

Kai Arzheimer | Vorlesung Forschungsmethoden

Ereignisdatenanalyse

Übersicht

Grundbegriffe/Probleme
Parametrische Modelle
Cox Proportional Hazard Model
Software: Stata
Zusammenfassung



Harold Macmillan, PM 1957-1963

The greatest challenge in politics: „events, my dear boy, events“

Was sind Ereignisdaten?

- ▶ Alternative Begriffe: Failure-/Survival- etc. Analyse
- ▶ Einfacher Fall: Zeitdauer bis Ereignis eintritt
- ▶ Komplexer Fall: Zeit, die Objekt in einem von mehreren (evtl. reversiblen) Zuständen zubringt
- ▶ Politikwissenschaftliche Beispiele
 - ▶ Lebensdauer einer Regierung
 - ▶ Beginn/Dauer von Krieg und Frieden
 - ▶ Karrieren von Mandatsträgern
- ▶ Alternative Analysemodelle: Panelanalyse, Logit-Analyse, ...
- ▶ "Taking time seriously ..."

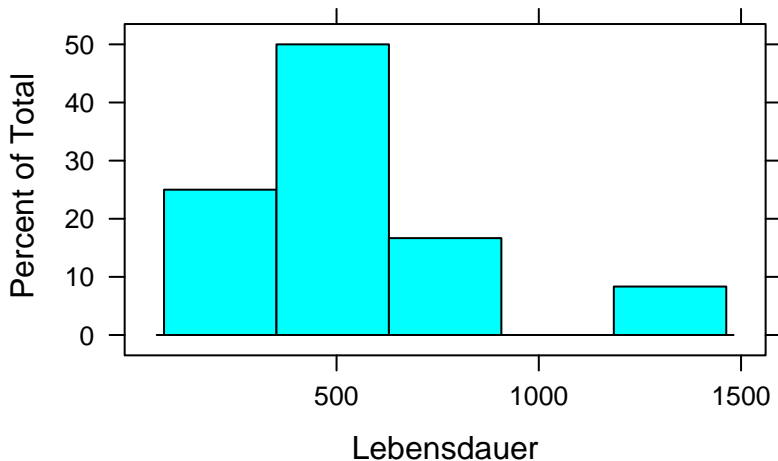
Terminologie

- ▶ Stammt aus der Medizin bzw. Ingenieurwissenschaften
- ▶ „Ereignis“: Tod eines Patienten, Versagen einer Maschine
- ▶ Verwirrend:
 - ▶ „Risiko“ für das Ende eines Krieges
 - ▶ „Versagen“ = Ende einer Kandidatur = Einzug ins Parlament
 - ...
- ▶ **Relevante Variable: Zeitdauer bis zum Eintritt eines Zustands**

Ein Beispiel: Kabinette in Italien bis 2009

Beginn	Ende	Dauer	PM	Ausrichtung
28/06/92	28/04/93	304	Giuliano Amato	links
28/04/93	10/05/94	377	Carlo Azeglio Ciampi	links
10/05/94	17/01/95	252	Silvio Berlusconi	rechts
17/01/95	17/05/96	486	Lamberto Dini	links
18/05/96	21/10/98	886	Romano Prodi	links
21/10/98	22/12/99	427	Massimo D'Alema	links
22/12/99	25/04/00	125	Massimo D'Alema (2. Amtszeit)	links
25/04/00	11/06/01	412	Giuliano Amato (2. Amtszeit)	links
11/06/01	23/04/05	1412	Silvio Berlusconi (2. Amtszeit)	rechts
23/04/05	17/05/06	389	Silvio Berlusconi (3. Amtszeit)	rechts
01/05/06	08/05/08	738	Romano Prodi (2. Amtszeit)	links
08/05/08	?	(385)	Silvio Berlusconi (4. Amtszeit)	rechts

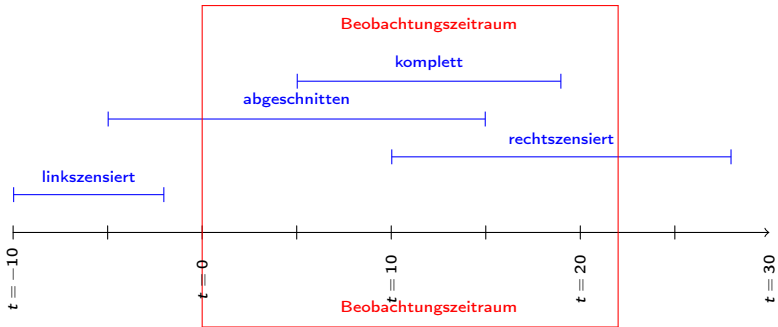
Ein Beispiel: Kabinette in Italien bis 2009



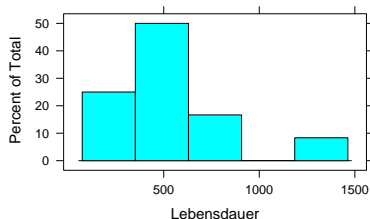
Warum besondere Modelle?

- ▶ Kovariaten nicht notwendigerweise über Zeit stabil
- ▶ Keine Werte $y < 0$
- ▶ Beobachtungen möglicherweise bimodal, „rechtszensiert“, „linkszensiert“, „abgeschnitten“ (truncated)
- ▶ Hochgradige Abweichung von Normalverteilung für ϵ
- ▶ Abweichungen sind *informativ*

Abgeschnitten, linkszensiert, rechtszensiert?



Was ist die abhängige Variable?



- ▶ T : Zufallsvariable (Dauer bis zum Eintritt des Ereignis); t : Zeitpunkt, Realisation von T
- ▶ t beginnt mit dem Beginn der Beobachtung
- ▶ Falls keine truncation/left-censoring mit Beginn der Exposition identisch
- ▶ Beschreibung durch (theoretische) Verteilungen/Funktionen

Was sind die wichtigen Funktionen/Verteilungen?

1. Dichtefunktion von T : $f(t)$
2. Kumulative Verteilungsfunktion: $F(t) = \Pr(T \leq t)$
3. Survivor-Funktion: $S(t) = 1 - F(t) = \Pr(T > t)$
4. **Hazard-Funktion** $h(t)$
5. (Kumulative Hazard-Funktion:
 $H(t) = \int_0^t h(u) du = -\ln(S(t))$)

Verhältnis zueinander?

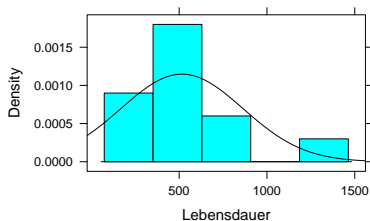
- ▶ Alle vier/fünf Funktionen sind äquivalent
- ▶ Verschiedene Parametrisierungen

Was sagt uns die Dichtefunktion?

- ▶ Alle Objekte starten gleichzeitig ...
 - ▶ Wie wahrscheinlich ist eine bestimmte Überlebenszeit?
 - ▶ Bzw. instantane Wahrscheinlichkeit des Ereignisses (Regierungsende) zu einer bestimmten Zeit?
- ▶ Unkonditional
- ▶ Dichteschätzung über Überlebenszeiten

Was sagt uns die Dichtefunktion?

- ▶ Alle Objekte starten gleichzeitig ...
 - ▶ Wie wahrscheinlich ist eine bestimmte Überlebenszeit?
 - ▶ Bzw. instantane Wahrscheinlichkeit des Ereignisses (Regierungsende) zu einer bestimmten Zeit?
- ▶ Unkonditional
- ▶ Dichteschätzung über Überlebenszeiten



Was sagt uns die kumulative Verteilungsfunktion?

- ▶ Bestimmtes Integral über Dichtefunktion
- ▶ Wieviele Objekte sind zu einem bestimmten Zeitpunkt bereits „gestorben“?
- ▶ Bzw. wie hoch ist die kumulative Wahrscheinlichkeit des Versagens?

Kumulative Verteilungsfunktion

$$F(t) = \Pr(T \leq t) = \int_0^t f(t) dt$$

Was sagt uns die Survivor-Funktion?

- ▶ Wie hoch ist die Wahrscheinlichkeit, bis zu einem bestimmten Zeitpunkt „zu überleben“
- ▶ Bzw. äquivalent: Wieviele der ursprünglichen Objekte sind zu einem bestimmten Zeitpunkt „noch am Leben“ (im Ausgangszustand)

Survivor-Funktion

$$S(t) = 1 - F(t) = \Pr(T > t) = 1 - \int_0^t f(t) dt$$

Was sagt uns die Hazard-Funktion?

- ▶ „Intensity“, „age-specific failure rate“, aktuelles Risiko:
(instantane) *Rate* des Versagens (z. B. Kabinettsauflösung,
wenn Objekt bis dahin durchgehalten hat)
 - ▶ Diskrete Zeit: bezogen auf Zeitraum
 - ▶ Kontinuierliche Zeit: Grenzwert für Länge des Zeitraums $\rightarrow 0$
- ▶ Konditionale Variante der Dichtefunktion

Hazard-Funktion

$$\begin{aligned}h(t) &= \lim_{\Delta t \rightarrow 0} \frac{\Pr(t + \Delta t > T > t | T > t)}{\Delta t} \\ &= \frac{f(t)}{S(t)}\end{aligned}$$

Wie war das mit dem Grenzwert?

► Dichtefunktion:

- Im Intervall $[t = 10; t = 20]$ endet bestimmter Prozentsatz (kumulative Verteilungsfunktion) \rightarrow Risiko des Scheiterns in diesem Intervall
- Wieviel scheitern exakt bei $t = 10$?
- Intervall immer schmaler machen $[10; 11] \rightarrow [10; 10 + 10^{-10}]$
 $\rightarrow [10; 10 + 10^{-100}]$
- Breite des Intervalls $\rightarrow 0$;

Wie war das mit dem Grenzwert?

- ▶ Dichtefunktion:
 - ▶ Im Intervall $[t = 10; t = 20]$ endet bestimmter Prozentsatz (kumulative Verteilungsfunktion) \rightarrow Risiko des Scheiterns in diesem Intervall
 - ▶ Wieviel scheitern exakt bei $t = 10$?
 - ▶ Intervall immer schmaler machen $[10; 11] \rightarrow [10; 10 + 10^{-10}] \rightarrow [10; 10 + 10^{-100}]$
 - ▶ Breite des Intervalls $\rightarrow 0$;
- ▶ Hazard: Wahrscheinlichkeit(sdichte) des Scheiterns an diesem Punkt / Wahrscheinlichkeit(sdichte) noch im Spiel zu sein \rightarrow instantanes Risiko des Scheiterns
- ▶ Niedriges Risiko: man kann auch davor gescheitert sein

Was ist die inhaltliche Bedeutung der Hazard Rate?

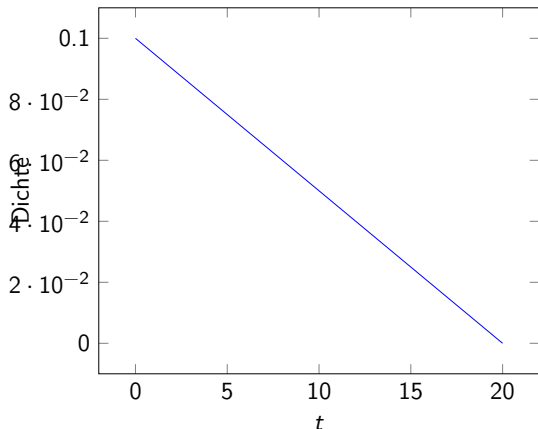
- ▶ Hazard – Risiko zu einem gegebenen *Zeitpunkt* (Grenzwert)
- ▶ Steht im Zentrum moderner Ereignisdatenanalyse
- ▶ Hat eine bestimmte Form
 - ▶ Kann über die Zeit konstant sein, steigen, fallen
 - ▶ Bei *konstantem* Hazard wird das Überleben über die Zeit immer unwahrscheinlicher
 - ▶ Fällt der Hazard auf null, ist das weitere Überleben zunächst gesichert, wenn man es bis hierhin geschafft hat
 - ▶ Menschliche Mortalität hat eine badewannenförmige Hazard-Funktion

Was ist die Bedeutung der kumulativen Hazard-Funktion?

- ▶ Beschreibt das Risiko, das im Lauf der Zeit akkumuliert wird („Umdrehungen“)
- ▶ Wie oft würde ein Objekt (im Mittel) „sterben“, wenn Scheitern wiederholbar?
- ▶ Äquivalent: Überlebenswahrscheinlichkeit, da $S(t) = \exp(-H(t))$
- ▶ Bzw. Wahrscheinlichkeit, *mindestens einmal* zu scheitern ($F(t) = 1 - S(t)$)

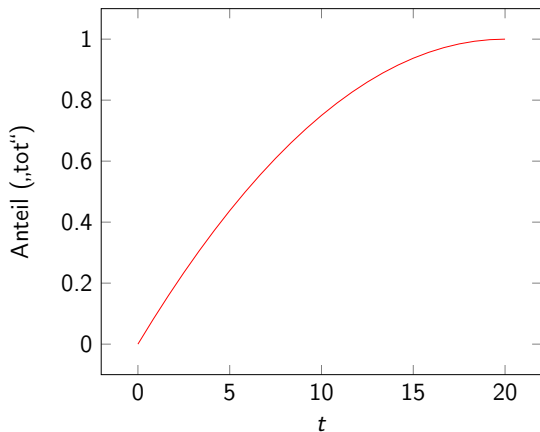
Ein konstruiertes Beispiel

- ▶ lineare Dichtefunktion
- ▶ $f(t) = 0.1 - 0.005 \times t$



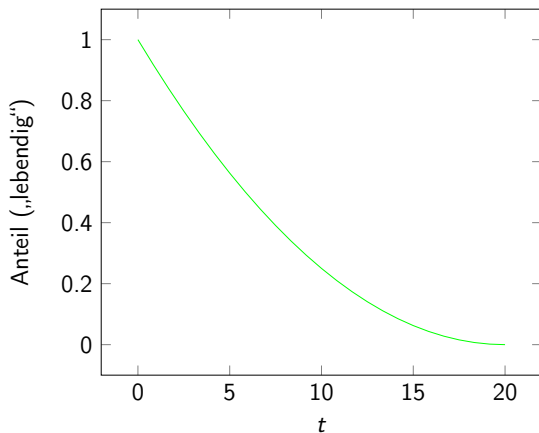
Kumulierte Verteilungsfunktion

► $F(t) = \int_0^t f(t) dt = 0.1 \times t - 0.0025t^2$



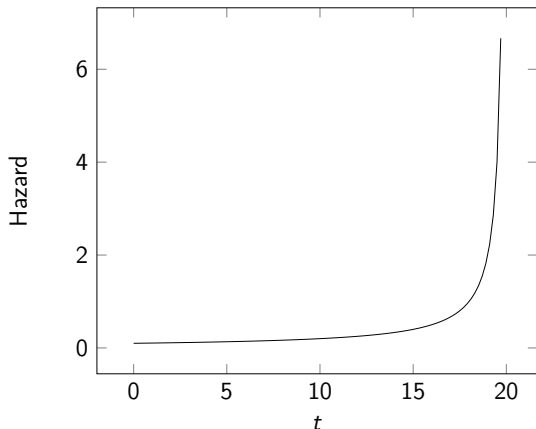
Survivor-Funktion

► $1-F(t)$

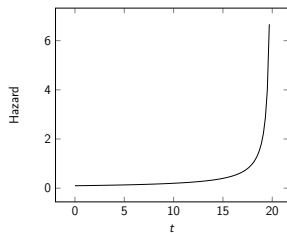
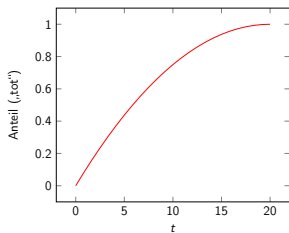
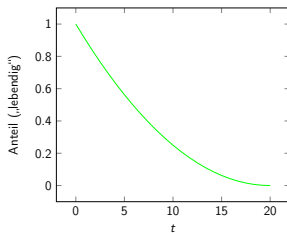
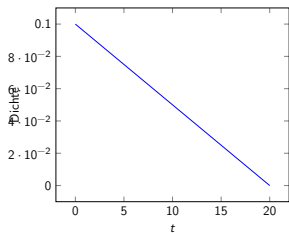


Hazard-Funktion

- ▶ $h(t) \frac{f(t)}{S(t)}$
- ▶ Ausfallrate, bezogen auf überlebende Objekte



Im Überblick



Survivor-, Dichte- und Hazard-Funktion

1. $h(t) = \frac{f(t)}{S(t)}$

2. $f(t) = h(t) \times S(t)$

3. $S(t) = \frac{f(t)}{h(t)}$

- ▶ Moderne Anwendungen *parametrisieren* Modell in Hazard-Variante
- ▶ $h_j(t) = \text{somefunction}(h_0; \beta_0 + \mathbf{x}_j\beta_x)$
- ▶ Fehler ϵ aus dem linearen Modell wandert in Hazard-Funktion
- ▶ Wahl von „somefunction“ = $g(\cdot)$ → Likelihood-Funktion
- ▶ Meistens *proportional*: $h_j(t) = h_0(t) \exp(\beta_0 + \mathbf{x}_j\beta_x)$

Diskrete vs. kontinuierliche Zeit

- ▶ Kann das Ereignis zu einer beliebigen Zeit auftreten?
- ▶ In politikwissenschaftlichen Anwendungen praktisch immer diskrete Zeiten (z. B. Tage)
- ▶ Survivor-Funktion in der Praxis mit „Treppen“
- ▶ Hazard-Rate leichter zu verstehen

Parametrische, semiparametrische, non-parametrische Modelle

- ▶ Parametrisch: Annahmen über Verteilungs-/Hazard-Funktion
 - ▶ „Baseline Hazard“ hat bestimmte Form
 - ▶ Form/Lage durch Kovariaten verändert
 - ▶ Effizienzgewinne
 - ▶ Bias, wenn Annahmen falsch

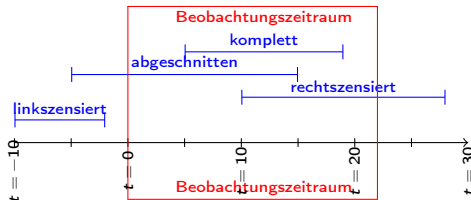
Parametrische, semiparametrische, non-parametrische Modelle

- ▶ Parametrisch: Annahmen über Verteilungs-/Hazard-Funktion
 - ▶ „Baseline Hazard“ hat bestimmte Form
 - ▶ Form/Lage durch Kovariaten verändert
 - ▶ Effizienzgewinne
 - ▶ Bias, wenn Annahmen falsch
- ▶ Semi-parametrisch:
 - ▶ Keine Annahmen über baseline hazard h_0 (Hazard-Funktion)
 - ▶ Aber parametrische Wirkung der x -Variablen
 - ▶ Serie von binären Analysen

Parametrische, semiparametrische, non-parametrische Modelle

- ▶ Parametrisch: Annahmen über Verteilungs-/Hazard-Funktion
 - ▶ „Baseline Hazard“ hat bestimmte Form
 - ▶ Form/Lage durch Kovariaten verändert
 - ▶ Effizienzgewinne
 - ▶ Bias, wenn Annahmen falsch
- ▶ Semi-parametrisch:
 - ▶ Keine Annahmen über baseline hazard h_0 (Hazard-Funktion)
 - ▶ Aber parametrische Wirkung der x -Variablen
 - ▶ Serie von binären Analysen
- ▶ (Non-parametrisch):
 - ▶ Keine Annahmen
 - ▶ Weniger effizient

Was passiert mit zensierten Beobachtungen?



- ▶ Rechtszensierung relativ leicht handhabbar
- ▶ Schätzung der Parameter mittels Maximum Likelihood
- ▶ Likelihood eines Falles: Dichtefunktion
 - ▶ Für vollständig beobachtete Fälle keine Probleme
 - ▶ Rechtszensierte Fälle enthalten Information über das Überleben bis *zum Ende der Beobachtung* → Survivor-Funktion
 - ▶ Für Rechtszensierte Fälle → in Likelihood-Funktion einschließen

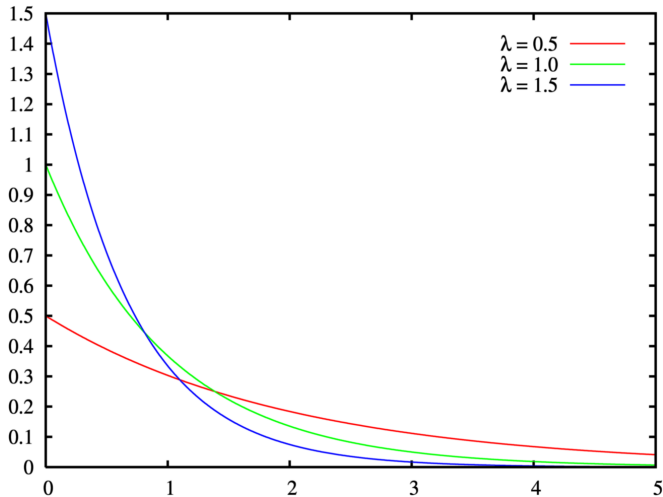
Wie funktioniert das Exponential-Modell?

- ▶ Parametrisierung via Hazard-Rate
- ▶ $h(t) = \exp(\dots)$
- ▶ Exponential-Modell: Hazard-Rate ist konstant (flache Linie), Kovariaten verschieben Niveau der Linie
- ▶ „Gedächtnisloser“ Prozeß → exponentieller Zerfall
- ▶ Survivorfunktion: $\exp(-\lambda t)$
- ▶ Dichtefunktion: $\lambda \exp(-\lambda t)$

Wie funktioniert das Exponential-Modell?

- ▶ Parametrisierung via Hazard-Rate
- ▶ $h(t) = \exp(\dots)$
- ▶ Exponential-Modell: Hazard-Rate ist konstant (flache Linie), Kovariaten verschieben Niveau der Linie
- ▶ „Gedächtnisloser“ Prozeß → exponentieller Zerfall
- ▶ Survivorfunktion: $\exp(-\lambda t)$
- ▶ Dichtefunktion: $\lambda \exp(-\lambda t)$
- ▶ Einfach, aber für uns nicht realistisch

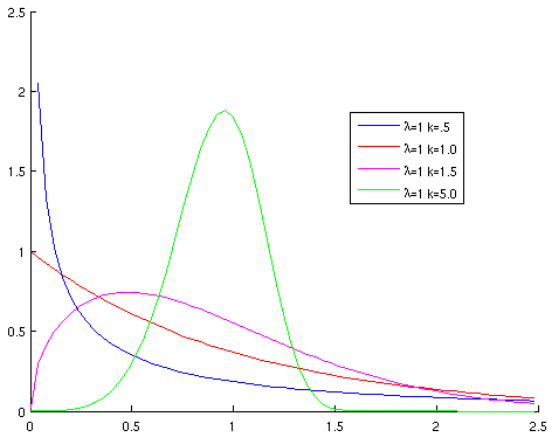
Dichtefunktion (Überlebenszeit) mit verschiedenen λ



Weibull-Modell

- ▶ Eng mit Exponential-Modell verwandt ($\alpha = 1 \rightarrow$ Exponentialverteilung)
- ▶ Weibull-Verteilung
 - ▶ Flexibel
 - ▶ $k > 0 =$ „shape“, $\lambda > 0 =$ „scale“
 - ▶ $f(x; \lambda; k) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} \exp(-(x/\lambda)^k)$
 - ▶ Entspricht konstanten, steigendem oder fallendem Hazard
- ▶ Parametrisierung: $h(t) = \exp(\mathbf{X}\boldsymbol{\beta} + \alpha \ln(t))$
- ▶ α legt die Form der Verteilung fest

Weibull: Mögliche Dichteverteilungen



Was gibt es sonst noch? Wie werden die Modelle parametrisiert?

- ▶ Flexiblere Verteilungen als Weibull (z. B. generalised Gamma)
- ▶ Aber wir haben meistens keine starken Annahmen über Form des Hazard → relativ uninteressant für uns
- ▶ Verschiedene Programme, verschiedene Parametrisierungen
- ▶ Zwei Varianten:
 1. Log-lineares Modell (Accelerated Failure Time)
 - ▶ Logarithmus der Zeit bis failure als abhängige Variable
 - ▶ Linearer Prädiktor inkl. ϵ (nicht normalverteilt)
 2. Proportional Hazard: Veränderung des instantanen Risikos

Wie sieht das Cox Proportional Hazard Model aus?

- ▶ $h(t) = h_0(t) \exp(\mathbf{X}\beta)$
- ▶ **Baseline-Hazard**: wird nicht-parametrisch aus den Daten geschätzt
- ▶ **Proportional**: Kovariaten haben multiplikativen Effekt auf Baseline-Hazard
- ▶ Basiert auf einem partiellen Likelihood-Verfahren
- ▶ Erweiterung ermöglicht Umgang mit variierenden Kovariaten

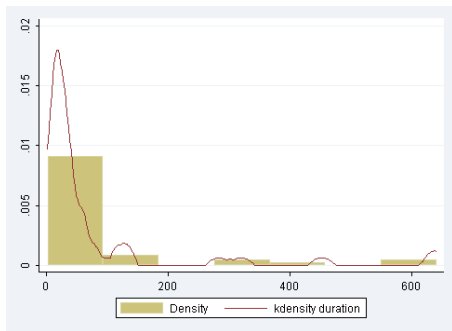
Mögliche Probleme mit dem Cox-Modell?

1. Unpräziser, vor allem bei kleinen Datensätzen
2. Ties problematisch (wenn häufig)
3. Nicht angemessen, wenn Form der Zeitabhängigkeit von Interesse
4. Schwächere Theorie

Was gibt es?

- ▶ Früher: Spezialisierte Software
- ▶ Heute: Unterstützung in sehr vielen Standardpaketen
- ▶ R + Stata (Buch)

Dauer von 54 UN-Missionen 1948-2001



- ▶ $\bar{y} = 74$, Median 25.5, aber 15 noch nicht abgeschlossen (rechtszensiert)
- ▶ Korrigierte Schätzmethoden: Median 30 Monate, Mittelwert 228 Monate
- ▶ Drei Konflikttypen: civil war, interstate conflict, internationalised civil war

Weibull: Hängt Dauer vom Konflikttyp ab?

Weibull regression -- log relative-hazard form

No. of subjects =	54	Number of obs =	54
No. of failures =	39		
Time at risk =	3994		
Log likelihood =	-84.655157	LR chi2(2) =	17.67
		Prob > chi2 =	0.0001

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
civil	.8879245	.3832017	2.32	0.020	.1368629	1.638986
interst	-1.401441	.5117817	-2.74	0.006	-2.404515	-.3983673
_cons	-3.459909	.4952858	-6.99	0.000	-4.430652	-2.489167
/ln_p	-.2145617	.1237889	-1.73	0.083	-.4571834	.02806
p	.806895	.0998846			.6330642	1.028457
1/p	1.239319	.1534138			.97233	1.579619

Was bedeutet das?

- ▶ $p < 1$ Hazard reduziert sich mit der Zeit (Signifikanz?)
- ▶ Relative Hazard Parametrisierung
 - ▶ Für Referenzkategorie ist hazard = $\exp(-3.46)0.8t^{0.8-1} \exp(0)$
 - ▶ Für interstate conflict ist hazard = $\exp(-3.46)0.8t^{0.8-1} \exp(-1.4 \times 1)$
 - ▶ $\exp(-1.4) = 0.25$, d. h. hazard ist zu jedem Zeitpunkt 75% niedriger
 - ▶ Civil war: $\exp(0.89) = 2.44 \rightarrow$ hazard 144% höher
- ▶ Einsätze bei interstate conflict dauern am längsten, bei Bürgerkriegen am kürzesten (höheres Risiko des „Scheiterns“
→ schnelleres Ende)

Weibull: Accelerated failure time

Weibull regression -- accelerated failure-time form

No. of subjects =	54	Number of obs =	54
No. of failures =	39		
Time at risk =	3994		
Log likelihood =	-84.655157	LR chi2(2) =	17.67
		Prob > chi2 =	0.0001

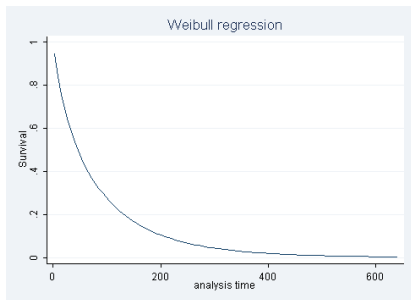
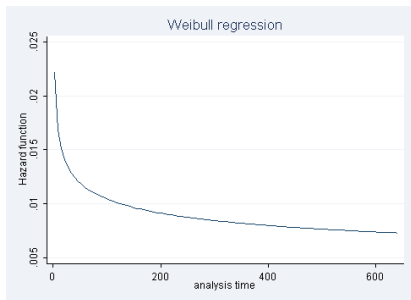
_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
civil	-1.100421	.4457861	-2.47	0.014	-1.974146	-.2266966
interst	1.736832	.6165459	2.82	0.005	.5284242	2.94524
_cons	4.28793	.2652436	16.17	0.000	3.768062	4.807798
/ln_p	-.2145617	.1237889	-1.73	0.083	-.4571834	.02806
p	.806895	.0998846			.6330642	1.028457
1/p	1.239319	.1534138			.97233	1.579619

Was bedeutet das?

- ▶ Andere Parametrisierung,
- ▶ Identische Likelihood, identische Ergebnisse (weil $h(t) \Leftrightarrow f(t)$)
- ▶ Umgekehrte Vorzeichen: niedrigerer hazard – längere Zeitdauer
- ▶ Civil war < internationalised civil war < interstate conflict

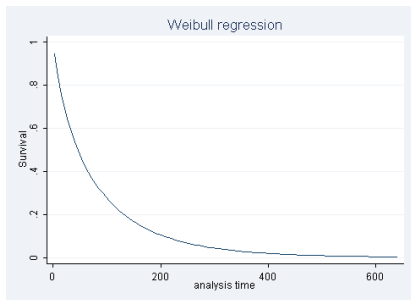
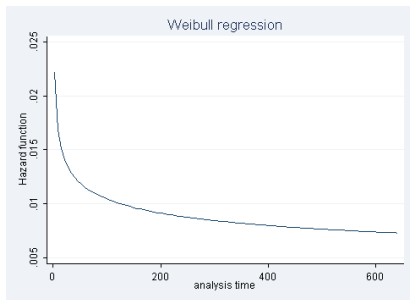
Geht das besser?

