quality-of-life* während der Lebensdauer (Tunes da Silva et al., 2009).
- Multivariate Analyse von Lebensdauern:
  - → Frailty-Modelle
  - → Mehr Information in z.B. Hougaard (2000)
- Modellierung von multiplen Ereignissen pro Individuum bzw. von unter einander abhängigen Ereignissen:
  - → Vorlesung *Ereignisanalyse* im WiSe 2010/2011
  - → Mehr Information in z.B. Therneau und Grambsch (2000)

Outline

5.1.1 Modelle und Datenlage (1)

- Zeitachse zerlegt in q Intervalle



- Anstelle der stetigen Zeit beobachtet man die diskrete Zeit $T \in \{1, \ldots, q\}$, mit

$$T = t \iff \text{Ausfall / Tod / Ereignis in } [a_{t-1}, a_t)$$

- Diskrete Hazardfunktion:

$$\lambda(t \,|\, \mathbf{x}) = P(T = t \,|\, T \geq t, \mathbf{x}), \quad t = 1, 2, \ldots$$

Modelle und Datenlage (2)

- Interpretation der diskreten Hazardfunktion: *Bedingte Wahrscheinlichkeit* für Ereignis in $[a_{t-1}, a_t)$, gegeben das Intervall wurde erreicht.
- Rechtszensierte Daten:

$$\{(t_i, \delta_i, \mathbf{x}_i); i = 1, \ldots, n\},$$

wobei

$$\delta_i = \begin{cases} 1 & , \quad \text{Ausfall in } [a_{t_i-1}, a_{t_i}) \\ 0 & , \quad \text{Zensierung in } [a_{t_i-1}, a_{t_i}) \end{cases}$$

Sichtweise als binärer Entscheidungsprozess

- Sei R_t die Menge der Individuen, die im Intervall $[a_{t-1}, a_t)$ unter Risiko stehen:

$$R_t = \{i : t \leq t_i\}.$$

- Es wird angenommen, dass die Zensierung am Ende des Intervalls passiert.
- Definiere binäre Ereignisindikatoren für $i \in R_t$

$$y_{it} = \begin{cases} 1 & , \quad \text{für } t = t_i \text{ und } \delta_i = 1 \\ 0 & , \quad \text{sonst} \end{cases}$$

- Somit ist für $i \in R_t$

$$\lambda_i(t|\mathbf{x}_i) = P(y_{it} = 1 \,|\, \mathbf{x}_i) \qquad\qquad t = 1, \ldots, q.$$

Zeitdiskrete Survivalmodelle (1)

- Idee: Setze binäres Regressionsmodell für $P(y_{it} = 1 \,|\, \mathbf{x}_i)$ an.
- Beispiel: Logit-Modell

$$\pi_{it} \;=\; P(y_{it} = 1 \,|\, \mathbf{x}_{it}) \;=\; \frac{\exp(\eta_{it})}{1 + \exp(\eta_{it})} \;\equiv\; \lambda(t \,|\, \mathbf{x}_{it})$$

mit Prädiktor

$$\eta_{it} \quad = \quad \underbrace{\beta_{0t}}_{\text{Zeit-Effekt}} + \underbrace{\mathbf{x}'_{it}\,\boldsymbol{\beta}}_{\text{Kovariaten-Effekt}}$$

$$\Longleftrightarrow \quad \frac{\pi_{it}}{1 - \pi_{it}} \;=\; \underbrace{\exp(\beta_{0t})}_{\gamma_{0t}} \exp(\mathbf{x}'_{it}\,\boldsymbol{\beta})$$

Baseline-Hazardfunktion

- Datendarstellung für ein solches Regressionsmodell $\longrightarrow$ Tafel.

Zeitdiskrete Survivalmodelle (2)

- Für Intervallbreiten $\longrightarrow 0$ und $q \longrightarrow \infty$ konvergiert das zeitdiskretes Logitmodell gegen das zeitstetiges Cox-Modell (Fleming und Harrington, 1991).
- Somit: Für ein feines Gitter approximiert das Logit-Modell mit Parametern $(\beta_{01}, \ldots, \beta_{0q}, \boldsymbol{\beta})'$ das Cox-Modell mit Baseline-Hazardrate $\lambda_0(t)$ und Prädiktor $\mathbf{x}'\boldsymbol{\beta}$.
- Alternative: Probit-Modell

$$\pi_{it} \;=\; P(y_{it} = 1 \,|\, \mathbf{x}_{it}) \;=\; \Phi(\eta_{it}) \;\equiv\; \lambda(t \,|\, \mathbf{x}_{it})$$

- Weitere Modelle in Fahrmeir und Tutz (2001).

Zeitdiskrete Survivalmodelle (3)

- Gruppiertes Cox-Modell (aka. komplementäre log-log-Modell)
- Sei T_{stet} zugrundeliegende stetige Lebensdauer mit Cox multiplikativer Hazardfunktion

$$\lambda_{stet}(t \mid \mathbf{x}_i) = \lambda_0(t) \exp(\mathbf{x}_i' \beta) \,.$$

- Die diskrete Hazardfunktion der entsprechenden diskreten Überlebenszeit ist dann ($\rightarrow$ Tafel)

$$\lambda(t \mid \mathbf{x}_i) = 1 - \exp(-\exp(\beta_{0t} + \mathbf{x}_i' \beta)) \;\equiv\; \pi_{it},$$

  wobei $\beta_{0t} = \log(\exp(\theta_t) - \exp(\theta_{t-1}))$ und $\theta_t = \log\{\Lambda_0(a_t)\}$.
- $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0q})'$ ist also eine diskrete Reparametrisierung von $\lambda_0(t)$ im zugrundeliegenden Cox-Modell.

5.1.2 Inferenz (1) – GLMs

- Parameter: $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta})'$
- Für q klein (d.h. große Gitter- / bzw. Intervallbreiten): Parametrische Likelihood-Inferenz mit Daten wie in 5.1.1
- Die Likelihood entspricht der des binären Regressionsmodells

$$\hat{\boldsymbol{\theta}}_{ML} \sim N(\boldsymbol{\theta},\, F^{-1}(\hat{\boldsymbol{\theta}}_{ML})),$$

  wobei $F(\cdot)$ die beobachtete Fisher-Information ist.
- Maximum-Likelihood-Schätzung mit GLM-Software
- Für q groß: Parametrische ML-Inferenz ist instabil, da $\dim(\boldsymbol{\theta}) = q + \dim(\boldsymbol{\beta})$ groß.

Inferenz (2) – GAMs

- Fasse β_0 als „glatte“ Folge bzw. Funktion $g_0(t) := \beta_{0t}$ auf.
- Modelliere $g_0(t)$ z.B. durch Polynom oder Regressions-Spline

$$g_0(t) = \sum_{j=1}^{m} \gamma_j \cdot \underbrace{B_j(t)}_{\text{Basisfunktionen}}$$

- Bei zusätzlicher Penalisierung: Nicht- bzw. semiparametrische Inferenz

$$\pi_{it} = h(\underbrace{g_0(t) + \mathbf{x}_{it}'\beta}_{\substack{\text{semiparametrischer} \\ \text{Prädiktor}}}) \ ,$$

  $g_0(t)$ durch Glättungs-Spline oder P-Spline mit Software für GAMs schätzen; geht auch für $m \approx 30 - 40$.

Inferenz (2) – GAMs

- Erweiterung zu allgemeinerem semiparametrischem Prädiktor:

$$\eta_{it} = \underbrace{g_0(t)}_{\text{Baseline-Effekt}} + \underbrace{g_1(t)\, z_{i1} + \cdots + g_q(t)\, z_{iq}}_{\text{Zeitvariierende Effekte}} + \underbrace{f_1(v_i)}_{\text{stetige Kovariable, z.B. Alter}} + \cdots + \mathbf{x}_i'\boldsymbol{\beta}$$

- Notation:

$g_0(t), \ldots, g_q(t)$	glatte Funktionen von t
$f_1(v_1), \ldots$	glatte Funktionen stetiger Kovariablen
$\mathbf{x}'\boldsymbol{\beta}$	üblicher linearer Teil des Prädiktors

- Inferenz mit GAM bzw. Variierenden-Koeffizienten-Modellen, z.B. mit dem Paket `mgcv` in R.

Beispiel: Kindersterblichkeit in Nigeria

- Analyse aus Adebayo und Fahrmeir (2005) mit Response

  $T \in \{1, \ldots, 36\}$ (Über-)Lebensdauer von Kind in Monaten

- Modell (ohne räumlichen Effekt $f_{\text{spat}}(\mathbf{s}_i)$):

$$\eta_{it} = g_0(t) + g_1(t)\, bf_{it} + g_2(t)\, ma_{1i} + g_3(t)\, ma_{2i} + \mathbf{x}_i'\beta$$

- Kovariablen: Indikator für Stillen

$$bf_t = I(\text{im Monat } t \text{ wird gestillt})$$

Effekt-Codierung für Alterskategorien <22, 22–35, >35)

$$ma_1 = \begin{cases} 1 & magb < 22 \text{ Jahre} \\ -1 & magb > 35 \text{ Jahre} \end{cases}$$

$$ma_2 = \begin{cases} 1 & magb \in \{22 - 35 \text{ Jahre}\} \\ -1 & magb > 35 \text{ Jahre} \end{cases}$$

Beispiel: Kindersterblichkeit in Nigeria

(S. Adebayo, L. Fahrmeir, 2005, Statistics in Medicine 24,709-728)

Daten: 1999 Nigeria Demographic and Health Survey (NDHS).
Teildatensatz zu 3552 Kindern, Alter 1-36 Monate

Zielvariable: Lebensdauer in Monaten
Zensierung: Alter > 36 Monate = 3 Jahre

Kovariablen: Siehe Table 2.

Table 2: *Descriptive information about selected covariates on the 3552 children involved in the survey.*

	Bio-demographic and health factors		
Place of Residence:	Urban (29.3%)	Rural (70.7%)	
Sex:	Male (50.9%)	Female (49.1%)	
Tetanus injection:	Yes (55.2%)	No (38.9%)	Missing (5.9%)
Mother's education:	Primary (72.8%)	$>$ Primary (27.2%)	
Mother assisted:	Yes (85.6%)	No (11.8%)	Missing (2.6%)
Place of delivery:	Hospital (36.9%)	Others (60.4%)	Missing (2.7%)
Preceding birth interval:	$\geq$ 2 Years (70.2%)	$<$ 2 Years (8.8%)	Missing (21.0%)
Antenatal visit:	$\geq$ 1 visit (57.7%)	None (38.6%)	Missing (3.7%)
Mother's age at birth:	$<$ 22 years (25.8%)	22-35 years (64.0%)	$>$ 35 years (10.2%)
Birth Order:	First born (20.9%)	Others (79.1%)	
Weight at birth:	$\leq$ Average (63.9%)	$>$ Average (31.7%)	Missing (4.4%)
Had cough:	Yes (20.3%)	No (66.9%)	Missing (12.8%)
Had diarrhea:	Yes (13.9%)	No (74.0%)	Missing (12.1%)
Had fever:	Yes (27.5%)	No (60.3%)	Missing (12.2%)
	Socio-economic factors		
Mother's working status:	Working (50.1%)	Not working (49.4%)	Missing (0.5%)
Household size:	$\leq$ 5 (42.0%)	$\geq$ 6 (58.0%)	
Toilet facility:	Any type (73.2%)	None (25.1%)	Missing (1.7%)
Electricity:	Yes (41.7%)	No (56.7%)	Missing (1.6%)
Water:	Controlled (76.4%)	Uncontrolled (22.2%)	Missing (1.4%)
Condition of wall and floor:	Good (40.2%)	Poor (57.3%)	Missing (2.5%)

Zusätzliche räumliche Information: Distrikt, in dem die Mutter lebt.

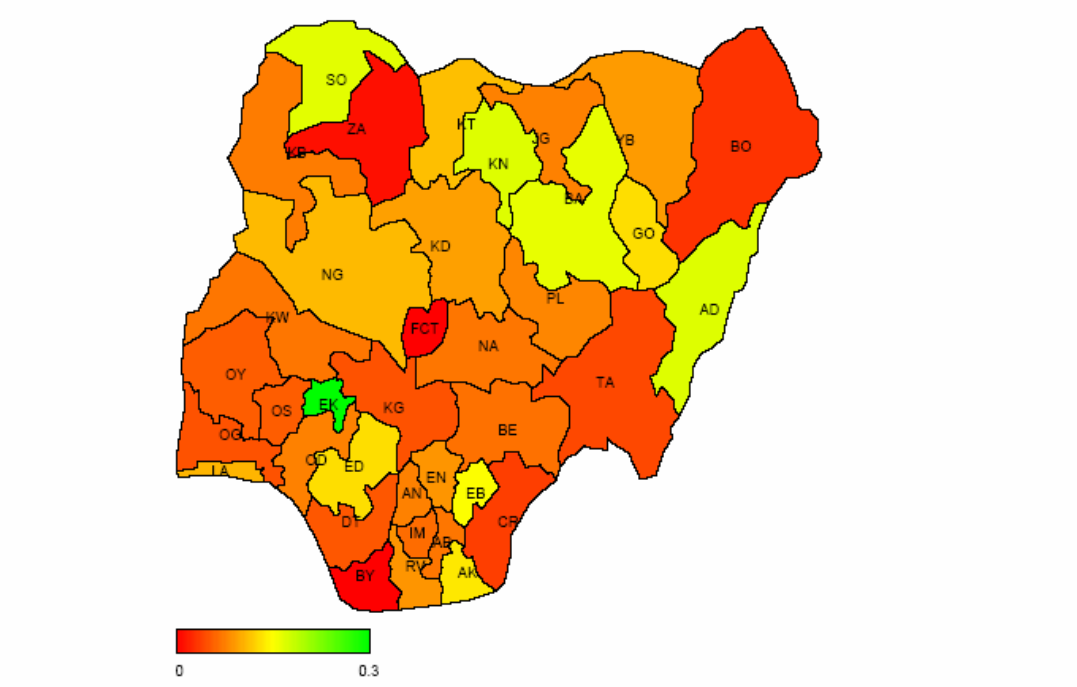


Figure 1: *Map of Nigeria showing the geographical distribution of mortality rates in proportion for the 37 districts.*

Analysen mit zeitdiskreten Survival-Modellen

Table 4: *Posterior estimates of the fixed effects in model M7.*

Variable	mean	std. error	5%	10%	90%	95%
urban	0.024	0.080	-0.137	-0.082	0.126	0.177
male	0.053	0.066	-0.079	-0.033	0.135	0.181
At most primary	0.166	0.102	-0.030	0.036	0.295	0.379
Assisted	-0.151	0.098	-0.342	-0.271	-0.025	0.056
Hospital	-0.143	0.101	-0.337	-0.269	-0.017	0.059
Long birth interval	-0.422	0.082	-0.571	-0.526	-0.322	-0.250
Assist*antenat	-0.159	0.078	-0.308	-0.265	-0.056	-0.013

Table 5: *Posterior estimates of the odds ratios for model M7.*

Variable	mean	std. error	5%	10%	90%	95%
Urban	1.063	0.170	0.801	0.849	1.287	1.356
Male	1.122	0.149	0.899	0.936	1.311	1.393
At most primary	1.424	0.296	1.016	1.075	1.802	1.965
Assisted	0.754	0.153	0.531	0.582	0.951	1.024
Hospital	0.767	0.157	0.537	0.584	0.967	1.056
Long birth interval	0.436	0.073	0.338	0.350	0.525	0.567
Assist*antenat	0.737	0.116	0.566	0.589	0.894	0.939

Table 7: *Posterior estimates of the fixed effects of socio-economic factors in model M7 .*

Variable	mean	std. error	5%	10%	90%	95%
Mother currently working	0.165	0.079	0.017	0.063	0.267	0.330
Household ≤ 5	0.224	0.074	0.089	0.131	0.323	0.361
Has toilet	-0.129	0.079	-0.292	-0.233	-0.030	0.021
Electricity	0.118	0.089	-0.039	0.013	0.233	0.301
High quality of wall and floor	-0.243	0.199	-0.651	-0.513	-0.0001	0.131
Water in the residence	-0.152	0.086	-0.302	-0.261	-0.049	0.031

