Ereignisanalyse

Petra Stein / Marcel Noack

12. Juli 2007

Inhaltsverzeichnis

1 Einleitung **3**

2 Grundlagen **6**
2.1 Regression für Längschnittdaten? 6
2.2 Unterscheidungen: A vs. B 7
2.2.1 Distributional- vs. Regressionsansätze 7
2.2.2 Repeated vs. nonrepeated events 7
2.2.3 Einzelne Ereignisse vs. multiple Fälle von Ereignissen . 7
2.2.4 Parametrische vs. nichtparametrische Methoden 8
2.2.5 Diskrete vs. stetige/kontinuierliche Zeit 8
2.3 Begriffe 8
2.3.1 Zustände - State 9
2.3.2 Ereignis - Event 9
2.3.3 Verweildauer - Duration 9
2.3.4 Risiko-Periode - Risk Period 9
2.3.5 Risikomenge - Risk Set 10
2.4 Mathematische Grundlagen 10
2.4.1 Dichtefunktion $f(t)$ & Verteilungsfunktion $F(t)$ 10
2.4.2 Survivalfunktion $S(t)$ 12
2.4.3 Hazardrate $h(t)$ 13
2.4.4 Verknüpfungen 14
2.5 Zensierung 15

3 Diskrete Zeit **17**
3.1 Discrete Time Logit Models 17
3.1.1 Approximation stetiger Zeit durch Modelle mit diskreter Zeit 17
3.2 Mathematische Konzepte 18
3.3 Logitmodell für diskrete Zeit 19
3.4 Deskriptiv: Nichtparametrische Verfahren 21
3.4.1 Life Table Methode: Verweildauer in Intervallen 22

3.4.2 Product-Limit Estimation / Kaplan-Meier 25
3.4.3 Nachteile nichtparametrischer Verfahren 28

4 Stetige Zeit **30**
4.1 Parametrische Modelle der Zeitabhängigkeit 30
4.1.1 Exponential Hazard Rate Models 31
4.1.2 Piecewise Constant Exponential Models 33
4.1.3 Weibull-Modell 35
4.1.4 Gompertz-Makeham Modell 36
4.1.5 Intermezzo I 38
4.1.6 Log-Logistisches Modell 39
4.1.7 Log-Logistsches Standardmodell 39
4.1.8 Log-Normale Modelle 40
4.1.9 Intermezzo II 41
4.1.10 Sichelmodell / Sickle-Model 43
4.1.11 Letzte parametrische Bemerkung 44
4.2 Semi-Parametrische Modelle: die Regression nach Cox 45
4.2.1 Cox-Modell, Notation nach Allison 46
4.2.2 Cox-Modell, Notation nach Yamaguchi 47

A Variablen: diskret & stetig **49**
A.1 Diskret Variablen 49
A.2 Stetige Variablen 52

B Dichtefunktion & Verteilungsfunktion **54**

C Grundlagen der Analysis **55**
C.1 Ausgangsfunktion $f(x)$ 55
C.2 Stammfunktion $F(x)$ 56
C.3 Erste Ableitung $f'(x)$ 58
C.4 Beispiel einiger Funktionen 60
C.4.1 Beispiel: Integration von $\frac{3}{4\sqrt{x^5}}$ 62
C.4.2 Beispiel: Ableitung von $\frac{2}{3}\sqrt{x^3}$ 62

D Herleitung der logistischen Regressionsgleichung **63**

Literatur **66**

Kapitel 1

Einleitung

Der Begriff Ereignisanalyse bezeichnet eine Reihe statistischer Verfahren, die zur Untersuchung von Zeitintervallen zwischen aufeinander folgenden Ereignissen oder Zustandswechseln verwendet werden. Die von den Untersuchungseinheiten - z.B. Parteien, Personen, Staaten oder Regierungen - eingenommen Zustände sind in der Regel abzählbar, also nicht unendlich. Es handelt sich also um einen *diskreten Zustandsraum*. Diese Zustände oder Ereignisse können zu jedem beliebigen Zeitpunkt eintreten. Die Zeit, mit der wir es zu tun haben ist folglich *stetig*.
Die Ereignisanalyse ist eine Methode mit einem breiten Anwendungsfeld. Untersucht werden kann die Zeitdauer bis zu einem Regierungswechsel in Land x oder der Wechsel der Parteipräferenz bei Person y. Überlebenszeiten von Patienten in medizinischen Studien, beispielsweise nach Herzoperationen oder Chemotherapie, die Dauer von Lernprozessen in der Psychologie, die Zeitspanne bis zu einem transregionalen Umzug in der räumlichen Mobilitätsanlyse, die Dauer der "Herrschaft" eines Löwen über sein Rudel in der Biologie oder die Dauer von Arbeitslosigkeit in ökonomischen Untersuchungen sind nur ein kleiner Ausschnitt möglicher Anwendungsfelder.
Die Statistik bietet heute eine grosse Anzahl an Möglichkeiten zur Analyse von Ereignisdaten. Sie umfassen:

- Deskriptive Verfahren: Sterbetafel-Methode oder Kaplan-Meier- Schätzung
- Semiparametrisches Regressionsmodell von Cox
- Parametrische Verfahren mit und ohne Zeitabhängigkeiten: Exponentialmodell, Piecewise-Constant-Modell, Gompertz- (Makeham-) Modell, Weibull-Modell, log-logistisches-Modell

Da die Ereignisanalyse in den letzten zwei Jahrzehnten sehr eng mit der Lebensverlaufsforschung verbunden gewesen ist, stehen dort die Veränderung und die Interaktion der verschiedenen Dimensionen des Lebenslaufs im Vordergund. Es hat sich gezeigt, dass die Methoden der Ereignisanalyse besonders geeignet sind, folgende drei konzeptionelle Dimensionen zu beschreiben:

1. **Selbstreferentielle Prozesse**: Der Verlauf der Entwicklung eines Individuums in einem bestimmten Bereich bezieht sich immer auf in diesem Lebensbereich bereits kummulierte Erfahrungen. Die Vorgeschichte der Person ist also immer in die gerade aktuellen Entscheidungen involviert. Vorerfahrungen und bereits in der Vergangenheit getroffene Entscheidungen begrenzen dabei den Spielraum der in der Zukunft möglichen Ereignisse.

2. **Multidimensionale Prozesse**: Der Lebensverlauf entwickelt sich in mehreren, wechselseitig aufeinander bezogenen Bereichen. Jeder Bereich ist ein Teilprozess des Lebensverlaufs, so beispielsweise die Bildungskarriere, die Krankengeschichte, der Familienverlauf, der Erwerbsbiographie oder das bisherige Wahlverhalten. Diese verschiedenen Lebensbereiche sind dabei in der Regel *nicht unabhängig* voneinander. Ein Beispiel hierfür ist das Zusammenspiel von Erwerbsprozess und Bildungskarriere oder Krankengeschichte. Der Lebensverlauf setzt sich hier also nicht aus dem selbstreferentiellen Bezug auf frühere Zustände zusammen, sondern durch die parallele Interdependenz vieler verschiedener Lebensbereiche in der Vergangenheit. Auch die unterschiedliche Gewichtung der einzelnen Bereiche im Hinblick auf das Alter einer Person ist dabei nicht zu vernachlässigen. So ist ersichtlich, dass die Krankengeschichte für einen jugendlichen oder "Twenty-something" im Normalfall weniger bedeutend ist, als für eine Person jenseits der 70.

3. **Gesellschaftliche Mehrebenenprozesse**: Der Lebensverlauf ist in solche hochgradig differenzierten Prozesse eingebettet. So haben beispielsweise Einfluss:

    - *Andere Personen* mit denen mehr oder weniger enge Interaktionsbeziehungen bestehen, beispielsweise Eltern, Lebenspartner, Kinder, Freunde etc.
    - *Verschiedene soziale Gruppen* deren Mitglied man ist, also elterliche Familie, eigene Familie, Sport- oder sonstige Vereine, Bezugsgruppen, "Peer-Groups"
    - *Veränderungen gesellschaftlicher Institutionen und sozialer Organisationen* wie staatliche Institutionen, Arbeitsorganisationen etc.

- *Wandel der Lebensbedingungen*, beispielsweise soziale oder regionale Kontexte.
- *Generelle Rahmenbedingungen*, so die historisch gewachsenen, sich verändernden gesellschaftlichen Strukturen, die die sozialkulturellen, politischen, rechtlichen, kulturellen und ökonomischen Rahmenbedingungen für die Lebensorganisation darstellen.

Zusammengefasst lässt sich sagen, dass es sich bei Verläufen (z.B. Lebensverläufe) um komplexe, nichtlineare Prozesse handelt, die durch *Selbstreferenz, zeitlich lokale Interdependenz sowie vertikale Interdependenz zwischen verschiedenen sozialen Prozessen* beeinflusst werden.

Kapitel 2

Grundlagen

2.1 Regression für Längschnittdaten?

Eine Annäherung an die Analyse von Ereignisdaten über das Standardverfahren der *multiplen Regression* ist leider nicht unproblematisch. Nach Allison (1984) führen die bei Ereignisdaten vorhandenen Zensierungen und zeitveränderlichen unabhängigen Variablen zu ernsten Problemen, wenn man statistische Standardverfahren anwenden möchte. Solche Verfahren können zu einem starken bias, oder zu enormen Datenverlust führen. Als Beispiel für diese Probleme nennt Allison eine Studie über Ex-Häftlinge: In dieser Studie wurde untersucht, ob Personen die aus dem Gefängnis entlassen wurden, in einem Ein-Jahresintervall wieder im Gefängnis landen. Obwohl das exakte Datum der Verhaftungen der in diesem Jahr rückfällig gewordenen bekannt war, wurde für den gesamten Zeitraum ein Dummy als abhängige Variable gebildet, der angab, ob das entsprechende Individuum verhaftet wurde oder nicht. Einmal davon abgesehen, dass die Verwendung einer multiplen Regression bei dieser Art von abhängiger Variable fragwürdig erscheint (Stichwort logistische Regression), nimmt man durch die (willkürlich) Bildung eines Dummys viel Informationsverlust in Kauf. Beispielsweise lassen Individuen, die direkt in der ersten Woche oder am ersten Tag nach der Entlassung wieder rückfällig werden theoretisch anders beschreiben, als jemand der nach 11 oder 12 Monaten rückfällig wird. Die Länge des Zeitintervalls von Freilassung bis zur nächsten Verhaftung zu nutzen ist aber auch nicht unproblematisch, da für alle Personen, die 12 Monate nach Entlassung nicht wieder im Gefängnis gelandet sind die Informationen zensiert sind. Es zeigt sich, dass eine grosse Anzahl von Zensierungen zu einem grossen bias führt. Selbst wenn kein einziger Fall zensiert wäre, würde sich das Problem, zeitveränderliche unabhängige Variablen zu integrieren, als schwerwiegend erweisen.

2.2 Unterscheidungen: A vs. B

2.2.1 Distributional- vs. Regressionsansätze

In den Anfängen widmete sich die Ereignisanalyse vornehmlich der Verteilung der Zeit vor einem Ereignis oder der Zeit zwischen zwei Ereignissen. Wie wir noch sehen werden, ist dies die Hauptidee hinter der Life-Table Methode (3.4.1). Mit der Weiterentwicklung der Ereignisanalytischen Verfahren verschob sich der Focus immer mehr auf die Regressionsmodelle, in denen das Auftreten eines Ereignisses von der Linearkombination einer oder mehrerer erklärender Variablen abhängt.

2.2.2 Repeated vs. nonrepeated events

In einigen Wissenschaften ist das interessierende Ereignis nicht wiederholbar. So interessiert in der Biostatistic oftmals der Tod des Individuums, dass natürlich nur einmal sterben kann. In den Ingenieurwissenschaften ist die Lebenszeit eines Bauteils von Interesse, das ebenso nur einmal "kaputt gehen" kann, und danach ausgetauscht wird. Anders liegt der Fall in den Sozialwissenschaften. Auch hier gibt es Ereignisse, die nur einmalig auftreten können, so wie die Geburt des ersten Kindes oder die erste Heirat. Aber wie man sich an dieser Stelle schon denken kann, ist es durchaus möglich in seinem Leben mehr als ein Kind zu bekommen oder öfter als einmal zu heiraten. Diese Ereignisse sind also wiederholbar. Diese Modelle sind also für uns interessanter, allerdings auch komplizierter.

2.2.3 Einzelne Ereignisse vs. multiple Fälle von Ereignissen

In manchen Analysen ist es nicht problematisch, alle Ereignisse gleich zu behandeln. So ist es beispielsweise in einer medizinischen Studie möglich, nur zwischen "Patient hat überlebt" und "Patient hat nicht überlebt" zu unterscheiden. Sollte die Fragestellung allerdings spezieller sein, dann ist es sinnvoll, auch hier zu unterscheiden. Ist ein Patient beispielsweise nach einer neuen Chemotherapie an den Folgen der Behandlung, an Krebs oder an einer damit nicht in Verbindung stehenden Krankheit wie einem Schlaganfall oder Herzinfarkt verstorben, oder ist die Todesursache vollkommen anders wie ein Verkehrsunfall oder ein Verbrechen? In diesen Fällen spricht man von "konkurrierenden Risiken" oder Competing Risks. Sofern diese voneinander unabhängig sind, ist ihre statistische Behandlung einfach: Bei der Untersuchung der Übergänge in einen bestimmten Zielzustand werden alle

anderen Übergänge als Zensierungen behalten, also als Beendigung der Beobachtungsdauer, ohne dass das untersuchte Zielereignis eingetreten wäre. Sind die verschiedenen Zielzustände jedoch nicht unabhängig (z.B.: Arbeitslose entscheiden sich umso mehr für eine Weiterbildung, je länger sie keinen Job gefunden haben), ist dieses Verfahren nicht zulässig. Eine adäqute statistische Behandlung solcher abhängiger Risiken ist nach Mayerhofer noch nicht möglich. In der Biostatistik wurden Modelle für konkurrierende Risiken (*competing risks*) entwickelt. Auch diese Modelle sind komplizierter als das Basismodell.

2.2.4 Parametrische vs. nichtparametrische Methoden

In der Biostatistik sind nichtparametrische Verfahren sehr beliebt, die kaum oder keine Annahmen über die Verteilung der Eintrittszeitpunkte der Ereignisse machen. In der Sozialwissenschaft und den Ingeniuerwissenschaften sind dagenen parametrische Verfahren, die genaue Angaben über diese Verteilung macht, beliebt. Um diese Verteilung zu beschreiben, bedient man sich besonderer Verteilungen aus der Mathematik, so beispielsweise der Gompertzverteilung, der Weibullverteilung oder der Exponentialverteilung. Eine Brücke zwischen diesen beiden Ansätzen wird vom proportional hazards Modell nach Cox (4.2)geschlagen. Dieser Ansatz ist insofern parametrisch, als er ein Regressionsmodell mit funktionalem Term angibt, und in sofern nicht-parametrisch, als es keine genauere Annahme über die Verteilung des Eintretens der Ereignisse trifft.

2.2.5 Diskrete vs. stetige/kontinuierliche Zeit

Modelle die annehmen, dass die Zeit des Eintretens des Ereignisses exakt gemessen ist, sind als continous-time modells oder Modelle mit stetiger Zeit bekannt. In der Praxis sind diese Zeitpunkte immer diskret gemssen, egal wie klein die Intervalle sind. Allerdings ist es möglich, bei feinen Intervallen eine kontinuierliche Messung zu unterstellen. Sind die Intervalle in Monaten oder Jahren gemessen, ist es angebrachter von einer diskreten Messung auszugehen.

2.3 Begriffe

Um zu verstehen, welche Ideen hinter der Ereignisanalyse stehen, ist es unumgänglich, einige zentrale Grundbegriffe zu definieren:

2.3.1 Zustände - State

Unter *Zuständen* verstehen wir die Ausprägungen der abhängigen Variablen. Dafür müssen wir festlegen, welche Zustände wir unterscheiden wollen. An jedem Zeitpunkt nimmt jede Person exakt einen Zustand ein. Z.B. unterscheidet man in der Untersuchung von Heiratsverläufen

- Nie verheiratet
- Verheiratet
- Geschieden
- Verwitwet

Das Set der möglichen Zustände wird auch Zustandsraum oder state space genannt.

2.3.2 Ereignis - Event

Unter *Ereignissen* versteht man Veränderungen von einem Zustand in einen anderen, also von einem Ursprungszustand (origin state) in einen Zielzustand (destination state). Erwähnenswert ist, dass die Zahl der Ereignisse von der Zahl der Zustände abhängt. Wenn nur zwischen verheiratet und verwitwet unterschieden wird, gibt es das Ereigniss "Scheidung" sozusagen nicht.

2.3.3 Verweildauer - Duration

Die *Verweildauer* gibt an, wie lange ein Individuum in einem Zustand verharrt, also z.B. wie lang eine Person Single ist und nicht heiratet, oder wie lange eine Ehe dauert bis sich die Ehe geschieden wird, oder ein Partner stirbt.

2.3.4 Risiko-Periode - Risk Period

Natürlich können nicht alle Personen sämtliche Zustände zu jedem Zeitpunkt einnehmen. Um eine bestimmte Veränderung zu durchleben muss die Person in dem Ursprungszustand sein, der den Wechsel in den Zielzustand erlaubt. Z.B. kann ein Single kein Witwer werden. Die Periode, in der ein Individuum dem Risiko ausgesetzt ist, einen bestimmten Zustandswechsel durchzumachen nennt man die *Risiko-Periode*. Ein eng verwandtes Konzept ist das *Risiko-Set*. Es wird von der Zahl aller Individuen gebildet, die zu einem bestimmten Zeitpunkt dem Risiko ausgesetzt sind einen Zustandswechsel zu erleben.

2.3.5 Risikomenge - Risk Set

Die Anzahl der Fälle, die an einem bestimmen Zeitpunkt (stetig) oder in einem bestimmten Intervall (diskret) dem Risiko eines Ereignisses unterliegt. Dies ist die Menge der "noch lebenden" Individuen, also derjenigen, denen noch kein Ereignis wiederfahren ist. Beachtenswert ist, dass die Risikomenge kontinuierlich abnimmt. Es ist also zwar auf den ertsen Blick verwunderlich, dass die Hazardrate wächst, während die Menge derjenigen Individuen, bei denen ein Ereignis auftritt, kleiner wird. Auf den zweiten Blick ist es jedoch einsichtig, dass dem so ist, da die Hazardrate steigt, aber die Menge derjenigen, für die ein Ereignis möglich ist, immer kleiner wird. In absoluten Zahlen wird diese Menge kleiner, in relativen Zahlen wächst aber der Anteil derjenigen aus der Risikomenge, die ein Ereignis erlebt haben.

2.4 Mathematische Grundlagen

Wir nehmen an, dass es sich bei T um eine stetige Zufallsvariable handelt. Bei $f(t)$ handelt es sich um die *probability density function*. Sie gibt an, wie sich die Wahrscheinlichkeiten auf die möglichen Zufallsergebnisse verteilen. Also beispielsweise wie wahrscheinlich es ist, dass eine Person einen IQ von 120 besitzt.
$F(t)$ bezeichnet die *distribution function* von T. Sie gibt also an, wieviele Fälle kumuliert in Relation zu allen Fällen bisher aufgetreten sind.

2.4.1 Dichtefunktion $f(t)$ & Verteilungsfunktion $F(t)$

Wenn es sich bei T um eine stetige Zufallsvariable handelt, kann die Verteilung auch als Dichtefunktion ($f(t)$) beschrieben werden, die mit der Verteilungsfunktion $F(t)$ in folgendem Zusammenhang steht:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} = \frac{\partial F(t)}{\partial t}$$

$$F(t) = P(T \leq t) = \int_0^t f(u)du$$

Die Dichte, also der Flächeninhalt wird über Integralrechnung angegeben. Bilden wir die erste Ableitung des Integrals erhalten wir die Ausgangsgleichung. Also:

$$f(t) = F'(t)$$

Verdeutlichen wir uns dies an Hand der uns bekannten Standardnormalverteilung: In folgender Graphik sehen wir die Dichtefunktion $f(x)$, die berühmte *Gauss'sche Glockenkurve* und die Verteilungsfunktion $F(x)$ der Standardnormalverteilung. Sie hat *keinen* glockenförmigen Verlauf.

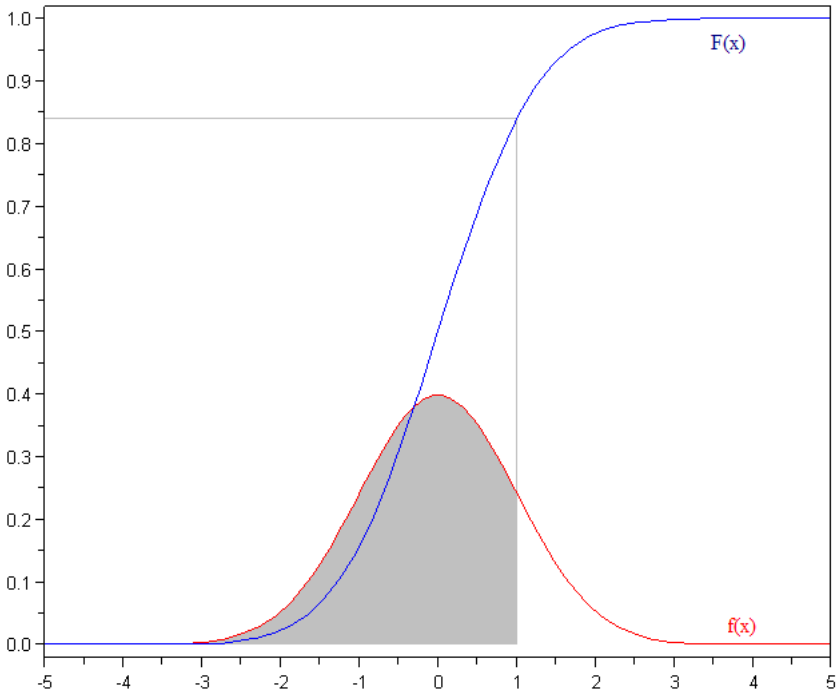

Abbildung 2.1: f(x) & F(x)

wobei:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} da$$

An der Dichtefunktion können wir sehen, wie die Wahrscheinlichkeit für die verschiedenen Ausprägungen verteilt sind. An unserer Verteilungsfunktion $F(x)$ können wir den grau eingefärbte Bereich ablesen, der uns hier angibt, wie viele Fälle von $-\infty$ bis in unserem Beispiel bis zum Z-Wert 1 liegen. In der Z-Tabelle sind diese Werte tabelliert.

2.4.2 Survivalfunktion $S(t)$

Die *survival function* oder *survival probability* gibt die Wahrscheinlichkeit dafür an, das vor dem Zeitpunkt t *kein* Ereigniss eintritt. Individuen, denen das Ereignis noch nicht wiederfahren ist haben "überlebt" (survived). Der Begriff überlebt kommt aus der Biostatistik, wo das interessierende Ereignis oftmals der Tod des Individuums ist. Bei $S(t)$ handelt es sich um eine fallende Funktion von t, mit $S(0) = 1$ und $S(t) = 0$ für $t \to \infty$. Dies bedeutet ausgesprochen, dass wir die Analyse mit 100% "Überlebenden" beginnen und sich nach unendlich langer Zeit ($t \to \infty$) bei jedem Individuum ein Zustandswechsel vom Urzustand in den Zielzustand vollzogen hat. Sie ist definiert als:

$$S(t) = 1 - F(t) = 1 - P(T \leq t) = P(T \geq t) = \int_t^\infty f(u)du$$

Die *distribution function* ist also das Komplement der *survival function*. Sie gibt die Wahrscheinlichkeit dafür an, dass ein Ereignis vor dem Zeitpunkt t statt findet. Folgender Zusammenhang besteht zwischen *survival function* und *distribution function*, der in den Graphiken 2.2 und 2.3 verdeutlicht werden soll:

$$F(t) + S(t) = 1$$

$$P(T \leq t) + P(T \geq t) = P(\Omega) = 1$$

$$\int_0^t f(u)du + \int_t^\infty f(u)du = \int_0^\infty f(u)du = 1$$

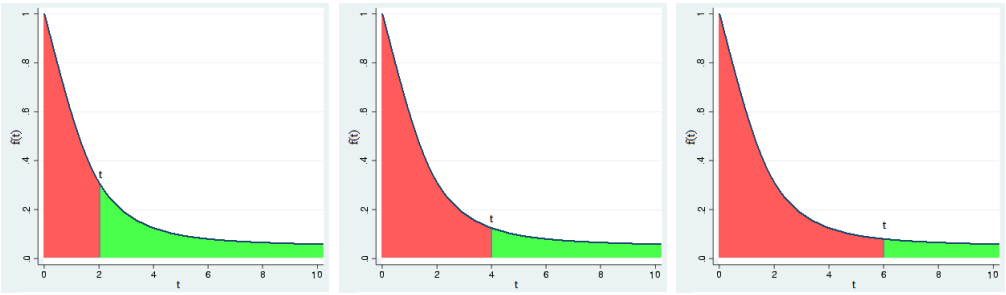

Abbildung 2.2: Eintrittswahrscheinlichkeit Ereignis

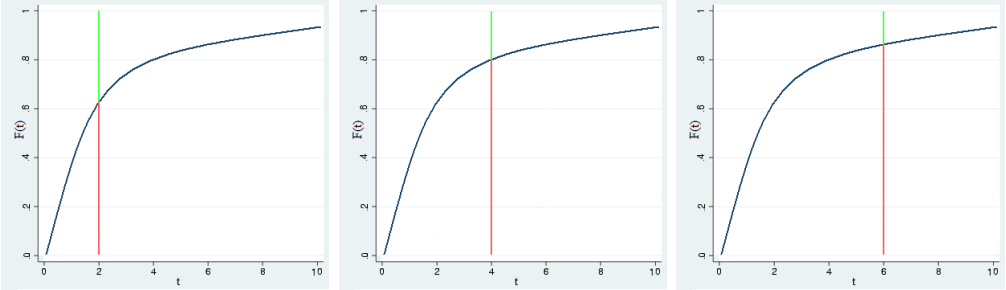

Abbildung 2.3: Flächeninhalt unter Kurve = Integral: survival function & distribution function

2.4.3 Hazardrate $h(t)$

Ein weiteres wichtiges Konzept ist die *hazard rate* oder *hazard function*. $P(t \leq T < t + \Delta t | T \geq t)$ bezeichnet die Wahrscheinlichkeit dafür, dass ein Ereignis im Intervall mit dem Zeitpunkt t als unterer Grenze und dem Zeitpunkt $t + \Delta t$ als oberer Grenze statt findet, sofern dieses Ereignis nicht schon vor dem Zeitpunkt t statt gefunden hat. Es soll also eine Veränderung von einem Anfangszustand in einen Zielzustand stattfinden. Beispielsweise von unverheiratet in verheiratet oder von verheiratet in geschieden. Sie gibt das augenblickliche "Risiko" für solch einen Zustandswechsel an.

$$P(t \leq T < t + \Delta t | T \geq t); \text{ wobei gilt } t < t + \Delta t$$

Dies ist die Wahrscheinlichkeit dafür, dass ein Ereignis eintritt, unter der Bedingung, dass vorher kein Ereigniss (keine Zustandsänderung) eigetreten ist, also im Intervall zwischen 0 und t. Ein Beispiel hierfür ist, dass sich jemand nur im interessierenden Intervall scheiden lassen kann, wenn er noch verheiratet ist, und sich nicht in einem beliebigen anderen vorherigen Intervall hat scheiden lassen. $\lim_{\Delta t \to 0}$ bedeutet, dass die Breite des Intervalls gegen Null geht, da Δt gegen 0 strebt, also obere und untere Grenze unendlich Nahe beieinander liegen. Dies ist möglich, da es sich bei T um eine stetige Zufallsvariable handelt. Das zeitliche Intervall wird also *sehr* kurz. Es zeigt sich jedoch das Problem, dass die Wahrscheinlichkeit in einem infinitesimal kleinen Intervall zu liegen Null ist.

$$\lim_{\Delta t \to 0} P(t \leq T < \Delta t + t | T \geq t) = 0$$

Um dies zu umgehen betrachten wir die Ratio aus Übergangswahrscheinlichkeit und der Grösse des Intervalls. So kommen wir zu der Wahrscheinlichkeit von Veränderungen in der abhängigen Variable pro Zeiteinheit:

$$\frac{P(t \leq T < \Delta t + t | T \geq t)}{\Delta t}$$

Dies erlaubt uns, folgenden Grenzwert zu definieren:

$$\lim_{\Delta t \to 0} \frac{P(t \leq T < \Delta t + t | T \geq t)}{\Delta t}$$

Hier haben wir nun das zentrale Konzept der *hazard rate* oder auch *transition rate* $h(t)$ vor uns:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

Der "hazard" gibt die Wahrscheinlichkeit an, dass das Ereigniss in einem sehr kurzen zeitlichen Intervall - sofern das Ereignis nicht schon vorher eingetreten ist - statt findet. Aus diesem Grund ist die *hazard rate* auch als "instantaneous risk" bekannt. Der Term

$$H(t) = \int_0^t h(u) du$$

Cumulative Hazard Function $H(t)$

wird *cumulative hazard function* genannt. Es gilt:

$$H(t) = -\ln S(t)$$

2.4.4 Verknüpfungen

Es ist möglich, die aufgeführten Begriffe $h(t)$, $S(t)$, $f(t)$ sowie $F(t)$ durch die jeweils anderen Begriffe auszudrücken. Es gilt

$$P(t \leq T < t + \Delta t | T \geq t) = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)}$$

Also können wir $h(t)$ wiefolgt umschreiben:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \frac{1}{P(T \geq t)}$$

$$= \frac{f(t)}{S(t)}$$

Ebenso lässt sich also Ausdrücken:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lim\limits_{\Delta t \to 0} \frac{P(t \leq T < \Delta t + t)}{\Delta t}}{P(T \geq t)}$$

Es gilt ebenfalls:

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\}$$

$$f(t) = \lim_{\Delta t \to 0} \frac{F(t') - F(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{p(t \leq T < t + \Delta t)}{\Delta t}$$

$$f(t) = h(t) \cdot S(t) = h(t) \cdot \exp\left\{-\int_0^t h(u)du\right\}$$

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} = F'(t)$$

$$F(t) = \int_0^t f(u)du$$

2.5 Zensierung

Beobachtungen von Ereignisgeschichten sind normalerwiese zensiert. Zensierung bedeutet, dass die Information über die Verweildauer in einem Zustand nicht vollständig ist. Man spricht von vollständiger Linkszensierung, wenn der Beginn und das Ende einer Episode *vor* dem Beobachtungsfenster liegen. Teilweise linkszensiert ist eine Episode wenn nur der Beginn vor dem Beobachtungsfenster liegt, aber wir nicht wissen, wann diese Episode begonnen hat.

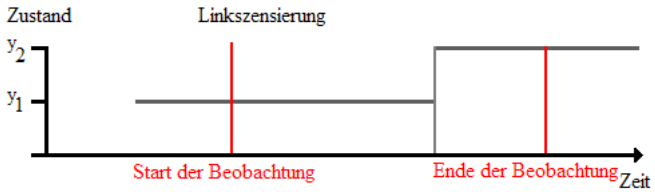

Abbildung 2.4: Teilweise linkszensiert

Linkszensierung ist ein schwerwiegendes methodischen Problem in der Ereignisanalyse, da die Information der unbekannten Episode oder Verweildauer nicht in das Modell mit einbezogen werden kann. Es entsteht ein Selektionsproblem, weil die Wahrscheinlichkeit, dass diese Episode beobachtet wird, vom Beginn und der Dauer dieser Episode abhängt. Es sind dann solche Episoden systematisch unterrepräsentiert, die entweder sehr kurz sind, oder die lange vor der Beobachtung begonnen haben. Nach *Blossfeld und Rohwer* sind nur solche Daten zu analysieren, bei denen die Annahme der Markov-Eigenschaften -d.h. wenn der Prozess nur vom Ausgangszustand, nicht aber von der Verweildauer im Ausgangszustand abhängt- gerechtfertigt ist.
Der Normalfall in der Ereignisanalyse ist jedoch die Rechtszensierung. In diesem Fall kennen wir den Anfang der Episode und deren Vorgeschichte, das Ende jedoch ist nicht bekannt. Dies ist immer dann der Fall, wenn zum Zeitpunkt der letzten Befragung die Episode noch nicht abgeschlossen war.

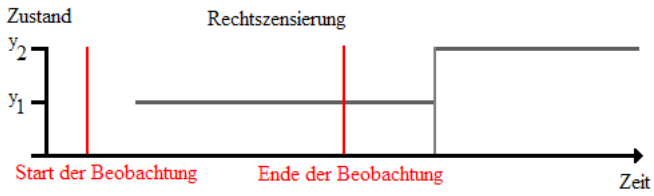

Abbildung 2.5: Rechtszensierung

Dies ist zum Beispiel dann der Fall, wenn jemand zum Ende des Beobachtungsfensters noch immer verheiratet ist. In diesem Fall ist die Ehedauer rechtszensiert. Da dieses rechtszensierende Ereignis im Normallfall unabhängig vom beobachteten Prozess eintritt, ist die statistische Handhabung dieser Rechtszensierungen methodisch unproblematisch.

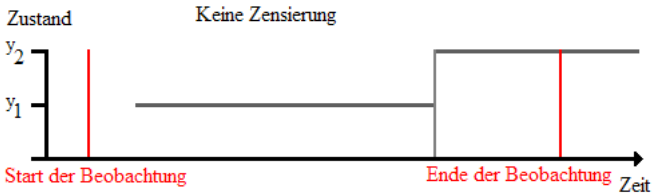

Abbildung 2.6: Keine Zensierung

Kapitel 3

Diskrete Zeit

3.1 Discrete Time Logit Models

In discrete time logit models wird angenommen, dass die Ereignisse nur zu bestimmten diskreten Zeitpunkten auftreten. Die Anwendung von discret time models kann mehreren Zwecken dienen:

1. Durch Modelle mit diskreter Zeit können Modelle mit stetiger Zeit approximiert werden.
2. Modelle mit diskreter Zeit nach Cox haben gegenüber Modellen mit stetiger Zeit Vorteile bei der Behandlung von Ties.
3. Der zu Grunde liegende Zeitprozess ist tatsächlich diskret.
4. Ein binärer Prozess -der bestimmten Anforderungen genügt- wird angenommen und durch Daten einer Panelbefragung analysiert.

Zuerst fokussieren wir uns auf Ereignisse, die nicht wiederholbar sind, also nur einmalig auftreten. Solche Ereignisse sind beispielsweis die Geburt des ersten Kindes oder die erste Heirat. Die nachfolgende Beschreibung basiert auf Cox und Brown

3.1.1 Approximation stetiger Zeit durch Modelle mit diskreter Zeit

Drei Überlegungen sind relevant für die Anwendung von Modellen mit diskreter Zeit um damit Modelle für stetiger Zeit zu approximieren.
Erstens die Einheit der Zeitmessung. Die Ereignisse, die wir erhalten sind höchst selten auf einer feinen Skala gemessen, sondern eher grob. So kennen

wir vielleicht das Alter eines Befragten in Jahren, aber nicht in Jahren, Monaten und Tagen. In solchen Fällen ist es natürlicher, diskrete Zeit zu Grunde zu legen.
Zweitens spielt die Anzahl der Ties in der Analyse eine Rolle. Man spricht von Ties, wenn die Ereignisse zweier oder mehrerer Personen gleichzeitig statt finden. Sind viele Ties vorhanden, kann dies zu einem ernsthaften bias in den Parameterschätzungen führen, wenn "Cox method for proportional hazards" (4.2) für stetige Zeit genutzt wird.
Drittens ist die Frage, ob es adäquat ist, durch solche Modelle zu approximieren von Bedeutung. Dies hängt mit der bedingten Wahrscheinlichkeit, ein Ereignis an einem diskreten Zeitpunkt zu beobachten, zusammen. Diskrete Modelle sind nur geeignet Modelle mit stetiger Zeit zu approximieren, wenn die bedingt Wahrscheinlichkeit angemessen klein ist.

3.2 Mathematische Konzepte

Nehmen wir an, bei T handelt es sich um eine diskrete Zufallsvariable, die den Zeitpunkt eines Ereignisses angibt. $T = t$ bedeutet, dass das Ereignis zum Zeitpunkt t eintritt. Die Wahrscheinlichkeit eines Ereignisses ist gegeben durch:

$$f(t_i) = P(T = t_i),\ i = 1, 2, \ldots$$

wobei t_i mit $t_1 < t_2 < \ldots$ den i^{ten} diskreten Zeitpunkt bezeichnet.
Die Survivorfunktion ist gegeben durch

$$S(t_i) = P(T \geq t_i) = \sum_j f(t_j)$$

Sie gibt die Wahrscheinlichkeit an, vor dem Zeitpunkt t_i kein Ereignis zu erleben. Das Risiko zum Zeitpunkt t_i ist als bedingte Wahrscheinlichkeit des Eintretens des Ereignisses zum Zeitpunkt t_i, unter der Bedingung, dass es nicht schon vorher eingetreten ist, definiert.

$$h(t_i) = \lambda_i = P(T = t_i | T \geq t_i) = \frac{f(t_i)}{S(t_i)}$$

anders geschrieben:

$$S(t_i) = \prod_{j=1}^{i-1} (1 - \lambda_j)$$

Jede parametrische Spezifikation der bedingten Wahrscheinlichkeit für λ_j ist ein Hazardmodell mit diskreter Zeit.

3.3 Logitmodell für diskrete Zeit

Im ersten Schritt ist zu spezifizieren, wie die Hazardrate von den unabhängigen Variablen abhängt. $P(t)$ bezeichnet hier die Wahrscheinlichkeit, dass ein Individuum ein Ereignis zum Zeitpunkt T hat, sofern es das Ereignis nicht schon vor T erlebt hat. Der Einfachheit halber nehmen wir an, dass wir es mit zwei unabhängigen Variablen zu tun haben. x_1 ist über die Zeit konstant, z.B. in einem wissenschaftlichen Kontext das Prestige des beschäftigenden Instituts, und $x_2(t)$ kann mit den Zeitpunkten seinen Wert wechseln, beispielsweise die Anzahl der Publikationen zum Zeitpunkt T.
Als erste Annäherung (für t Zeitpunkte) können wir $P(t)$ als Linearkombination der unabhängigen Variablen schreiben:

$$P(t) = a + b_1 x_1 + b_2 x_2(t)$$

Problematisch ist hier, dass $a+b_1x_1+b_2x_2(t)$ jeden beliebigen Wert annehmen kann, $P(t)$ jedoch auf den Wertebereich $0 \leq P(t) \leq 1$ eingeschränt ist. Was also tun? Das Logitmodell für diskrete Zeit nutzt das Konzept des *Logit* oder der *log-Odds*. Unter Odds versteht man die Ratio zweier wechselseitig exklusiven Zustände. Die Odds für die Wahrscheinlichkeit $P(t)$ sind wie folgt definiert:

$$\text{Odds} = \frac{P(t)}{1 - P(t)}$$

Man kann erkennen, dass je grösser $P(t)$ auf dem Bruchstrich wird, $1 - P(t)$ unter dem Bruchstrich immer kleiner wird. Da $P(t)$ immer noch auf den oben angegebenen Wertebereich beschränkt ist, also nicht negativ werden kann, sind die Odds auf das Intervall zwischen 0 und $+\infty$ fesgelegt. Um den kompletten Wertebereich von $-\infty$ bis $+\infty$ zu erschliessen, muss also noch ein Schritt getan werden. An dieser Stelle kommt der Begriff des Logit ins Spiel:

$$\text{Logit} = \ln\left[Odds\right] = \ln\left\{\frac{P(t)}{1 - P(t)}\right\}$$

Logits sind als logarithmierte Odds definiert, also Log-Odds. Wir verwenden aber nicht irgendeinen Logarithmus, sonder den Logrithmus Naturalis. Dies ist die Bezeichnung des Logarithmus zur Basis e, also $log_e = \ln$.
Also:

$$\ln\left\{\frac{P(t)}{1 - P(t)}\right\} = a + b_1 x_1 + b_2 x_2(t)$$

Diese Transformation ist nicht die einzige, die zu diesem Ergebnis führt, aber sie ist die gängigste. Die Koeffizienten b_1 und b_2 geben die Veränderung des Logit für jede Änderung von x_1 oder x_2 um eine Einheit an.
Dieses Modell schränkt uns immer noch ein, da es annimmt, dass Änderungen in der Hazardrate unter Einfluss von x_1 und x_2 auftreten. Oftmals ist es aber sinnig anzunehmen, dass die Hazards sich autonom über die Zeit verändern. Bei Jobwechseln kann man erwägen, dass der Hazard eines Wechsels mit verstreichender Zeit abnimmt. Dies lässt sich folgendermaßen in die Gleichungen einbauen:

$$\ln\left\{\frac{P(t)}{1-P(t)}\right\} = a(t) + b_1 x_1 + b_2 x_2(t)$$

Das Intercept, oder anders gesprochen die Regressionskonstante wird an jedem Zeitpunkt (t) einfach neu geschätzt.
Die Notation unterscheidet sich leider von Lehrbuch zu Lehrbuch. Da das zu Grunde liegende Konzept soweit klar sein sollte wird ab hier dies ausführlichere, aber auch verwirrendere Notation von Yamaguchi übernommen. So definiert Yamaguchi (1991) das Logitmodell folgendermaßen:

$$\frac{\lambda(t_i;\mathbf{X})}{1-\lambda(t_i;\mathbf{X})} = \frac{\lambda_0(t_i)}{1-\lambda_0(t_i)} + \exp\left\{\sum_k b_k X_k\right\}$$

Für mich scheinen diese beiden Notationen fogendermaßen in Einklang zu bringen zu sein:
Zuerst exponieren wir die einzelnen Komponenten.

$$\ln\left\{\frac{\lambda(t_i;\mathbf{X})}{1-\lambda(t_i;\mathbf{X})}\right\} = \ln\left\{\frac{\lambda_0(t_i)}{1-\lambda_0(t_i)}\right\} + \ln\left\{\exp\left\{\sum_k b_k X_k\right\}\right\}$$

Beim Logarithmus Naturalis handelt es sich um den Logarithmus zur Basis e, also der Eulerschen Zahl ($\ln = \log_e$). Dies ist die Umkehrfunktion zu $\exp\{x\} = e^x$. Es folgt also:

$$\ln\left\{\frac{\lambda(t_i;\mathbf{X})}{1-\lambda(t_i;\mathbf{X})}\right\} = \ln\left\{\frac{\lambda_0(t_i)}{1-\lambda_0(t_i)}\right\} + \sum_k b_k X_k$$

An einer späteren Stelle des Buches schreibt Yamaguchi selber, dass sich

$$\frac{\lambda(t_i;\mathbf{X})}{1-\lambda(t_i;\mathbf{X})} = \frac{\lambda_0(t_i)}{1-\lambda_0(t_i)} + \exp\left\{\sum_k b_k X_k\right\}$$

auch in Form einer logistischen Regression darstellen lässt, und zwar:

$$\ln\left\{\frac{\lambda(t_i;\mathbf{X})}{1-\lambda(t_i;\mathbf{X})}\right\} = a_i + \sum_k b_k X_k$$

wobei $a_i = \ln\frac{\lambda_0(t_i)}{1-\lambda_0(t_i)}$, also eine einfache Substitution.
Eine Herleitung der logistischen Regressionsgleichtung finden wir im Appendix auf Seite 63
Wenn die Kovariate alle zeitunabhängig sind, also sich über die Zeit nicht ändern, liegt ein proportional Odds model vor. In diesem Fall bilden die Odds $(\frac{\lambda(t_i;\mathbf{X})}{1-\lambda(t_i;\mathbf{X})})$, dass ein Ereignis eintritt eine konstante Ratio
Substituieren wir $\ln\left\{\frac{\lambda_0(t_i)}{1-\lambda_0(t_i)}\right\}$ durch a erhalten wir:

$$\ln\left\{\frac{\lambda(t_i;\mathbf{X})}{1-\lambda(t_i;\mathbf{X})}\right\} = a + \sum_k b_k X_k$$

wobei $\lambda(t_i;\mathbf{X})$ die bedingte Wahrscheinlichkeit angibt, ein Ereignis zum Zeitpunkt t_i für einen bestimmten Kovariatvektor $\mathbf{X} = (X_1, \ldots, X_k)$ zu erhalten. bei b_k, $k = 1, \ldots, K$ handelt es sich um Parameter. Die *baseline hazard function* $\lambda_0(t_i)$ mit $i = q, \ldots, I$ ist durch die bedingte Wahrscheinlichkeit der Fälle charakterisiert, für die $\mathbf{X} = \mathbf{0}$ gilt. Ebenso kann man hier sehen, dass die Wahrscheinlichkeit, ein Ereignis zu erleben, für jeden Fall, der nicht zur baseline group gehört an jedem Zeitpunkt um $\exp\{\sum_k b_k X_k\}$ höher liegt, da dieser Term in der baseline group wegen $\mathbf{X} = \mathbf{0}$ wegfällt.
Bei immer feiner werdenden Messungen der Zeit wird die Ratio zweier Odds

$$\frac{\frac{\lambda(t_i;\mathbf{X})}{1-\lambda(t_i;\mathbf{X})}}{\frac{\lambda_0(t_i)}{1-\lambda_0(t_i)}}$$

der Ratio zweier Raten immer ähnlicher:

$$\frac{\lambda(t_i;\mathbf{X})}{\lambda_0(t_i)}$$

und nähert sich einem proportional hazards model für stetige Zeit an. Also: wenn die bedingten Wahrscheinlichkeiten genügend klein sind, dann liefert uns das Logit-Modell eine Approximation des proportional hazards model für stetige Zeit.

3.4 Deskriptiv: Nichtparametrische Verfahren

Nichtparametrische Verfahren sind Verfahren, bei denen keine Annahmen über die Verteilung der Wartezeit gemacht wird. Hierzu zählen die Life Ta-

ble Method ("Sterbetafelschätzung") als auch die Kaplan-Meier-Schätzung (Product-Limit Estimation). Die Life-Table Methode hat ihren Ursprung in der Demographie und zählt zu den bekanntesten und lange Zeit beliebtesten Methoden der Ereignisanalyse. Erwähnenswert ist, dass eine der bekanntesten Regressionsmethoden für Ereignisdaten (Die Regression nach Cox (4.2)) von der Grundidee hinter der Life-Table Methode inspiriert ist.
Der wesentlicher Unterschied zwischen diesen beiden nichtparametrischen explorativen Verfahren ist, dass die Sterbetafel-Schätzung für gruppierte Wartezeiten und die Produkt-Limit-Schätzung für exakte Wartezeiten konzipiert ist. Neben einer ersten allgemeinen Beschreibung des Veränderungsprozesses besteht auch die Möglichkeit, anhand eines Vergleichs der geschätzten Überlebensfunktionen und Hazardraten einzelner Subgruppen, einen Überblick über mögliche Erklärungsfaktoren zu gewinnen.

3.4.1 Life Table Methode: Verweildauer in Intervallen

Wie bereits erwähnt, sind bei der Life-Table Methode keine Annahmen über die Verteilung von T notwendig. Errechnet werden die Survivorfunktionen zu Beginn des jeweiligen Intervalls sowie für jedes Intervall die Dichte- und Hazardfunktion (und deren Standardfehler). Nachteile dieser Methode sind, dass diskrete Zeitintervalle nötig sind und dass sie eine grosse Anzahl an events benötigt, um reliable zu sein. Um die diskreten Intervalle zu erhalten, wird die Zeitachse punktweise aufgesplittet.

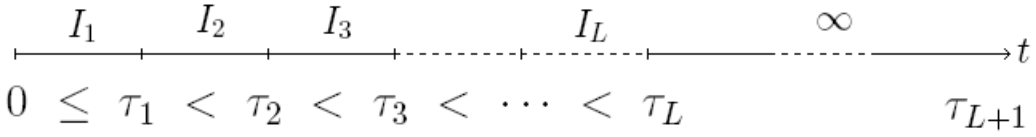

Abbildung 3.1: Einteilung in diskrete Intervalle

Mit der Konvention: $\tau_{L+1} = \infty$ existieren L Intervalle, von denen jedes die linke Grenze beinhaltet, aber nicht die Rechte. Es gilt:

$$I_l = \{t | \tau_l \leq T \leq \tau_{l+1}\},\ l = 1, \cdots, L$$

Terminologie:

- N_l Zahl der Fälle, die in Intervall I_l eintreten.
- E_l Zahl der Ereignisse / Übergänge im Intervall I_l
- Z_l Zahl der Zensierungen im intervall I_l

- $\mathcal{R}_l$ Risk Set / Risikomenge im Intervall I_l
- R_l Zahl der Elemente in $\mathcal{R}_l$

Wenn wir die Zahl der Fälle, die im jeweiligen Intervall ein Ereignis (Übergang in den Zielzustand) erfahren, mit E_l benennen und die Zahl der Fälle mit Zensierungen in einem Intervall mit Z_l, so lässt sich zunächst die Risikomenge $\mathcal{R}_l$, also die Zahl der Fälle, die im jeweiligen Intervall dem Risiko eines Ereignisses unterliegt, berechnen. Hier wird wiederum die Zahl der Fälle benötigt, die zu Beginn eines Intervalls noch nicht ausgeschieden ist (durch ein Ereignis oder durch Zensierung). Diese ist für das erste Intervall gleich N (der Gesamtzahl der Fälle), für alle folgenden Intervalle gilt:
Rekursive Bestimmung von N_l. Es gilt für das erste Intervall:

$$N_1 = N$$

Für das zweite Intervall:

$$N_2 = N_1 - E_1 - Z_1$$

Generell gilt:

$$N_l = N_{l-1} - E_{l-1} - Z_{l-1}$$

Berücksichtigung von Zensierung in I_l: Zur Berechnung der Risikomenge sind nun Annahmen über die Verteilung der zensierten Fälle während des Intervalls zu machen. Üblicherweise wird angenommen, dass die Zensierungen gleichmäßig über das gesamte Intervall verteilt sind; daraus folgt, dass die Zahl der Fälle zu Beginn des Intervalls um die Hälfte der Zensierungen während dieses Intervalls zu reduzieren ist, um die Risikomenge zu erhalten. Die Risikomenge R wird also folgendermaßen bestimmt:

$$R_l = N_l - \frac{1}{2} \cdot Z_l$$

Wird dies nicht angenommen gilt allgemein folgendes:

$$R_l = N_l - \omega \cdot Z_l \ , \ \omega = (0 \leq \omega \leq 1)$$

wobei für $\omega = \frac{1}{2}$ die vorherige annahme wieder gilt.
Die bedingte Wahrscheinlichkeit für einen Übergang im Intervall I_l ist wie folgt definiert:

$$q_l = \frac{E_l}{R_l}$$

Folglich lautet die bedingte Wahrscheinlichkeit für keinen Übergang im Intervall I_l, also das Intervall zu überleben:

$$p_1 = 1 - q_l = 1 - \frac{E_l}{R_l}$$

Die Überlebenswahrscheinlichkeit zu Beginn von I_l, also die Survivorfunktion lautet:

$$S_l = 1;\ S_l = p_{l-1} \cdot S_{l-1}$$

die durchschnittliche Überlebenswahrscheinlichkeit im Intervall I_l ist wie folgt definiert:

$$\bar{S}_l = \frac{S_l + S_{l+1}}{2}$$

Die durchschnittliche Wahrscheinlichkeitsdichte im Intervall I_l ergibt sich durch

$$f_l = \frac{S_l - S_{l+1}}{\tau_l - \tau_{l+1}},\ l = 1, \cdots, L-1$$

sowie die Hazard-Rate

$$h_l = \frac{f_l}{\bar{S}_l}$$

die auch in anderer Form dargestellt werden kann:

$$h_l = \frac{1}{\tau_{l+1} - \tau_l} \cdot \frac{q_l}{1 - \frac{q_l}{2}} = \frac{1}{\tau_{l+1} - \tau_l} \cdot \frac{E_l}{R_l - \frac{E_l}{2}}$$

Life-Tables sind für den Vergleich mehrerer Gruppen anwendbar. In nachfolgender Graphik 3.2 sehen wir ein Beispiel aus Arias (2003) über die Anwendung von Life Tables.

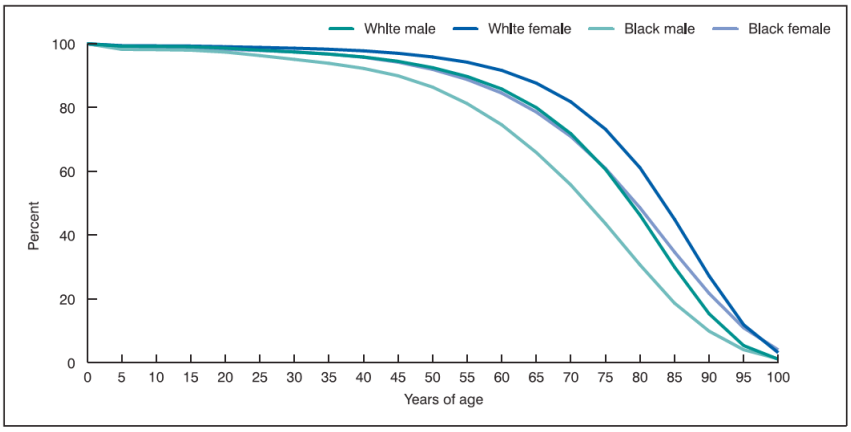

Figure 2. Percent surviving by age, race, and sex: United States, 2003

Abbildung 3.2: Vergleich mehrerer Gruppen

Hier wurden vier Subgruppen gebildet, die nun untereinander hinsichtlich ihrer Lebenserwartung verglichen werden können. Allerdings wird oftmals betont, dass die Interpretation von Life Table Tabellen nicht immer einfach ist, und sich daher Graphiken zur Erleichterung anbieten.
Im Unterschied zur Sterbetafelschätzung verwendet der Kaplan-Meier-Schätzer direkt die Wartezeiten; eine Klassifizierung in Intervalle und eine Annahme über die Verteilung der Ereignisse und Rechtszensierungen pro Intervall wird nicht vorgenommen.

3.4.2 Product-Limit Estimation / Kaplan-Meier

Der Unterschied zu der Life-Table Methode ist die direkte Verwendung der Wartezeiten. Es ist also unnötig, eine Zusammenfassung der Zeit in Intervallen vorzunehmen. Statt dessen wird die Risikomenge für jeden Zeitpunkt, an dem ein Ereignis statt findet, berechnet. Graphik 3.3 zeigt ein Beispiel, entmommen aus , http://www.thieme-connect.com/ejournals/pdf/dmw/doi/10.1055/s-2002-32819.pdf. Eine Sortierung der Zeitpunkte mit Ereignissen ist erforderlich:

$$\tau_1 < \tau_2 < \tau_3 < \cdots < \tau_L$$

wobei τ_1 den Zeitpunkt bezeichnet, an dem das erste Ereignis stattfindet, τ_2 den Zeitpunkt, an dem das zweite Ereignis staffindet, und so weiter.
Terminologie:

- E_l Zahl der Episoden mit Ereignissen zum Zeitpunkt τ_l. Es gilt: $\tau_0 = 0$
- Z_l Zahl der Zensierugen im Intervall $\tau_{l-1} \leq t < \tau_l$. Dies bedeutet, dass wenn Zensierung und Ereigniss zum selben Zeitpunkt stattfinden wird, angenommen, dass die Zensierung etwas später statt findet.
- R_l Risikomenge zum Zeitpunkt τ_l, d.h.: mit einer Startzeit $t_{\text{Start}} < t_l$ und einer Endzeit $t_{\text{Ende}} \geq t_l$

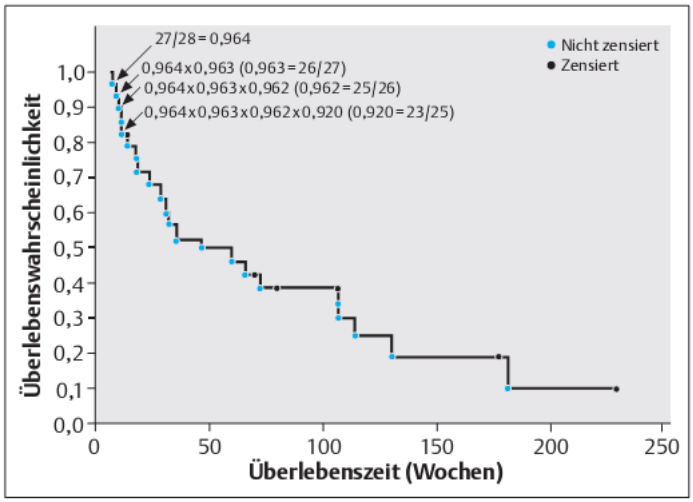

Abb. **1** Kaplan-Meier Kurve für die Überlebenszeit der 28 Zungenkrebspatienten mit diploidem Tumor. Es wird die Wahrscheinlichkeit gezeigt, dass ein Patient eine Zeit (in Wochen) überlebt.

Abbildung 3.3: Beispiel: Kaplan-Meier Kurve

Die Risikomenge R_l zum Zeitpunkt τ_l enthält Episoden, die zu diesem Zeitpunkt zensiert sind. Es wird angenommen, dass eine zensierte Episode die Information enthält, dass, inklusive des Endzeitpunktes kein Ereignis aufgetreten ist. Nach *Blossfeld und Rohwer* wird in der Literatur oftmals davon ausgegangen, dass die Zensierung einen infinitesimalen Betrag rechts der Endzeit der Beobachtung statt findet.

Es gilt für einen Zeitpunkt mit Ereignis:

$$q_l = \frac{E_l}{R_l} \quad p_l = 1 - q_l = 1 - \frac{E_l}{R_l}$$

sowie für einen Zeitpunkt ohne Ereignis:

$$q_l = 0 \quad p_l = 1 - q_l = 1$$

Der Product-Limit-Estimator für $S(t)$ ist definiert als:

$$\hat{S}(t) = p_1 \cdot p_2 \cdot p_3 \cdot \cdots \cdot p_{l-1} = \prod_{l:\tau_l<t} p_l = \prod_{l:\tau_l<t} 1 - \frac{E_l}{R_l}$$

Noch einmal: bei E_l handelt es sich um die Zahl der Episoden mit Ereignis zum Zeitpunkt τ_l. Anders gesprochen handelt es sich hierbei um die Anzahl der Personen, die in diesem Intervall "ausfallen". Bei R_l handelt es sich um

die Risikomenge zum Zeitpunkt τ_l, also die Personen die "noch leben", und nicht zensiert wurden. Zur Verdeutlichung hier ein kurzes Beispiel, zur Verdeutlichung fallen hier jeweils mehrere Personen an einem Zeitpunkt aus der Analyse, stören wir uns nicht daran:
Wir starten mit n=125 Personen. Zum ersten Zeitpunkt, an dem ein Ereignis stattfindet, fallen 5 Personen gleichzeitig aus. Also: $\frac{5}{125} = 0.04$. Diesen relativen Anteil der Ausfälle ziehen wir nun von der Gesamtheit ab, also von 1. Wenn wir diese Zahlen mit 100 multiplizieren, bekommen wir Prozentwerte heraus. Also sind von 100% (125 Personen) am ersten Zeitpunkt 4% (5 Personen) ausgefallen. Zum zweiten Zeitpunkt fallen 10 Personen aus. Also $\frac{10}{120} = 0.08\bar{3}$. Es sind also zum zweiten Zeitpunkt von der Anzahl der Personen vom ersten Zeitpunkt $8.\bar{3}$% nicht mehr "dabei". Im dritten Schritt fallen 15 Personen aus, also $\frac{15}{110} = 0.1\bar{3}6$. In jedem Schritt wird also der relative Anteil der Ausfälle, gemessen an der Anzahl der Verbliebenen zum Zeitpunkt τ_1 angegeben.

$$\hat{S}(t) = p_0 \cdot p_1 \cdot p_2 \cdot p_3$$

$$\hat{S}(t) = (1 - q_0) \cdot (1 - q_1) \cdot (1 - q_2) \cdot (1 - q_3)$$

$$\hat{S}(t) = \left(1 - \frac{E_0}{R_0}\right) \cdot \left(1 - \frac{E_1}{R_1}\right) \cdot \left(1 - \frac{E_2}{R_2}\right) \cdot \left(1 - \frac{E_3}{R_3}\right)$$

wobei:

$$\hat{S}(t) = \left(1 - \frac{0}{125}\right) \cdot \left(1 - \frac{5}{125}\right) \cdot \left(1 - \frac{10}{120}\right) \cdot \left(1 - \frac{15}{110}\right)$$

$$\hat{S}(t) = (1 - 0) \cdot (1 - 0.04) \cdot (1 - 0.08\bar{3}) \cdot (1 - 0,1\bar{3}6)$$

$$\hat{S}(t) = 1 \cdot 0.96 \cdot 0.91\bar{6} \cdot 0,86\bar{3}$$

Dies bedeutet nun inhaltlich: Zum Zeitpunkt 0, also am Anfang sind alle Personen "lebend". Da wir den Wert 1 erhalten, kann man diesen Zeitpunkt also bedenkenlos wegfallen lassen. Zum Zeitpunkt des ersten Ereignisses bleiben 0,96 oder 96% übrig. Zum zweiten Zeitpunkt bleiben $91.\bar{6}$% der Überlebenden des ersten Zeitpunktes erhalten. Die Berechnung für von $91.\bar{6}$% von 96% erfolgt über $0.96 \cdot 0.91\bar{6} \approx 0.87\bar{9}$. Für den dritten Zeitpunkt multipliziert man dieses Ergebnis mit dem Wert des dritten Zeitpunktes: $0.87\bar{9} \cdot 0.863 \approx 0.96 \cdot 0.91\bar{6} \cdot 0.863 \approx 0.75\bar{9}$. Kürzer geschrieben:

$$\hat{S}(t) = p_1 \cdot p_2 \cdot p_3 \cdot \cdots \cdot p_{l-1} = \prod_{l:\tau_l < t} p_l = \prod_{l:\tau_l < t} 1 - \frac{E_l}{R_l}$$

Hierbei handelt es sich um eine Treppenfunktion mit den Stufen zu den Zeitpunkten τ_l.
Zusätzlich zur Schätzung der Survivor-Funktion bietet die Product-Limit-Estimation eine simple Schätzung für die cumulated hazard rate:

$$\hat{H}(t) = -log(\hat{S}(t))$$

wobei:

$$H(t) = \int_0^t h(u)du$$

$$S(t) = \int_t^\infty f(u)du = \exp\left\{-\int_0^t h(u)du\right\}$$

und demnach:

$$H(t) = -log(S)t))$$

$$\int_0^t h(u)du = -log\exp\left\{-\int_0^t h(u)du\right\}$$

$$\int_0^t h(u)du = (-1)\left(-\int_0^t h(u)du\right)$$

$$\int_0^t h(u)du = \int_0^t h(u)du$$

Die cumulated hazard rate ist wiederum eine Treppenfunktion. Sie ist nützlich für einfache graphische Überprüfungen der Verteilungsannahmen und der zu Grunde liegenden Verweildauern. Leider bietet sie keine direkte Schätzung der hazard rate. Man könnte den Zusammenhang

$$\hat{h}(t) = \hat{H}'(t)$$

als möglichen Weg nutzen, da generell folgender Zusammenhang zwischen Funktionen gilt: Wenn $F(x)$ die Stammfunktion der Funktion $f(x)$ ist, die über Integration ermittelt wird, dann ist die erste Ableitung $(f'(x))$ der Stammfunktion $(F'(x))$ mit der Ursprungsfunktion identisch. Dafür muss die Treppenfunktion jedoch erst geglättet werden.

3.4.3 Nachteile nichtparametrischer Verfahren

Mit der Anwendung nichtparametrischen Verfahren treten diverse Probleme auf.
Erstens wird mit einer wachsenden Anzahl von Subgruppen schnell ein Punkt erreicht, an dem ein Vergleich der survivor functions $S(t)$ nicht mehr sinnvoll ist, da n in den verschiedenen Subgruppen zu klein wird.

Zweitens ist selbst wenn n in den verschiedenen Gruppen groß genug, und wir für eine grosse Anzahl wichtiger Subgruppen Survivorfunktionen schätzen können, so ist der Vergleich dieser Funktionen schnell sehr komplex und die Interpretation äußerst schwierig.
Drittens ist es im Fall quantitativer Variablen notwendig, diese zu gruppieren, um die Survivorfunktionen schätzen zu können. Beispielsweise wird eine metrische Einkommensvariable in eine neue Variable mit weniger Ausprägungen eingeteilt, z.B. Trichotom (niedriges - mittleres - hohes Einkommen) oder Dichotom (niedriges - hohes Einkommen). Der Informationsverlust ist dementsprechend groß.

Kapitel 4

Stetige Zeit

Im wesentlichen können Zeitverändliche Raten auf drei Wegen modelliert werden:

1. durch Aufnahme von Polynom-Termen für die Zeit,
2. durch Modellierung perioden- oder zeitabschnitts-spezifischer Regressionskonstanten (gegebenenfalls auch periodenspezifischer Einflüsse)
3. durch Wahl einer geeigneten Verteilung für die Hazardrate.

Nur der erste Weg ist sowohl für stetige als auch für diskrete Verweildauern möglich. Alle übrigen Verfahren sind nur für stetige Zeit ausformuliert. Wir werden uns auf die zwei letztgenannten konzentrieren.

4.1 Parametrische Modelle der Zeitabhängigkeit

Obwohl die Modelle mit diskreter Zeit einen breiten Anwendungsbereich haben wird doch zum grössten Teil mit Modellen für stetige Zeit gearbeitet. Dabei sind die parametrischen Modelle populär. Sie werden so genannt, weil in ihnen jeder Aspekt des Modells spezifiziert ist, ausser den zu schätzenden Parametern. Es ist wichtig, sich vor Augen zu führen, dass die Wahl der Verteilung die Hazardrate determiniert(ebenso die Zeit bis zu einem Ereignis oder zwischen zwei Ereignissen), da diese in einem engen Zusammenhang stehen.

4.1.1 Exponential Hazard Rate Models

Hierbei handelt es sich um das einfachste Hazard-Rate Modell, das auch unter dem Namen Exponential Transition Rate Model bekannt ist. Die angenommene Dauer von T kann von einer Exponentialverteilung angegeben werden. Im Exponential-Modell wird also die Verweildauer bis zu einem Ereignis durch eine Exponentialverteilung beschrieben. Das Risiko, dass ein Ereignis eintritt, ist von den im Modell beinhalteten Kovariaten abhängig, ist aber über alle Zeitpunkte unverändert konstant. Ein einziger Parameter -b- determiniert das Modell. Die Schätzung erfolgt über die Maximum-Likelihood Methode.
Es gilt:

Basic Exponential Model

$$f(t) = b \cdot \exp\{-bt\}, \; b > 0$$

$$h(t) = b$$

Achtung! Die Hazardrate ist Konstant über die Zeit.

$$S(t) = \exp\{-bt\}$$

wobei $b = \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}$, also:

$$h(t) = \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}$$

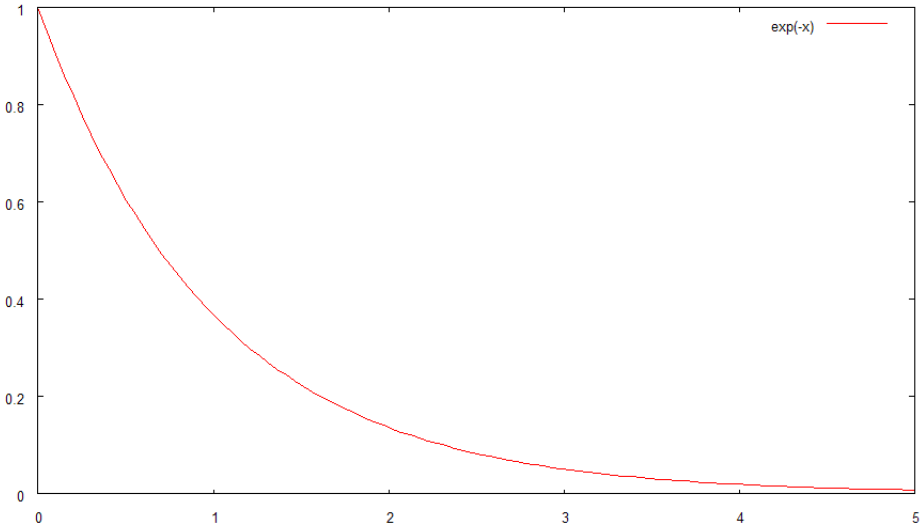

Abbildung 4.1: Dichte- & Survivor-Funktion im Exponential Hazard Rate Modell

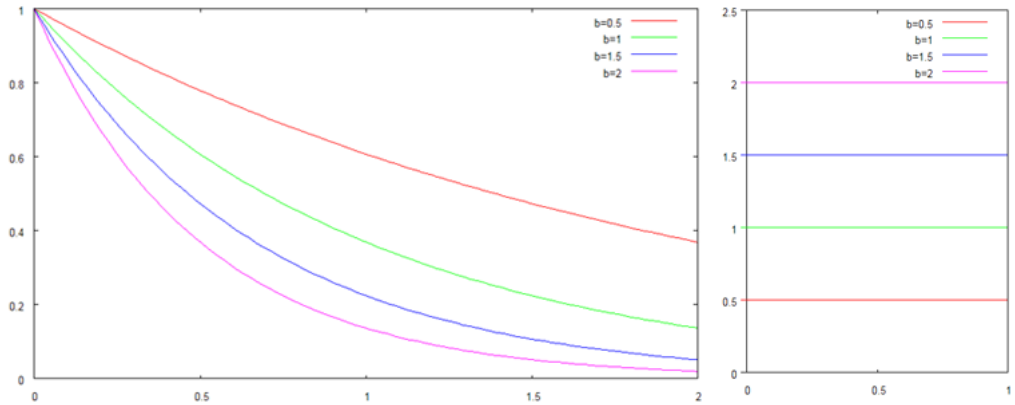

Abbildung 4.2: Survivorfunktion (variabel) und Hazardrate (konstant)

Die generelle Definition des Modells für Übergänge vom Ursprungszustand j in den Zielzustand k ist:

$$r_{jk}(t) \equiv r_{jk} = \exp\{\beta_{jk0} + B_{jk1}\beta_{jk1} + \ldots\} = \exp\{B_{jk}\beta_{jk}\}$$

r_{jk} = Zeitkonstante hazard rate vom Ursprungszustand j in den Zielzustand k Die *exit rate* (Rate des Verlassens des Ursprungszustand j in einen anderen Zielzustand k) ist definiert als:

$$r_j = \sum_{k \in \mathcal{D}_j}$$

wobei $\mathcal{D}_j$ das Set aller möglichen Zielzustände bezeichnet, die von j aus erreichbar sind.
Die Survivorfunktion $S(t)$ für die Verweildauer im Ursprungszustand j kann durch die exit rate ausgedrückt werden:

$$S_j(t) = \exp\left\{-\int_0^t r_j d\tau\right\} = \exp\{-tr_j\}$$

Annahme: $r_{jk}(t)$ kann zwischen verschiedenen Konstellationen von Kovariaten variieren, aber ist zeitkonstant. Mit anderen Worten: Es wird angenommen, dass der Prozess nicht Zeit-abhängig ist.
Der Zusammenhang zwischen der hazard rate und dem (Zeilen-) Vektor der Kovariaten A_{jk} ist als ist log-linear spezifiziert um sicherzustellen, dass die Schätzungen der hazard rate nicht negativ werden.
Der (Spalten-) Vektor der unbekannten Parameter α_{jk} und der Vektor der beobachteten Kovariaten A_{jk} sind spezifiziert im Hinblick auf den Ursprungszustand j und den Zielzustand k. Im Vektor der Parameter ist ein Term α_{jk0} enthalten, der auch dann geschätzt werden kann, wenn keine Kovariaten im Modell enthalten sind. Ein Modell ohne Kovariate wird geschätzt über:

$$r(t) \equiv r = \exp\{\beta_0\}$$

Solch ein simples Modell behandelt die Daten als ein Sample homogener Episoden. Es wird also von allen Merkmalen abstrahiert, die die Individuen unterscheiden, sie heterogen machen. Wir sind aber daran interessiert, Unterschiede zwischen verschiedenen Konstellationen von Merkmalen vereint in ihren Trägern zu entdecken. Der einfachste Weg dies zu erreichen ist, zeitkonstante Kovariaten mit ein zu beziehen. Bei zeitkonstanten Kovariaten sind die Werte dieser Kovariate für jedes Individuum über die Zeit unveränderlich. Es gibt zwei Arten zeitkonstanter Kovariaten: erstens solche, die -normalerweise- im Leben des Individuums konstant sind wie beispielsweise Geschlecht, soziale oder ethnische Herkunft (ascribed statuses). Zweitens solche, die vorher erlangt wurden und danach konstant bleiben, so wie beispielsweise höchster Bildungsabscluss oder Alter bei erster Heirat (statuses attained prior to).

4.1.2 Piecewise Constant Exponential Models

Hierbei handelt es sich um eine Abwandlung des einfachen Exponentialmodells, dass in der Anwendung äusserst nützlich sein kann. Nach Blossfeld und Rohwer (2002) ist seine Anwendung in zwei Fällen besonders in Betracht zu ziehen. *Erstens*, wenn der Forscher nicht in der Lage ist, wichtige zeitabhängige erklärende Variablen zu messen und in das Modell mit einzubeziehen

oder *zweitens*, wenn keine klare Vorstellung über die Form der Zeitabhängigkeit des Prozesses vorliegt. In dieser Art von Modell sind die Hazardraten stückweise konstant. Das bedeutet, dass die kontinuierliche Zeitachse in verschiedene, abzählbare Intervalle zerlegt wird. Innerhalb dieser Intervalle sind die Hazardraten jeweils konstant, unterscheiden sich jedoch in der Regel (aber nicht notwendiger Weise) zwischen den Intervallen.
In diesem Modell werden verschiedene intervallspezifische Konstanten geschätzt. Es gilt also:

Piecewise Constant Exponential Models

$h(t_l) = \exp\{\beta_{0l} + \beta_1 X_1 + \ldots + \beta_k X_k\}$
wobei der Index l anzeigt, dass für beliebige vom Anwender anzugebende Intervalle l jeweils eine spezifische Konstante geschätzt wird, die die "Basishöhe" der Hazardrate in diesem Intervall angibt.

Es wird angenommen, dass die Hazardraten piecewise constant, also frei übersetzt stückweise konstant sind. Dies bedeutet, dass konstant in jedem Intervall eines Sets aus Zeitintervallen.

Modelle mit Periodenspezifischen Effekten

Modelle mit Periodenspezifischen Effekten

$h(t_l) = \exp\{\beta_{0l} + \beta_{1l} X_1 + \ldots + \beta_{kl} X_k\}$
Hier werden für jedes Intervall l neben einer eigenen Regressionskonstante auch die Regressionskoeffizienten geschätzt.

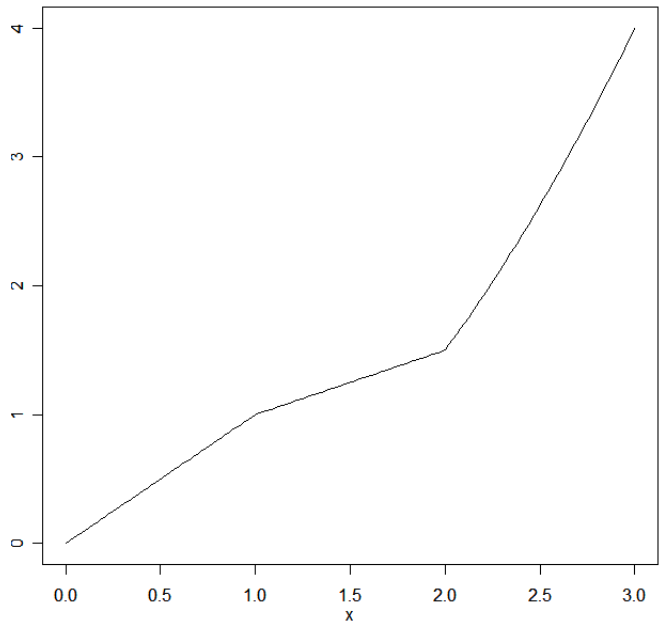

Abbildung 4.3: Beispiel einer Piecewise-Funktion

4.1.3 Weibull-Modell

In diesem Modell kann die Hazardrate nur auf eine ganz bestimmte Art monoton fallen oder steigen (siehe Graphik 4.4). In single transition Fällen wird dieses Modell durch Annahme einer Weibull-Verteilung für die Dauer der Episoden erlangt.

Weibull Modell

$$f(t) = ab^a t^{a-1} \exp\{-(bt)^a\} \ , \ a, b > 0$$

$$h(t) = ab^a t^{a-1}$$

$$S(t) = \exp\{-(bt)^a\}$$

für a=1 erhalten wir das Exponentialmodell.

wobei $b = \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}$, also:

$$h(t) = a \cdot \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}^a \, t^{a-1}$$

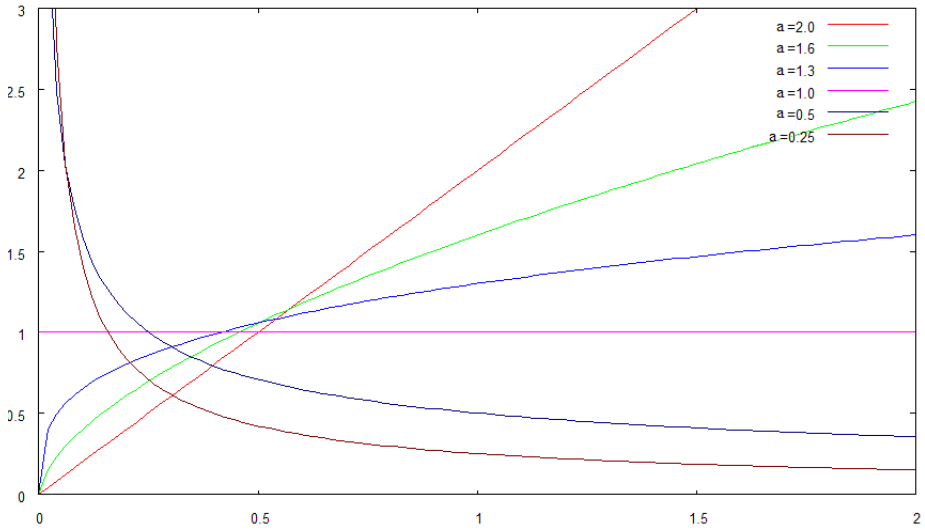

Abbildung 4.4: Hazardrate im Weibull-Modell

Bei $a > 1$ liegt eine steigende Hazardrate vor, bei $a < 1$ fällt sie. Ist $a = 1$ erhalten wir das Exponentialmodell, dass über eine Konstante Hazardrate verfügt.

4.1.4 Gompertz-Makeham Modell

Auch in diesem Modell kann die Hazardrate über die Zeit nur auf eine bestimmte Art monoton steigen oder fallen. Dies war auch schon beim Weibull-Modell der Fall. In Graphik 4.5 sehen wir den Unterschied. Beide Hazardraten, im Gompertz-Makeham Modell können nur monoton steigen oder fallen. Trotzdem sehen sie sich nicht gerade ähnlich. Das Modell ist definiert über:

Gompertz-Makeham Modell

$$f(t) = \exp\left\{-bt - \frac{a}{c}(\exp\{ct\}) - 1\right\}(b + a\exp\{ct\})$$

$$h(t) = b + a\exp\{ct\}$$

$$S(t) = \exp\left\{-bt - \frac{a}{c}(\exp\{ct\}) - 1\right\}$$

für $c = 0$ reduziert sich das Gompertz-Makeham Modell zum einfachen Exponentialmodell mit

$$f(t) = b \cdot \exp\{-bt\}, \; b > 0$$

$$S(t) = \exp\{-bt\}$$

wobei $b = \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}$, also:

$$h(t) = \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\} + a\exp\{ct\}$$

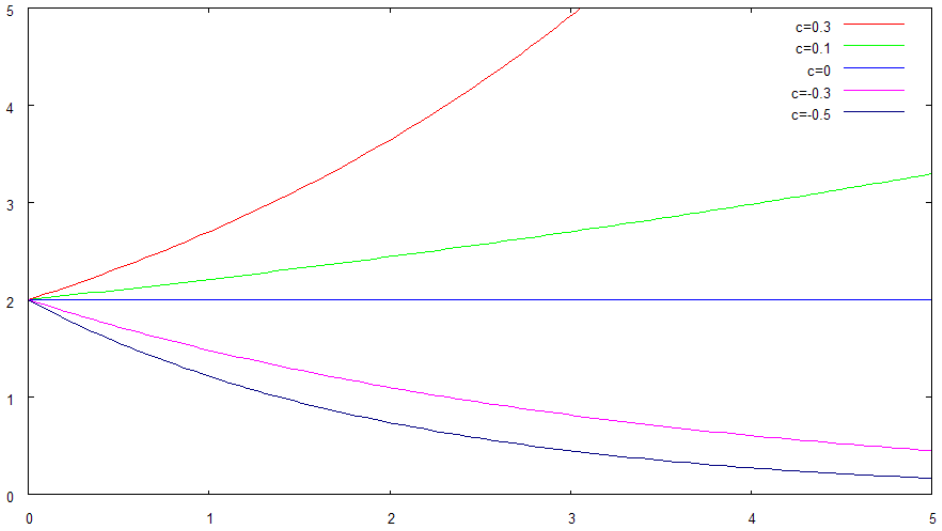

Abbildung 4.5: Hazardrate im Gompertz-Makeham-Modell

Für $b < 0$ fällt die Hazardrate, für $b > 0$ steigt die Hazardrate, für $b = 0$ erhalten wir das Exponentialmodell mit konstanter Hazardrate.

4.1.5 Intermezzo I

Nach Allison (1984) unterscheiden sich das Exponentialmodell, das Gompertz-Makeham Modell und das Weibull Modell nur dadurch, wie der Faktor Zeit in die Formeln eingebunden wird. Folgende Formeln werden angegeben, der übersichtlichkeit halber nur mit zwei erklärenden Variablen, die zeitkonstant sind:

Exponentialmodell	:	$\log(h(t)) = a + b_1x_1 + b_2x_2$
Gompertz-Makeham Modell	:	$\log(h(t)) = a + b_1x_1 + b_2x_2 + ct$
Weibull Modell	:	$\log(h(t)) = a + b_1x_1 + b_2x_2 + c \cdot \log(t)$

Wir haben es hier mit dem logarithmus der Hazardrate zu tun, da die rechten Seiten der Formeln negativ werden können. Dies macht aber bei (Eintritts-)Wahrscheinlichkeiten keinen Sinn. Deshalb wird der Logarithmus gebildet, um diesem Problem aus dem Weg zu gehen.
Wir sehen, dass im Exponentialmodell keine Zeitabhängigkeit der Hazardrate vorliegt. Sie ist Zeitkonstant. Im Gompertz-Makeham Modell hingegen verändert sich die Hazardrate linear mit der Zeit ($\log(h(t)) = a+b_1x_1+b_2x_2+ct$). Im Weibull Modell verändert sich die Hazardrate linear mit dem Logarithmus der Zeit ($\log(h(t)) = a+b_1x_1+b_2x_2+c \cdot log(t)$). Diese drei Modelle gehören alle der generellen Klasse von Modellen an, die als *proportional hazards models* bekannt sind.

4.1.6 Log-Logistisches Modell

Mit diesem Modell kann eine fallende oder eine zunächst steigende und dann fallende Hazardrate modelliert werden. Das Log-Logistische Modell ist definiert über:

4.1.7 Log-Logistsches Standardmodell

Log-Logistisches Standardmodell

$$f(t) = \frac{ab^a t^{a-1}}{(1+(bt)^a)^2}$$

$$h(t) = \frac{ab^a t^{a-1}}{1+(bt)^a}$$

$$S(t) = \frac{1}{1+(bt)^a}$$

wobei $b = \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}$, also:

$$h(t) = \frac{a \cdot \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}^a \, t^{a-1}}{1+(\exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}\, t)^a}$$

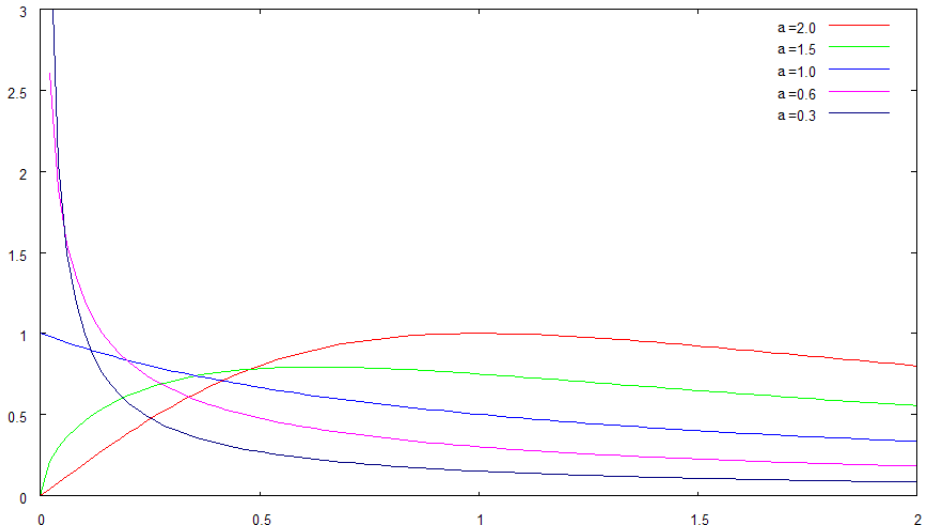

Abbildung 4.6: Hazardrate im Log-Logistischen Modell

Wir sehen, dass ein grösserer Wert für b von einer fallenden Hazardrate zu einer erst steigenden und dann fallenden Hazardrate führt

$$t_{\max} = \frac{1}{b}(a-1)^{\frac{1}{a}}$$

$$h_{\max} = b(a-1)^{1-\frac{1}{a}}$$

Erweitertes Log-Logistsches Modell

Erweitertes Log-Logistisches Modell

$$f(t) = c \cdot \frac{a(bt)^{a-1}}{(1+(bt)^a)^{\frac{c}{b}+1}} \quad a,b,c > 0$$

$$h(t) = c \cdot \frac{a(bt)^{a-1}}{1+(bt)^a}$$

$$S(t) = \frac{1}{(1+(bt)^a)^{\frac{c}{b}}}$$

wobei $b = \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}$, also:

$$h(t) = c \cdot \frac{a(\exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}\, t)^{a-1}}{1+(\exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}\, t)^a}$$

4.1.8 Log-Normale Modelle

Das Log-Normal-Modell unterstellt eine zunächst steigende und dann fallende Hazardrate. Im Log-Normalen Modell spielt die Normalverteilung eine wichtige Rolle. Sie ist definiert über

Dichtefunktion der Standardnormalverteilung:

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\}$$

Verteilungsfunktion der Standardnormalverteilung::

$$\Phi = \int_0^t \varphi(\tau) d\tau$$

und ist wie folgt in das Log-Normale Modell implementiert:

Log-Normal Modell

$$f(t) = \frac{1}{at}\varphi\left(\frac{\log(t)-b}{a}\right) \ , \ a > 0$$

$$h(t) = \frac{1}{at}\frac{\varphi(z_t)}{1-\Phi(z_t)} \text{ mit } z_t = \frac{\log(t)-b}{a}$$

$$S(t) = 1 - \Phi\left(\frac{\log(t)-b}{a}\right)$$

wobei $b = \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}$, also:

$$h(t) = \frac{1}{at}\frac{\varphi(z_t)}{1-\Phi(z_t)} \text{ mit } z_t = \frac{\log(t) - \exp\{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\}}{a}$$

4.1.9 Intermezzo II

Im Unterschied, zu den proportional hazards models, von denen wir in *Intermezzo I* (4.1.5) gelesen haben, gehören das log-normale und das log-logistische Modell einer anderen Klasse von Modellen an. Hierbei handelt es sich um *accelerated failure time models*, oder auch *location-scale models* genannt. Wenn T die Zeit beschreibt, bis ein Ereignis auftritt, dann kann diese Klasse von Modellen wie folgt beschrieben werden:

$$log(T) = a + b_1 x_1 + b_2 x_2 + \ldots + u$$

wobei u ein Zufälliger Zufallsterm bezeichnet, der unabhängig von x_i ist. Dieser Zufallsterm u ist für die Unterschiede zwischen den Mitgliedern dieser Modellfamilie zuständig. Verteilungen, die oftmals angenommen werden umfassen die Norma´lverteilung, log-gamma Verteilung, logistische Verteilung und die extreme-value Verteilung. Daraus ergeben sich die Verteilungen für T, die wir als log-normalen und die log-logistischen Modelle kennen, sowie das Gamma Modell, auf das nicht näher eingegangen wird. Ebenso treffen wir hier auf das Weibull-Modell, das in beide Klassen eingeteilt werden kann. Es kann gezeigt werden, dass das Weibull Modell (sowie sein Spezialfall, das Exponentialmodell) das einzige Modell ist, dass in beide Klassen fällt. Für verschiedene Verteilungen von u ergeben sich:

Verteilung von u	$\rightarrow$	Resultierende Verteilung
Normalverteilung	$\rightarrow$	Log-Normal
Log-Gamma	$\rightarrow$	Gamma
Logistisch	$\rightarrow$	Log-Logistisch
Extreme-Value	$\rightarrow$	Weibull

Das log-normale und das log-logistische Modell sind unter dem Blickwinkel etwas Besonderes, da sie (wie in den Graphiken 4.6 und **??** zu sehen) dazu geeignet sind, eine erst steigende und dann -nach einem Maximalwert- fallende Hazardrate zu modellieren. Ihre Hazardraten sind nicht monotone Funktionen der Zeit

4.1.10 Sichelmodell / Sickle-Model

Auch dieses Modell kann zunächst steigende und dann fallende Raten modellieren. Die Form ist glockenförmig, einer Sichel ähnlich. Es ist definiert über:

Sickle-Modell

$$f(t) = \exp\left\{-ba\left[a-(t+a)\exp\left\{-\frac{t}{a}\right\}\right]\right\} bt\exp\left\{-\frac{t}{a}\right\}$$

$$h(t) = bt\exp\left\{-\frac{t}{a}\right\} \;,\; a,b>0$$

$$S(t) = \exp\left\{-ba\left[a-(t+a)\exp\left\{-\frac{t}{a}\right\}\right]\right\}$$

wobei $b = \exp\{\beta_0+\beta_1X_1+\ldots+\beta_kX_k\}$, also:

$$h(t) = \exp\{\beta_0+\beta_1X_1+\ldots+\beta_kX_k\}\, t\exp\left\{-\frac{t}{a}\right\}$$

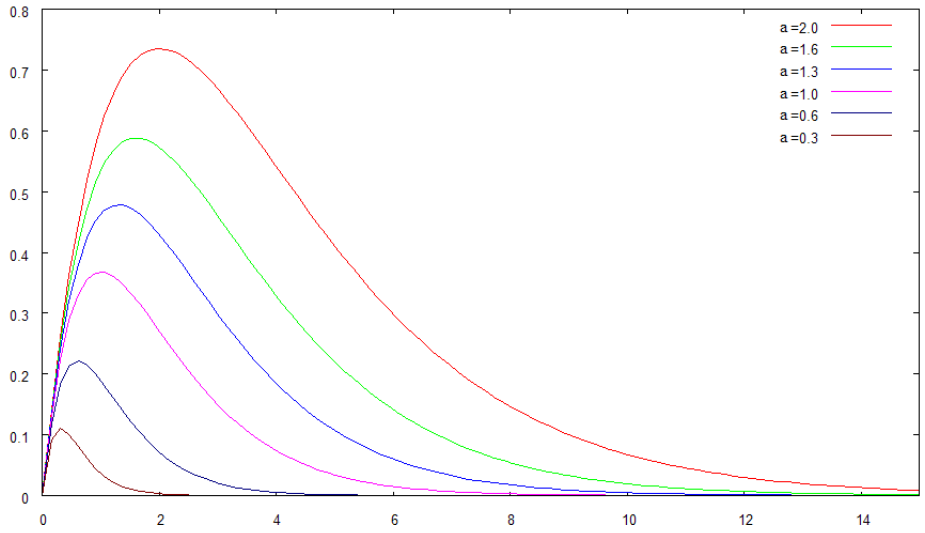

Abbildung 4.7: hazard rate im Sickle-Modell

Das Maximum der Rate liegt bei $t = a$ und der einzige Wendepunkt bei $t = 2a$. Eine Besonderheit dieses Modells ist, dass die Survivorfunktion nicht gegen 0 tendiert, sondern gegen

$$\exp-\left\{(\beta_0+\beta_1X_1+\ldots+\beta_kX_k)a^2\right\}$$

Mit anderen Worten, dieses Modell ist vor allem dann angemessen, wenn man annimmt, dass nicht für alle Personen ein Ereignis eintritt. Es ist z.B. für die Analyse von Ehescheidungen gut geeignet.

4.1.11 Letzte parametrische Bemerkung

Als Abschlussbemerkung sei erwähnt, dass nach Allison *kein* überzeugendes parametrisches Modell existiert, um eine u-förmige Hazardrate zu modellieren. Auch in anderen Lehrbüchern lässt sich dazu nichts finden. