LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Zusammenfassung Survival - Analyse

Institut für Statistik Bachelor Seminar SS 2017:

„Moderne statistische Methoden in der Epidemiologie"

Referent: Rina Nicaj

Betreuung: Dr. Michael Schomaker

13.06.2017

Gliederung

1. *Einleitung*
2. *Grundbegriffe der Lebenszeitanalyse*
3. *Zensierte Daten*
4. *Der Kaplan-Meier-Schätzer*
5. *Das Cox-Modell und Parametrische Modelle*
6. *Zusammenfassung und Fazit*

Gliederung

1. *Einleitung*
2. *Grundbegriffe der Lebenszeitanalyse*
3. *Zensierte Daten*
4. *Der Kaplan-Meier-Schätzer*
5. *Das Cox-Modell und Parametrische Modelle*
6. *Zusammenfassung und Fazit*

Survival Analysis

Die Ursprünge der Survival-Analyse kann man seit siebzehnten Jahrhundert aufspüren.

Die Sterbetafel Methode wurde von Aktuaren, Statistikern und Biomedizinischen Forschern eingesetzt.

Nach dem Krieg wurde der Term „ Lifetime Analysis“ zu „Survival Analysis“ geändert.

Nicht nur in medizinischen und biologischen Studien, sondern auch im Ingenieurwesen, der Ökonomie oder bei den Sozialwissenschaften spielt die statistische Analyse eine wichtige Rolle.

Survival Analysis

Die Survival-Analyse modelliert Überlebenszeiten.

→ *wie viel Zeit von einem Startzeitpunkt bis zum Auftreten eines bestimmten Ereignisses vergangen ist.*

Bei der Analyse von Überlebenszeiten kann sowohl die Form ihres grundsätzlichen Verlaufs von Interesse sein, als auch inwiefern ihr Verlauf systematisch von Einflussgrößen abhängt.

Survival Analysis

Beispiele der Survival-Analyse

→ *Zeitdauer, die ein Patient nach einer Behandlung weiter am Leben ist, bis ein bestimmtes Behandlungsgerät im Gebrauch einen Defekt aufweist.*

→ *Die Dauer, die ein Kleinkind benötigt, um ein vordefiniertes Entwicklungsziel zu erreichen*

→ *Arbeitslosigkeit bis zur Neueinstellung*

Grundbegriffe der Lebenszeitanalyse

1. *Einleitung*

2. *Grundbegriffe der Lebenszeitanalyse*
   - *Die Survival-Funktion*
   - *Die Hazard-Funktion*
   - *Die erwartete Restlebensdauer*

3. *Zensierte Daten*

4. *Der Kaplan-Meier-Schätzer*

Die Survival - Funktion

→ *Beschreibung von Lebenszeiten*

Die Survival-Funktion S zur Ausfallzeit T ist definiert

$$S(t) = P(T > t) = 1 - F(T) \quad , t \geq 0$$

S(t) → *Die Überlebensfunktion*

t → *Zeitpunkt*

T → *Die zufällige Variable, Überlebenszeit*

F(T) → *Verteilungsfunktion von T*

Eigenschaften der Survival - Funktion

→ *S(t) = 1− F(t), t ≥ 0*

→ *S ist eine monoton fallende Funktion mit S(0) = 1 und* $\lim t \to \infty$ *S(t) = 0*

→ *Ist T stetig verteilt mit positiver Dichte f auf (0,∞) , so ist S streng monoton fallend und es gilt :*

$$S(t) = \int_t^{\infty} f(u)\,du$$

→ *ist T diskret verteilt mit Werten in* $0 < t_1 < t_2 < \ldots$ *, d.h. diskreter Dichte* $p(t_j) = P(T = t_j)$*, j = 1,2 , ... , so ist S eine monoton fallende rechts-stetige Treppenfunktion:*

$$S(t) = \sum_{t_j > t} pt_j$$

Hazard Funktion

Die Hazard-Funktion ist für die Darstellung von Lebensverteilung geeignet

→ *wie sich das Ausfallrisiko in Abhängigkeit vom Alter im Verlauf der Zeit verändert*

Für stetige Überlebenszeit T folgt:

$$\lambda(t)=\lim_{\Delta\to 0}\left(\frac{P(t\leqslant T<t+\Delta|T\geqslant t)}{\Delta}\right), t\geqslant 0$$

Für diskrete Überlebenszeit T folgt:

$$\lambda(t_j)=P(T=t_j|T\geqslant t_j),\ j=1,2,3\ldots$$

Hazard Funktion

Kumulierte Hazard-Funktion

→ *für eine stetig verteilte Lebenszeit T*

$$\Lambda(t) = \int_0^t \lambda(u)\,du$$

→ *für eine diskret verteilte Lebenszeit T*

$$\Lambda(t) = \sum_{tj \leqslant t} \lambda(t_j)$$

Hazard Funktion

In der parametrischen Überlebenszeitanalyse verwendet man die Hazard-Funktion zur Bestimmung der Ausfallverteilung.

→ *qualitative Informationen über den Ausfallmechanismus*

Eine wachsende Hazard-Rate ergibt sich im Zusammenhang mit natürlichen Alterung.

Fallende Hazard-Funktionen sind weitaus weniger üblich, finden jedoch gelegentlich Gebrauch bei sehr frühen Sterbewahrscheinlichkeiten.

Die erwartete Restlebensdauer

→ *Für Individuen des Alters t gibt sie an, welche restliche Lebensdauer im Mittel noch verbleibt.*

Die erwartete Restlebensdauer ist definiert als

$$mrl(t) = E\,(T - t \mid T > t), \quad t \geq 0$$

Eigenschaften der erwartete Restlebensdauer

→ *Ist die Zufallsvariable T stetig verteilt, so gilt:*

1)

$$mrl(t) = \frac{1}{S(t)} \cdot \int_t^\infty S(u)\,du\,, t \geqslant 0$$

2) $E(T) = mrl(0)$

$$Var(T) = 2\int_0^\infty tS(t)\,dt - \left(\int_0^\infty S(t)\,dt\right)^2$$

Zensierte Daten

1. Einleitung

2. Grundbegriffe der Lebenszeitanalyse

3. Zensierte Daten

- *Rechts-zensierte*
- *Links-zensierte*
- *Intervallzensur*

4. Der Kaplan-Meier-Schätzer

Zensierte Daten

Der Erhebungszeitraum ist oft begrenzt, sodass nicht für alle Untersuchungseinheiten das Ereignis auch tatsächlich innerhalb des Beobachtungszeitraums auftritt.

Um zensierte Daten geeignet auswerten zu können, ist es wichtig zu wissen durch welchen Mechanismus sie hervorgerufen werden.

Drei Hauptarten der Zensur sind :

1. *Rechts-zensierte*
2. *Links-zensierte*
3. *Intervall-zensierten*

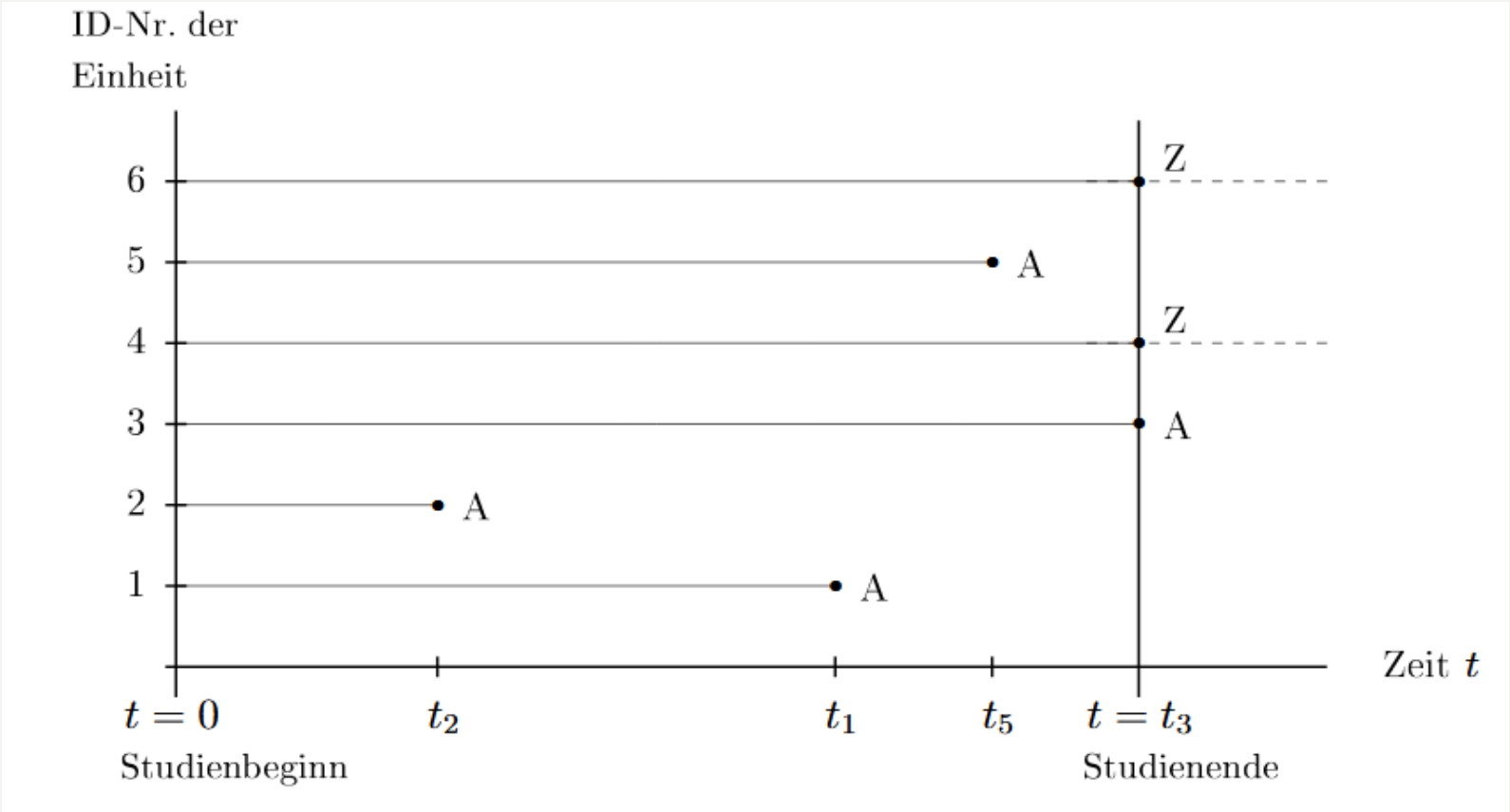
Rechts-zensierte

Es gibt zwei verschiedene Rechts-Zensuren:

1. *Die Überlebenszeit des letzten Beobachtungszeitpunkts überschreitet das Ende der Studie, nicht aber den exakten beobachteten Wert.*

2. *Ein frühzeitiger Abbruch aus der Studie*
   → Umzug, Tod ect.

Rechts-zensierte



Links-zensierte und Intervall-zensierte

Links-zensierte

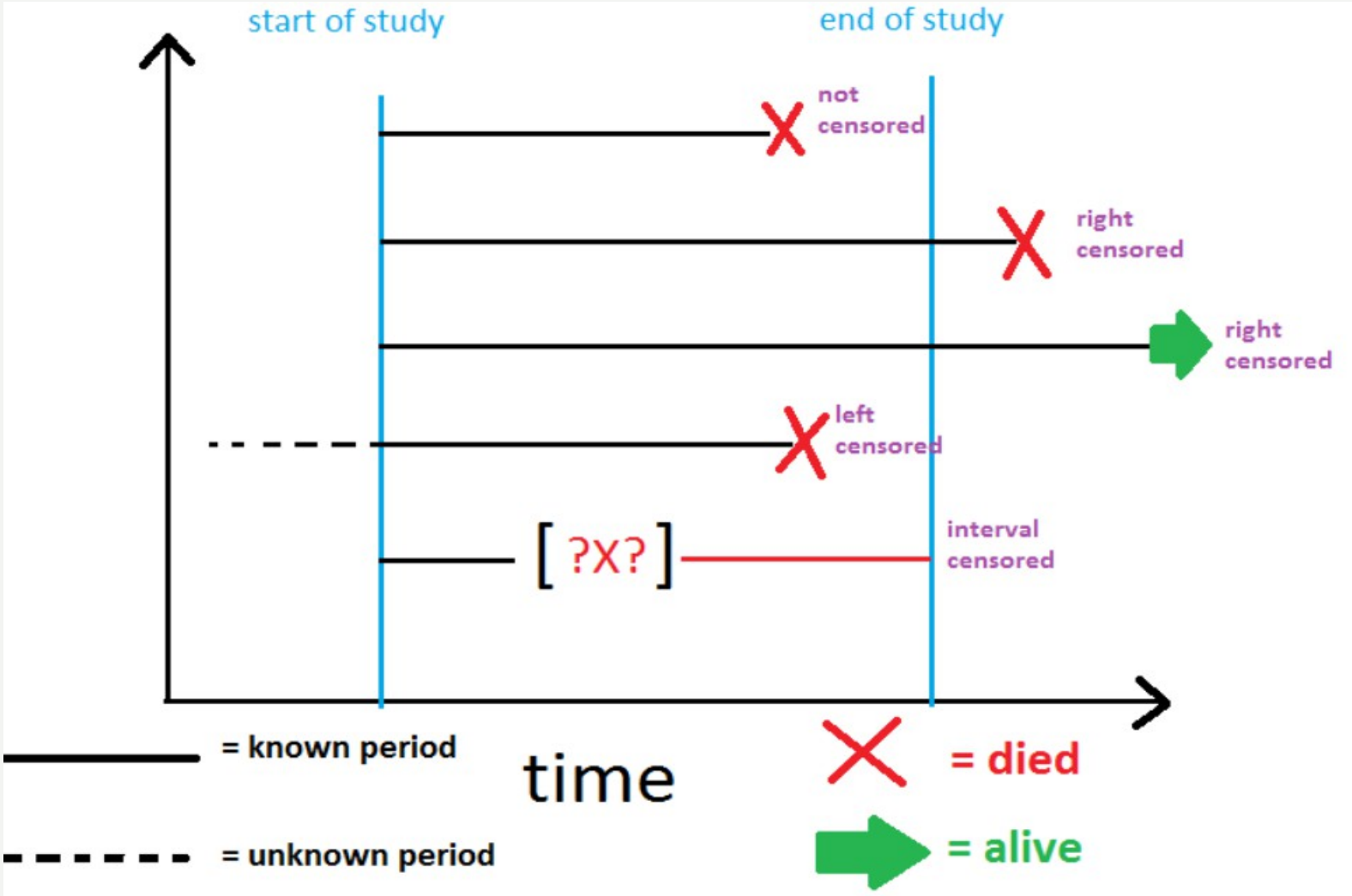
→ Das interessierende Ereignis ist bereits vor Beginn der Studie eingetreten

→ Eher bei empirischen Untersuchungen

Intervall-zensierte

→ Tritt das Ereignis unbeobachtet zwischen zwei Zeitpunkten a und b so spricht man von Intervall-zensierten Daten

Zensierte Daten



Der Kaplan – Meier - Schätzer

1. *Einleitung*

2. *Grundbegriffe der Lebenszeitanalyse*

3. *Zensierte Daten*

4. *Der Kaplan-Meier-Schätzer*

- *Der Kaplan-Meier-Schätzer*
- *Der KM-Schätzer als LM-Schätzer*
- *Die Varianz und Konfidenzintervall des KM-Schätzers*

Der Kaplan – Meier - Schätzer

Der Kaplan-Meier Schätzer schätzt die Wahrscheinlichkeit, dass bei einem Versuchsobjekt ein bestimmtes Ereignis innerhalb eines Zeitintervalls nicht eintritt.

→ *Nicht parametrisches Schätzer*

$$\hat{S}(t) = \prod_{t_i \leq t} [1 - \frac{d_i}{n_i}]$$

t_i → *ist die Dauer der Studie am Punkt i*

d_i → *ist die Zahl der Todesfälle bis Punkt i*

n_i → *ist die Anzahl der Personen in Gefahr kurz vor* t_i

Der Kaplan – Meier - Schätzer

Annahmen des KM - Schätzers

- *Zensierte Personen haben die gleiche Aussicht auf Überleben wie diejenigen, die weiterhin verfolgt werden.*
- *Überlebensaussichten sind gleich für früh wie für späte Rekruten in der Studie*
- *Das untersuchte Ereignis passiert zum angegebenen Zeitpunkt*.

Der KM–Schätzer als LM Schätzer

Unterliegt der Datensatz $(y, \delta) = ((y_1, \delta_1), \ldots, (y_n, \delta_n))$ *einem der beschreibenden Zensur-Mechanismen, die auf basierenden Likelihood Funktion beruhen, so gilt:*

$$L(P_t; (y, \delta)) = \prod_{i=1}^{n} f_{Ti}(y_i)^{\delta_i} S_{Ti}(y_i)^{1-\delta_i}$$

Likelihood Funktion durch Hazard-Funktion

$$L(P_t; (y, \delta)) = \prod_{j=1}^{k} \lambda(z_j)^{d_j} [1 - \lambda(z_j)^{(n_j - d_j)}]$$

Der Varianz der KM - Schätzer

Eine Schätzung für die Varianz von KM Schätzer erhält man mittels der Large-Sample - Theorie für ML - Verfahren

$$Var\,\hat{S}(t) = \hat{S}(t)^2 \sum_{t_i \leq t} \left(\frac{d_i}{n_i (n_i - d_i)} \right)$$

Die Approximation für die Varianz des Kaplan-Meier-Schätzers ist als Greenwood-Formel bekannt.

Konfidenzintervall der KM - Schätzer

Das Konfidenzintervall für die Überlebensfunktion wird nicht direkt mit der Varianzschätzung von Greenwood berechnet

→ *unmögliche Ergebnisse*

Ist der $\hat{S}(t)$ *ML-schätzer, so ist er unter schwachen Regularitätsbedingungen asymptotisch normalverteilt*

$$[\hat{S}(t_0)-\Phi^{-1}(1-(\frac{\alpha}{2}))\hat{\sigma}_s(t_o), \hat{S}(t_0)+\Phi^{-1}(1-(\frac{\alpha}{2}))\hat{\sigma}_s(t_o)]$$

Bessere Kondenzintervalle ergeben sich mittels Der log-logTransformation

$$[\hat{\psi}(t_0)-\Phi^{-1}(1-(\frac{\alpha}{2}))\hat{\sigma}_\psi(t_o), \hat{\psi}(t_0)+\Phi^{-1}(1-(\frac{\alpha}{2}))\hat{\sigma}_\psi(t_o)]$$

Das Cox Modell und Parametrische Modelle

4. *Der Kaplan-Meier-Schätzer*

5. *Das Cox-Modell und parametrische Modelle*
   Das Cox-Modell
   → *Modellierung*
   → *Interpretation geschätzte Parameter*
   Parametrische Modelle
   → *Spezielle Verteilung*
   → *Das Accelerated-Failure-Time Modell*

6. *Zusammenfassung und Fazit*

Das Cox Modell

Das Cox-Modell ist eine Methode, um die Wirkung von mehreren Variablen auf die Zeit bis zu einem bestimmtem Ereignis zu untersuchen.

→ *Mit semiparametrischen Verfahren*

→ *Ist an keinen speziellen Verteilungstyp gebunden*

→ *Kovariablen wirken sich direkt auf die Hazard-Rate eines Individuums aus*

Das Cox Modell

Annahmen der Cox Modell:

- *Nicht-informative Zensur*
- *Hazard-Rate zweier Individuen mit verschiedene Kovariablenwerten muss proportional zueinander sein.*

Vorausgesetzt, dass die Annahmen der Cox-Regression erfüllt sind, wird diese Funktion bessere Schätzungen der Überlebenswahrscheinlichkeiten und der kumulativen Hazard liefern, als die der Kaplan-Meier-Funktion.

Das Cox Modell

Sei T ≥ 0 eine Lebensdauer und $X=(X_1,\dots,X_p)'$ *ein p dimensionaler Vektor von erklärenden Variablen. Das Cox-Hazard-Modell postuliert, dass die Form der Hazard-Funktion in Abhängigkeit von X*

$$\lambda(t|X)=\lambda_0(t)\exp(\beta' X), t\geqslant 0$$

λ_0 → *Basis-Hazard-Funktion*

$\beta=(\beta_1,\dots,\beta_p)'$ → *ein Vektor von Regressionskoeffizienten falls X=0*

Das Cox Modell

Im Cox-Hazard-Modell lässt sich die Survival-Funktion bei stetiger Lebenszeit darstellen:

$$S(t|X) = S_0(t)^{\exp(\beta' X)}, t \geqslant 0$$

Das Cox Modell

Interpretation geschätzter Parameter

→ *Die Zeitunabhängigkeit des Risikoverhältnisses zweier unterschiedliche Beobachtung*

$$HR(t, x, \bar{x}) = \frac{\lambda(t|x)}{\lambda(t|\bar{x})} = \exp(\beta'(x - \bar{x})), t \geq 0$$

Das Cox Modell

Sind im Cox-Hazard-Modell stetige Kovariablen enthalten, so erfolgt die Interpretation der entsprechenden Regressionsparameter über konstante Intervalle.

Handelt es sich bei der stetigen Variable X um die einzige erklärende beobachtete Variable des Modells für ein Individuum des Werts X = x so ist seine Hazard-Rate

$$\lambda(t|X) = \lambda_0(t)\exp(\beta X), t \geqslant 0$$

Und für Hazard-Verhältnis

$$HR(t, x+c, x) = \exp(c\beta), t \geqslant 0$$

Parametrische Modelle

Parametrische Modelle bieten Alternativen zum Cox-Modell an, wenn die proportionalen Hazard Erhebung im Frage steht.

Wenn die Verteilungsannahme über die Überlebenszeiten gültig sind, dann werden Schätzungen der Parameter effizienter

→ *kleineren Standardfehlern im Vergleich zu nichtparametrischen Modellen.*

Parametrische Modelle

Exponential - Verteilung

→ *Grundmodell für die Überlebenszeit.*

Aufgrund ihrer Gedächtnislosigkeit und der daraus resultierenden konstanten Hazard-Rate ist die Exp-Verteilung in der modernen Lebenszeitanalyse allerdings nur sehr begrenzt anwendbar.

Dichte-Funktion $f(t) = \lambda \exp(-\lambda t)$

Survival-Funktion $S(t) = \exp(-\lambda t)$

Hazard-Funktion $\lambda(t) = \lambda$

Parametrische Modelle

Weibull - Verteilung $T \sim Weib(\lambda, \beta)$

→ *liefert ein Lebenszeit-Modell das in vielen verschiedenen Bereichen eingesetzt wird.*

Dichte - Funktion

$$f(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right]$$

Survival - Funktion

$$S(t) = \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right]$$

Hazard - Funktion

$$\lambda(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}$$

Parametrische Modelle

	Dichte-Funktion	*Survival-Funktion*	*Hazard-Funktion*
Exponential - Verteilung	$\lambda \exp(-\lambda t)$	$\exp(-\lambda t)$	λ
Weibull - Verteilung	$\frac{\beta}{\alpha}(\frac{t}{\alpha})^{\beta-1} \exp[-(\frac{t}{\alpha})^{\beta}]$	$\exp[-(\frac{t}{\alpha})^{\beta}]$	$\frac{\beta}{\alpha}(\frac{t}{\alpha})^{\beta-1}$
Log-Normal-Verteilung	$\frac{1}{(2\pi)^{1/2}\sigma t} \exp\left\{-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right\}$	$1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)$	
Log-Logistik-Verteilung	$\frac{(\frac{\beta}{\alpha})(\frac{t}{\alpha})^{\beta-1}}{[1+(\frac{t}{\alpha})^{\beta}]^2}$	$\left[1+\left(\frac{t}{\alpha}\right)\right]^{-1}$	$\frac{(\frac{\beta}{\alpha})(\frac{t}{\alpha})^{\beta-1}}{1+(\frac{t}{\alpha})^{\beta}}$

Das AFT Modell

Dient für zeitunabhängige Kovariablen

→ *parametrisches Regressionsmodell der sich erklärende Variablen auf das Vergehen der Lebenszeit eines Individuums auswirken.*

Sei T ≥ 0 eine Lebensdauer und $X=(X_1,\dots,X_p)'$ *ein p-dimensionaler Vektor von erklärenden Variablen.*

$$S(t|X)=S_0[\exp(\gamma' X)t], t\geqslant 0$$

$\gamma=(\gamma_1,\dots,\gamma_p)'$ → Vektor von Regressionskoeffizienten

S_0 → Survival-Funktion eines Individuums mit X = 0

$\exp(\gamma' X)$ → Beschleunigungsfaktor

Das AFT Modell

Mit Hilfe der Basis-Hazard-Funktion λ_0 und des Basis- p- Quantils $t_{0\,p}$ können im AFT-Modell die Hazard-Funktion und das p-Quantil zu einem beliebigen Kovariablenvektor X dargestellt werden :

$$\lambda(t|X) = \exp(\gamma' X)\lambda_0[\exp(\gamma' X)t], t \geq 0$$

$$t_0(X) = \exp(-\gamma' X)t_{0\,p}$$

Das AFT Modell

Ist T > 0 eine Lebensdauer Y=lnT und $X=(X_1,\dots,X_p)'$ *ein p-dimensionaler Vektor von erklärende Variablen, so wird im AFT- Modell der Einfluss der Kovariablen auf die Lebenszeit durch folgenden Zusammenhang beschreiben:*

$$Y=\ln(T)=\beta_0+\beta' X+bZ$$

$\beta=(\beta_1,\dots,\beta_p)'$ → Regressionsvektor

Z → Fehlverteilung

Das AFT Modell

Die Likelihood Funktion für die Stichprobe $((y_1, \delta_1, x_1), \ldots, (y_n, \delta_n, x_n))$ *ist im log-linearen Modell mit* $u(x) = \beta_0 + \beta' x$

$$L(u(x), b) = \prod_{i=1}^{n} [(\frac{1}{b}) f_{zi}(\frac{y_i - u(x_i)}{b})]^{\delta_i} S_{zi}(\frac{y_i - u(x_i)}{b})^{1-\delta_i}$$

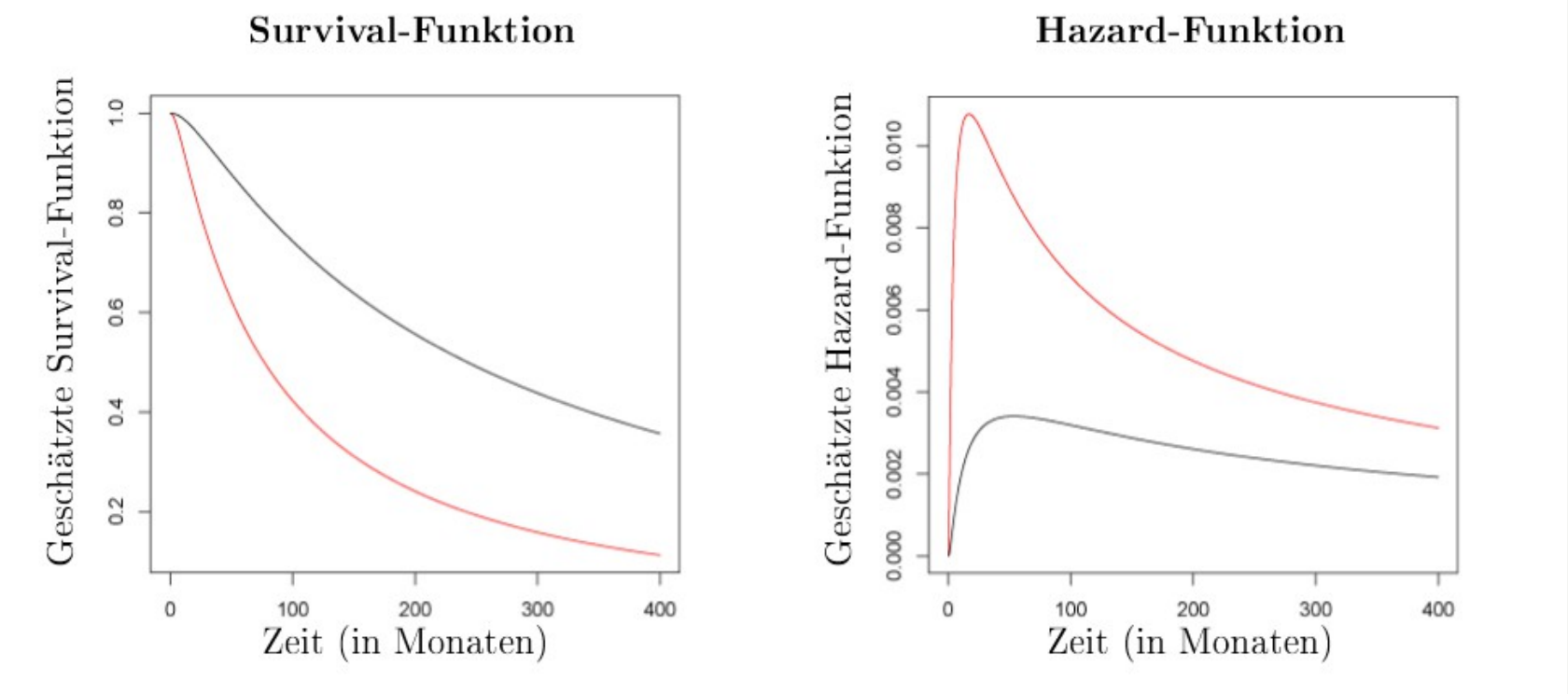
f_{zi} → Wahrscheinlichkeitsdichte

S_{zi} → Die Survival-Funktion der Fehlerwertverteilung

Das AFT Modell

Das Log-Normal-AFT-Modell bei einer log-normal-verteilten Zufallsvariable T enthält man das log-lineare Modell

$$\ln(T) = \beta_0 + \beta' X + bZ$$



Gliederung

1. *Einleitung*
2. *Grundbegriffe der Lebenszeitanalyse*
3. *Zensierte Daten*
4. *Der Kaplan-Meier-Schätzer*
5. *Das Cox - Modell und Parametrische Modelle*
6. *Zusammenfassung und Fazit*

Zusammenfassung

Bei der Survival- , Hazard-Funktion und erwartete Lebensdauer ist, falls eine der eingefügten Größen bekannt ist, sind die anderen zwei Größen eindeutig bestimmbar.

Die meisten Überlebensanalyse-Methoden konzentrieren sich auf die rechts zensierte Daten, da diese häufiger als links zensiert Daten auftreten.

Das Cox-Hazard-Modell hat gegenüber dem AFT Modell den Vorteil

Parametrische Modelle liefern bessere Schätzer für die interessierenden Größen