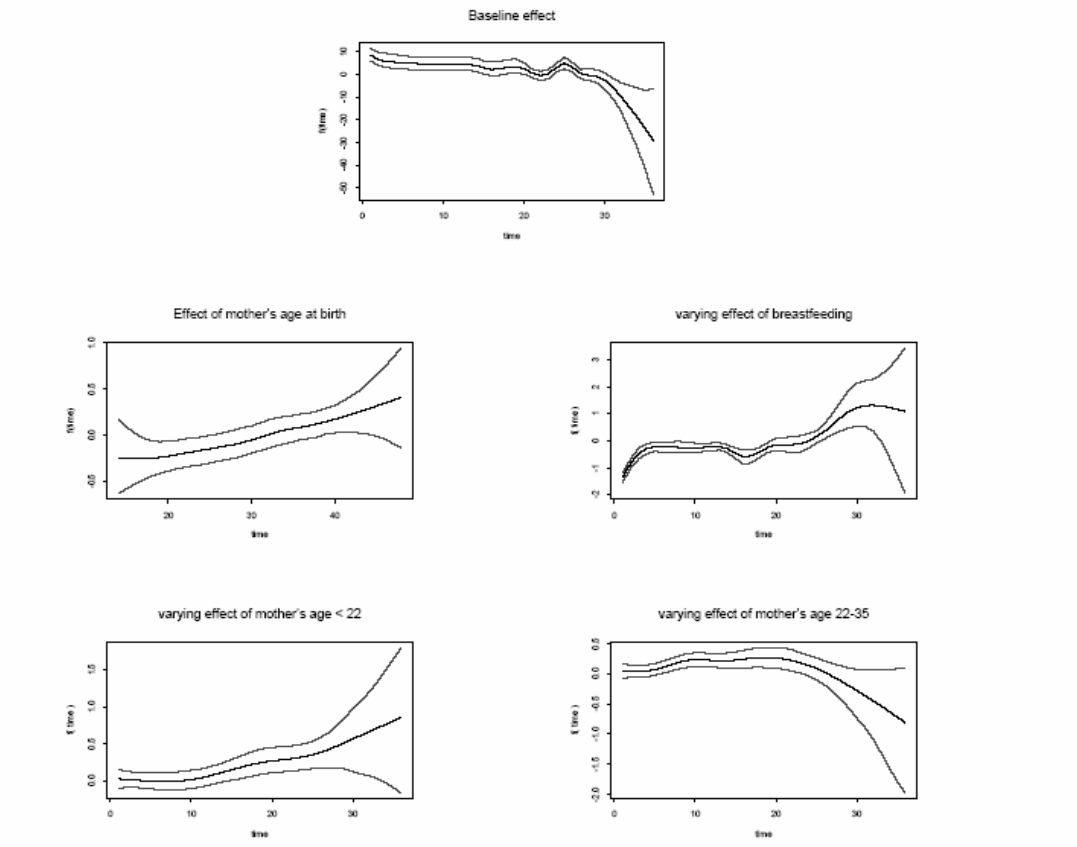


Figure 3: *Graphs of nonlinear effects from top to bottom: baseline effect, effect of mother's age at birth (only for model M6), varying effect of breastfeeding, varying effect of mother's age < 22 years and varying effect of mother's age 22-35 years respectively for model M7.*

Beispiel: Dauer von Arbeitslosigkeit in Westdeutschland

- Analyse in Fahrmeir et al. (2003).
- 1980 - 1995, ca. 16 000 Männer / 20 000 Frauen aus IAB (Institut für Arbeitsmarkt- und Berufsforschung)-Stichprobe
- Response: Dauer der Arbeitslosigkeit in Monaten
- Modellierung: Probitmodell (ohne räumlichen Effekt $f_{\text{spat}}(\mathbf{s}_i)$):

$$\eta_{it} = g_0(t) + f_1(a_i) + f_2^{trend}(k_{it}) + f_3^{saison}(k_{it}) + \mathbf{x}_i'\boldsymbol{\beta}$$

- Kovariablen:

a	Alter in Jahren (zu Beginn der Arbeitslosigkeit)
k_t	Kalenderzeit
x	enthält (folgende) Kovariablen in Effekt-Codierung

- Getrennte Analysen für Frauen und Männer.

Kovariablen (jeweils für Frauen und Männer)

- Nationalität: deutsch / nicht-deutsch (Referenzkategorie)
- Ausbildung: keine abgeschlossene Ausbildung / abgeschlossene Berufsausbildung (Referenzkategorie) / Universitätsabschluss
- Alter (in Jahren)
- Unterstützung: Arbeitslosengeld / Arbeitslosenhilfe (für jedes Monat der Arbeitslosigkeit)
- Anzahl vorangehender Arbeitslosenzeiten
- Ökonomische Sektoren: Landwirtschaft / Produktion (Referenzkategorie) / Energie, Wasser und Bergbau / Bau / Handel / Transport und Kommunikation / Finanzsektor / Dienstleistungen / Non-Profit-Organisationen und private Haushalte / Behörden und Sozialversicherung
- Landkreis der Wohnung

Resultate

Basierend auf zeitdiskretem Probit-Modell; ohne bzw. mit räumlich-zeitlicher Interaktion.

Variable	mean	std. dev.	10 % quant.	median	90 % quant.
German	0.03	0.01	0.02	0.03	0.04
foreign	-0.03	0.01	-0.04	-0.03	-0.02
no vocational training	-0.02	0.01	-0.03	-0.02	0.00
vocational training	0.03	0.01	0.02	0.03	0.04
university	-0.01	0.02	-0.04	-0.01	0.01
unemployment assistance	-0.17	0.01	-0.17	-0.17	-0.16
unemployment benefit	0.17	0.01	0.16	0.17	0.17
P=0	-0.09	0.01	-0.10	-0.09	-0.08
P=1,2	-0.01	0.01	-0.01	-0.01	0.00
P≥3	0.10	0.01	0.09	0.10	0.11
agriculture	0.24	0.02	0.21	0.24	0.27
manufacturing	-0.02	0.01	-0.04	-0.02	-0.01
energy	-0.19	0.04	-0.24	-0.19	-0.14
construction	0.15	0.01	0.13	0.15	0.17
trade	-0.02	0.01	-0.04	-0.02	-0.01
transport	0.06	0.02	0.04	0.06	0.08
financial sector	-0.24	0.04	-0.29	-0.24	-0.19
service industry	0.03	0.01	0.01	0.03	0.04
private	0.01	0.03	-0.03	0.01	0.06
public	-0.01	0.02	-0.04	-0.01	0.01

TABLE 1. a) Estimates of constant parameters: men.

Variable	mean	std. dev.	10% quant.	median	90% quant.
German	0.07	0.01	0.06	0.07	0.08
foreign	-0.07	0.01	-0.08	-0.07	-0.06
no vocational training	-0.04	0.01	-0.05	-0.04	-0.03
vocational training	-0.04	0.01	-0.05	-0.04	-0.02
university	0.07	0.02	0.05	0.07	0.10
unemployment assistance	-0.09	0.01	-0.10	-0.09	-0.09
unemployment benefit	0.09	0.01	0.09	0.09	0.10
P=0	-0.09	0.01	-0.10	-0.09	-0.08
P=1,2	-0.03	0.01	-0.04	-0.03	-0.02
P$\geq$ 3	0.12	0.01	0.11	0.12	0.13
agriculture	0.19	0.03	0.15	0.19	0.22
manufacturing	-0.15	0.01	-0.16	-0.15	-0.14
construction	-0.02	0.03	-0.06	-0.02	0.02
trade	-0.01	0.01	-0.02	-0.01	0.01
transport	0.09	0.02	0.06	0.09	0.12
financial sector	-0.18	0.03	-0.22	-0.18	-0.15
service industry	0.06	0.01	0.05	0.06	0.07
private	0.00	0.02	-0.02	0.00	0.03
public	0.02	0.02	0.00	0.02	0.04

TABLE 1. b) Estimates of constant parameters: woman.

Variable	mean	std. dev.	10 % quant.	median	90 % quant.
$P = 0$	-0.07	0.01	-0.08	-0.07	- 0.06
$P = 1,2$	0	0.01	-0.01	0	0.01
$P \geq 3$	0.07	0.01	0.06	0.07	0.08

a) men

Variable	mean	std. dev.	10 % quant.	median	90 % quant.
$P = 0$	-0.08	0.01	-0.09	-0.08	-0.08
$P = 1,2$	-0.03	0.01	-0.04	-0.03	-0.02
$P \geq 3$	0.11	0.01	0.10	0.11	0.12

b) women

TABLE 2. Estimates of the effect of the number of previous unemployment spells.

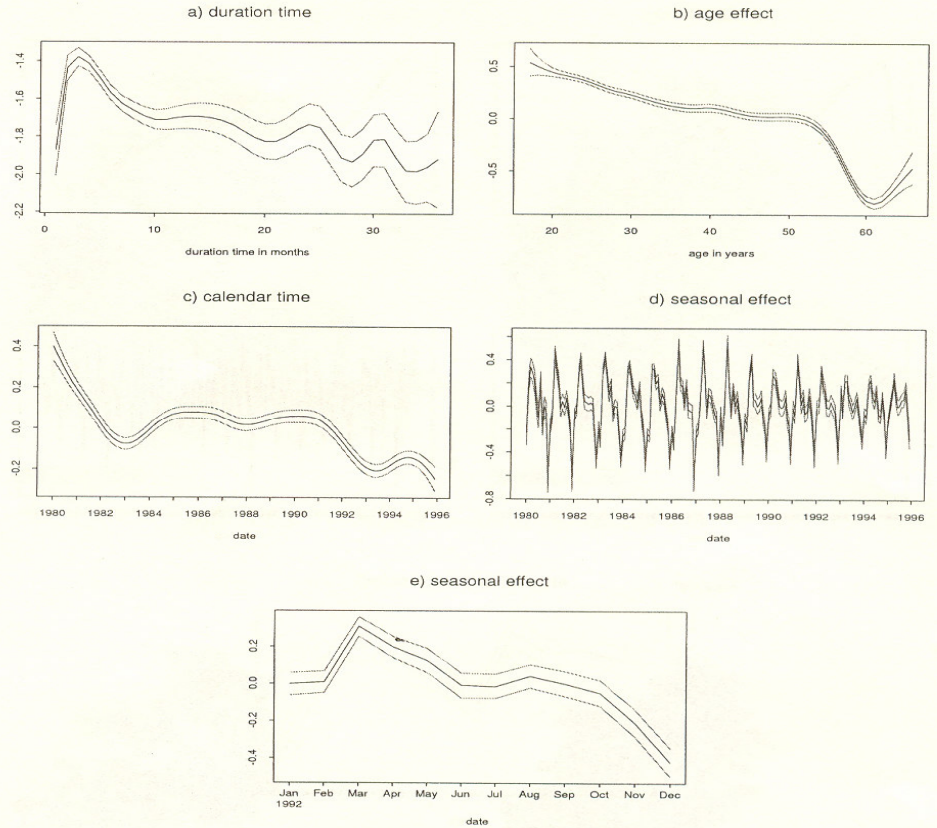


FIGURE 1. Estimated nonparametric functions and seasonal effect for males. Shown is the posterior mean within 80 % credible regions.

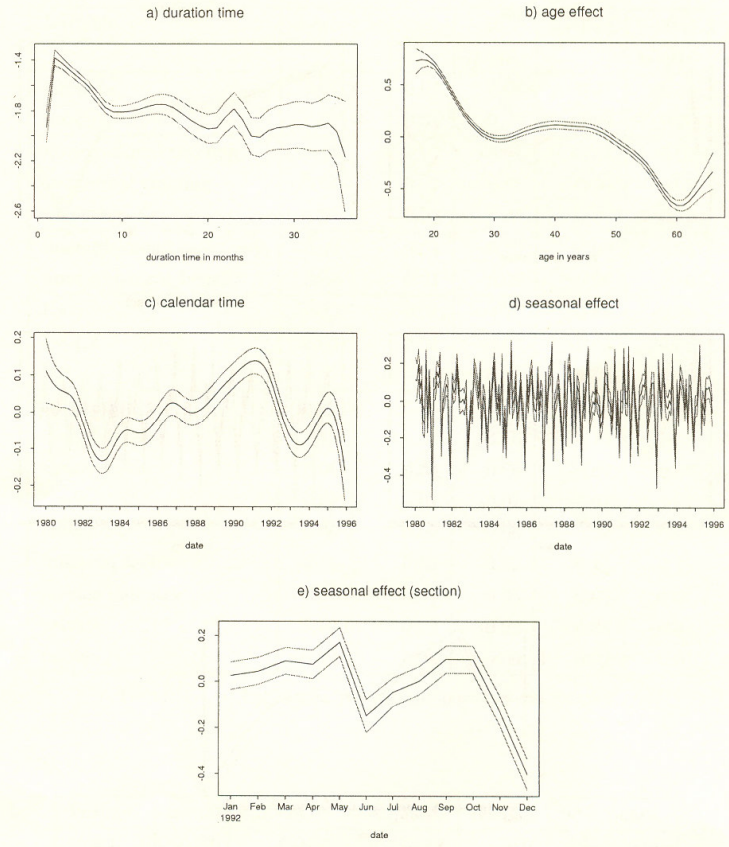


FIGURE 2. Estimated nonparametric functions and seasonal effect for females. Shown is the posterior mean within 80 % credible regions.

Outline

Frailty-Modelle

- Bis jetzt: Unabhängige Überlebenszeiten der Individuen.
- Diese Annahme ist nicht immer gegeben:
  - Analyse von Zwillingspaaren oder Tieren aus dem gleichem Wurf (genetische Komponente)
  - Analyse von Ehepaaren oder Patienten aus verschiedenen Krankenhäusern (gleiche Umgebung)
  - Eintreffen verschiedener nicht-tödlichen Krankheiten im gleichen Individuum.

Frailty-Modell

Frailty-Modelle modellieren die Assoziation zwischen Individuen durch einen unbeobachteten zufälligen Effekt.

Einschub: Linear Mixed Models (LMMs)

- Lineare gemischte Modelle (aka. linear mixed models) für Longitudinal- oder Clusterdaten besitzen folgende Form

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\boldsymbol{\gamma}_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

  $i = 1, \ldots, m; j = 1, \ldots, n_i$ (Fahrmeir et al., 2007).
- Dabei bezeichnet m die Anzahl der Cluster bzw. Individuen i, und n_i die Anzahl von (wiederholten) Beobachtungen pro Cluster oder Individuum.
- Für $z_{ij} \equiv 1$ erhält man das Modell mit einem zufälligen Intercept:

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \gamma_i + \epsilon_{ij}.$$

- Für die zufälligen Effekte wird meist $\gamma_i \overset{iid}{\sim} N(0, \tau^2)$ angenommen.

Frailties im Cox-Modell

- Cluster bzw. individuenspezifische zufällige Effekte werden in den linearen Prädiktor der Hazardrate einbezogen:

$$\begin{aligned}\lambda(t|\mathbf{x}_{ij}) &= \lambda_0(t)\exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + \gamma_i) \\ &= \lambda_0(t)\nu_i \exp(\mathbf{x}_{ij}'\boldsymbol{\beta}), \quad \nu_i = \exp(\gamma_i),\end{aligned}$$

  $i = 1, \ldots, m$ und $j = 1, \ldots, n_i$. Dabei ist m die Anzahl Cluster und n_i ist die Anzahl Beobachtungen des i'ten Clusters.

- Für die zufälligen Effekte wird z.B.

$$\gamma_i \overset{iid}{\sim} N(0, \theta^2)$$

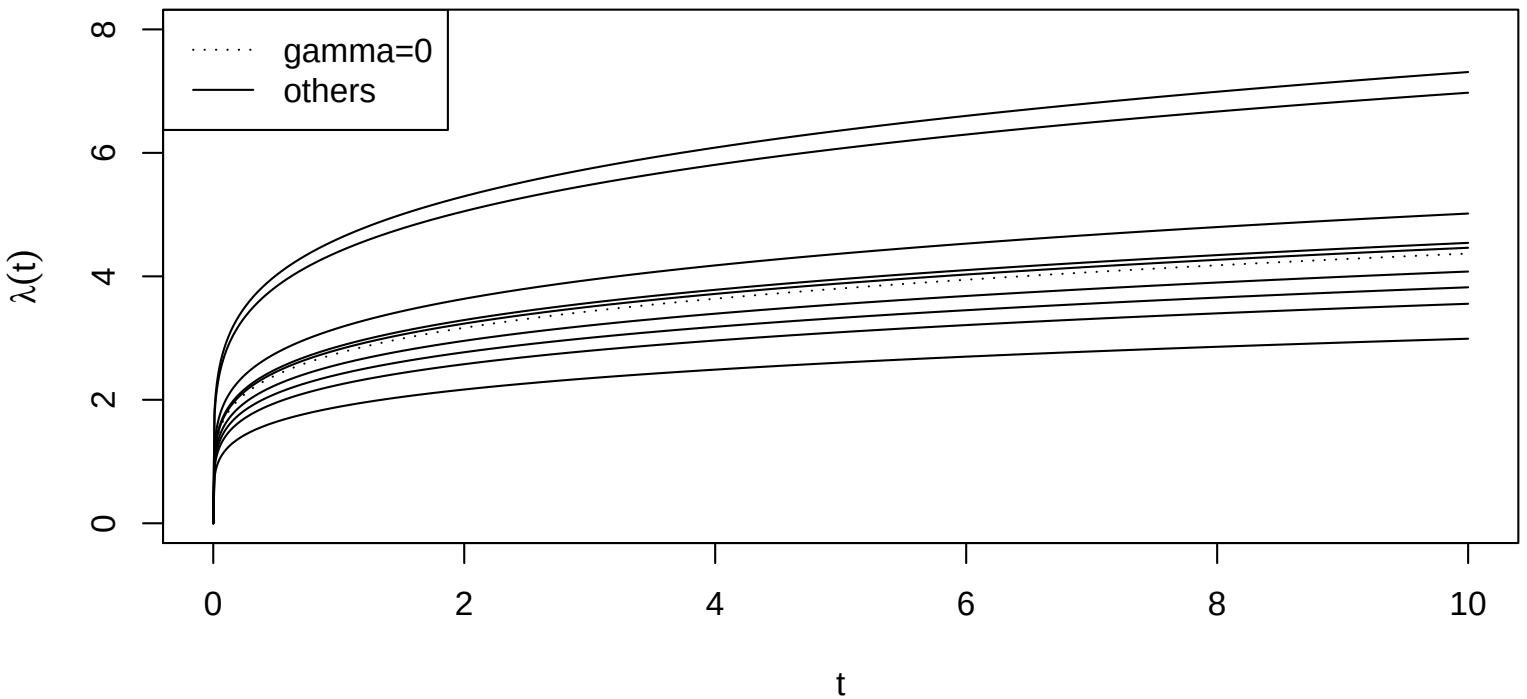
  bzw. $\nu_i \overset{iid}{\sim} \text{LogN}(0, \theta^2)$ oder $\nu_i \overset{iid}{\sim} \text{Ga}(\frac{1}{\theta}, \frac{1}{\theta})$ angenommen.

Beispiel: Random effects im Weibull-Modell

- Sei die Hazardrate wie im Weibull-Modell

$$\lambda(t) = \lambda\alpha(\lambda t)^{\alpha-1}\exp(\gamma),$$

  mit $\alpha, \lambda > 0$ und $\gamma \sim N(0, \sigma^2)$.
- Beispiel für $n = 10$ Individuen mit $\lambda = 2$ und $\alpha = 1.2$:



Clusterspezifische Effekte

Ursachen:

- Klinikeffekte der Klinik i mit n_i Patienten in einer klinischer Studie mit m beteiligten Kliniken (Multicenter-Analysen)
- Familieneffekte für n_i Mitglieder der Familie i.
- Räumliche Effekte für n_i erkrankte Personen aus Region i, $i = 1, \ldots, m$ in epidemologischer Studie oder n_i arbeitslose Personen aus Arbeitsmarktregion i, $i = 1, \ldots, m$.

Konsequenzen:

- Durch den gemeinsamen clusterspezifischen Effekt γ_i sind die Lebensdauern von Individuen aus Cluster i positiv korreliert.
- Im Gamma-Frailty ist die Korrelation – gemessen durch Kendall's τ – gleich $\theta/(\theta + 2)$.
- Frailty-Modelle bilden eine Möglichkeit zur Spezifikation von *multivariaten Survival-Modellen*.

Korrelierte zufällige Effekte (1)

- Statt i.i.d. Effekten γ_i kann auch die Annahme korrelierter Effekte sinnvoll sein
- Beispiel: wenn γ_i der Effekt der Region i ist aus dem die Beobachtungen $j = 1, \ldots, n_i$ entstammen.
- Eine übliche Annahme in der räumlichen Epidemiologie ist, dass der Vektor $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$ ein Gauß-Markov-Zufallsfeld (GMRF) bildet.
- Im GMRF nimmt man an, dass

$$\gamma_i | \boldsymbol{\gamma}_{-i} = \gamma_i | \boldsymbol{\gamma}_{N(i)} \sim N\left(\frac{1}{m_i} \sum_{l \in N(i)} \gamma_l, \frac{\tau^2}{m_i}, \right)$$

wobei die m_i Gewichte sind und $N(i)$ die Nachbarschaft von i definiert $\longrightarrow$ Vorlesung Räumliche Statistik

Korrelierte zufällige Effekte (2)

- Falls die genaue Lokation $\mathbf{s}_i = (x_i, y_i)$ von jedem Individuum bekannt ist kann γ_i auch als Realisation an der Stelle $\mathbf{s}_i$ eines Gaussian Random Fields mit Erwartungswert 0 und isotroper Korrelationsfunktion $c(\cdot; \tau)$ angesehen werden:

$$E(\gamma_j) = 0,$$
$$\text{Cov}(\gamma_i, \gamma_j) = c(||\mathbf{s}_i - \mathbf{s}_j||), \quad i \neq j.$$

- Mögliche Kovarianzfunktionen:
  - Exponential: $c(u; \tau) = \exp(-\tau u)$
  - Gauss: $c(u; \tau) = \exp(-\tau u^2)$

Korrelierte zufällige Effekte (3) – Beispiele

Beispiele solch *Geoadditiver Survival Analysen*:

- Arbeitslosigkeitsdauer in Westdeutschland
- Kindersterblichkeit in Nigeria

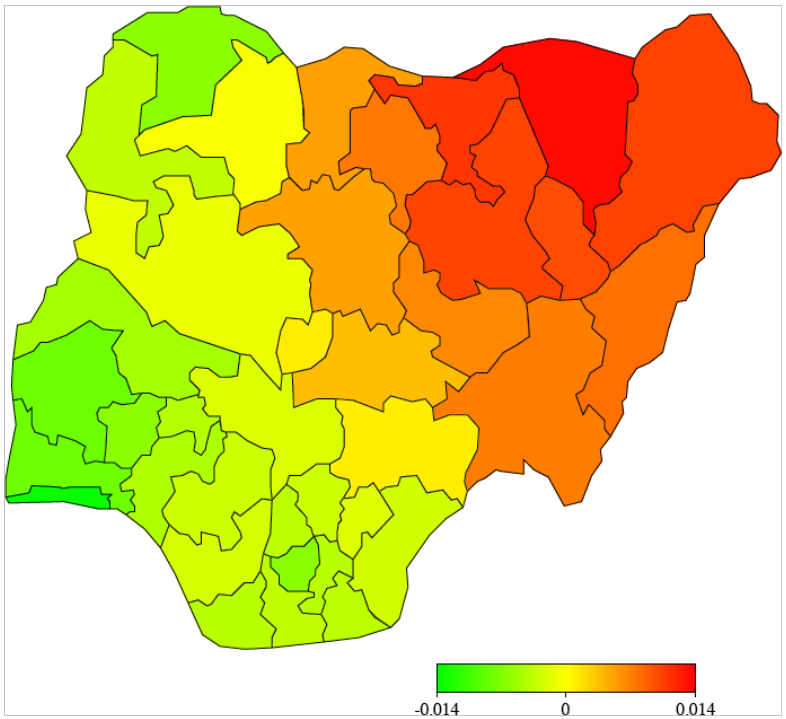


Figure: Räumlicher Frailty im Cox-Modell. Quelle: Kneib (2005).

Schätzkonzepte für Frailty-Modelle (1)

- Betrachtet wird *Shared-Frailty-Modell* mit linearem Prädiktor und einem zufälligem Intercept pro Cluster:

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}, \qquad i = \sum_{k=1}^{m} n_k,$$

  wobei $\mathbf{z}_i = (z_{i1}, \ldots, z_{im})'$ mit

$$z_{ij} = I(\text{Individuum } i \text{ ist Teil von Cluster } j).$$

- Unbekannte Parameter:
  - Die p festen Effekte $\boldsymbol{\beta}$
  - Varianzparameter bei LogN$(0, \theta^2)$-Verteilung bzw. θ Parameter der Ga$(\frac{1}{\theta}, \frac{1}{\theta})$-Verteilung
  - Die m zufälligen Effekte $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$

Schätzkonzepte für Frailty-Modelle (2)

- Die Funktion `coxph` implementiert einen in Therneau et al. (2003) beschriebenen penalisierten Likelihood-Ansatz.
- Betrachten Frailty-Modell mit $\nu_i \overset{iid}{\sim} F_\theta$. Sei $\nu_i = \exp(\gamma_i)$.
- Schätzung für dieses Modell durch Maximierung nach $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ der penalisierten Partial Loglikelihood (PPLL)

$$PPLL = PLL(\boldsymbol{\beta}, \boldsymbol{\gamma}) - g(\boldsymbol{\gamma}, \lambda),$$

wobei $PLL(\boldsymbol{\beta}, \boldsymbol{\gamma})$ die gewöhnliche PLL

$$\sum_{i=1}^{n} \int_0^\infty \left[Y_i(t)(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}) - \log \left\{ \sum_{k=1}^{n} Y_k(t) \exp(\mathbf{x}_k'\boldsymbol{\beta} + \mathbf{z}_k'\boldsymbol{\gamma}) \right\} \right] dN_i(t)$$

ist, und $g(\boldsymbol{\gamma}, \theta)$ eine Penalty-Funktion ist.

Schätzkonzepte für Frailty-Modelle (3)

- Das Gamma-Frailty-Modell $\mathrm{Ga}(\frac{1}{\theta}, \frac{1}{\theta})$ wird durch Verwendung der Penalty-Funktion

$$g(\boldsymbol{\gamma}, \theta) = \frac{1}{\theta} \sum_{i=1}^{m} (\gamma_i - \exp(\gamma_i))$$

  gelöst.
- Der Glättungsparameter θ wird in `coxph` über ein AIC-Kriterium bestimmt.
- Das $\mathrm{LogN}(0, \theta)$-Frailty-Modell wird durch folgende Penalty-Funktion gelöst:

$$g(\boldsymbol{\gamma}, \theta) = \frac{1}{2\theta} \sum_{i=1}^{m} \gamma_i^2.$$

- Details $\rightarrow$ Vorlesung *Gemischte Modelle* im SoSe 2009.

Beispiel

- Daten aus Therneau und Grambsch (2000) zur Untersuchung eines potentiell krebserzeugender Behandlungsform bei Ratten.
- Untersucht wurden jeweils drei Ratten aus 50 Würfen, wovon eine der drei die Behandlung bekommt. 40 der insgesamt 150 Ratten entwickelten einen Tumor.

```
R>data("rats")
R>(m.rats <- coxph(Surv(time, status) ~ rx + frailty(litter,
+     dist = "gauss"), data = rats))

Call:
coxph(formula = Surv(time, status) ~ rx + frailty(litter, dist = "gauss"),
    data = rats)

                          coef  se(coef) se2   Chisq DF   p
rx                        0.913 0.323    0.319  8.01  1.0 0.0046
frailty(litter, dist = "g                     15.57 11.9 0.2100

Iterations: 6 outer, 21 Newton-Raphson
     Variance of random effect= 0.412
Degrees of freedom for terms=  1.0 11.9
Likelihood ratio test=35.3  on 12.9 df, p=0.000711  n= 150
```

Literatur I

Adebayo, S. und L. Fahrmeir [2005]. Analysing child mortality in nigeria with geoadditive survival models. *Statistics in Medicine 24*, 709–728.

Fahrmeir, L., T. Kneib, und S. Lang [2007]. *Regression*. Springer.

Fahrmeir, L., S. Lang, J. Wolff, und S. Bender [2003]. Semiparametric bayesian time-space analysis of unemployment duration. *Allgemeines Statistisches Archiv 87*, 281–307.

Fahrmeir, L. und G. Tutz [2001]. *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd.Aufl.). Springer.

Fleming, T. R. und D. P. Harrington [1991]. *Counting Processes and Survival Analysis*. Wiley.

Hougaard, P. [2000]. *Analysis of Multivariate Survival Data*. Springer.

Kneib, T. [2005]. *Mixed model based inference in structured additive regression*. Dr. Hut Verlag. Dissertation, Available as `http://edoc.ub.uni-muenchen.de/5011/`.

Therneau, T. M. und P. M. Grambsch [2000]. *Modeling Survival Data – Extending the Cox Model*. Springer.

Therneau, T. M., P. M. Grambsch, und V. S. Pankratz [2003]. Penalized survival models and frailty. *Journal of Computational and Graphical Statistics 12*(1), 156–175.

Tunes da Silva, G., P. K. Sen, und A. C. Pedroso de Lima [2009]. Mean quality-adjusted survival using a multistate model for the sojourn times. *Journal of Statististcal Planning and Inference 138*(8), 2267–2282.