Geht das besser?



- ▶ Geschätzte hazard-/survival-Funktionen für internationalised civil war

Geht das besser?



- ▶ Geschätzte hazard-/survival-Funktionen für internationalised civil war
- ▶ Unter der Annahme, daß Zeiten mit geschätzten Parametern Weibull-verteilt sind

Welche Ergebnisse bringt das Cox model?

Cox regression -- Breslow method for ties

```
No. of subjects =          54          Number of obs =          54
No. of failures =          39
Time at risk    =          3994
Log likelihood  = -127.15763          LR chi2(2)    =          8.93
                                          Prob > chi2    =          0.0115
```

```
-----+-----
      _t |      Coef.  Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
      civil |   .7348046   .3781278    1.94   0.052   - .0063122   1.475921
      interst |  -.8556111   .5042314   -1.70   0.090   -1.843886   .1326643
```


Bzw.

Cox regression -- Breslow method for ties

```
No. of subjects =          54          Number of obs =          54
No. of failures =          39
Time at risk   =        3994
Log likelihood =   -127.15763
LR chi2(2)     =          8.93
Prob > chi2    =          0.0115
```

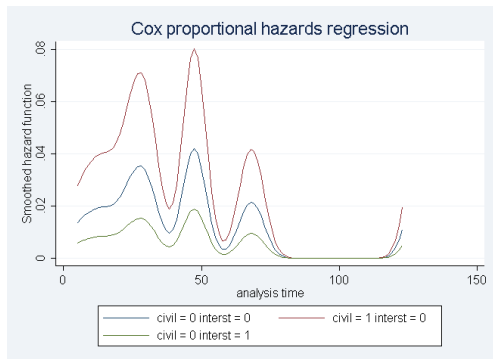
```
-----+-----
      _t | Haz. Ratio  Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
      civil |  2.085075   .7884246    1.94   0.052   .9937077   4.375065
      interst | .4250234   .2143101   -1.70   0.090   .1582014   1.141867
```

- ▶ $\exp(\text{Koeffizient}) = \text{hazard ratio (Voreinstellung)}$

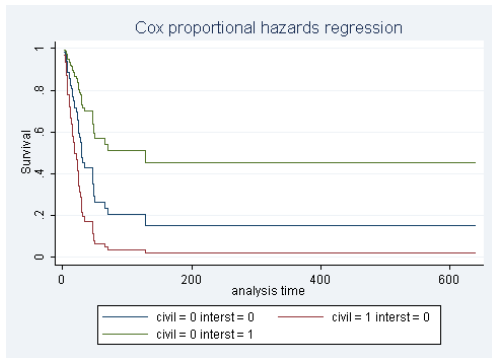
Was bedeutet das?

- ▶ Keine Konstante – keine Annahme über Form des base hazard
- ▶ **Aber:** Form des hazard für alle Fälle gleich!
- ▶ Hazard für civil war ca. zweimal größer als baseline hazard
- ▶ Hazard für interstate conflict weniger als halb so groß wie baseline hazard
- ▶ Identische Reihenfolge: civil war < internationalised civil war < interstate conflict

Geschätzte Hazard-Raten (geglättet)



Geschätzte Survivor-Funktionen



- ▶ Step-Funktionen
- ▶ Alle Missionen > 128 Monate rechtszensiert

Zusammenfassung

- ▶ Vielzahl von Besonderheiten bei der Analyse von Ereignisdaten
- ▶ „Normale“ Modelle führen fast unweigerlich in die Irre
- ▶ **Essentiell**: Unterschiede und Beziehungen zwischen $f(t)$, $S(t)$, $h(t)$
- ▶ Vielzahl von Analysemöglichkeiten, dynamisches Feld

Literatur für nächste Woche

- ▶ Cross-level inference – Schluß von einer Ebene (z. B. Aggregatdaten) auf eine niedrigere Ebene (Mikro)
- ▶ Problematisch
- ▶ Literaturempfehlung: Achen/Shively, Cross-Level Inference, Chicago 1995: Einleitung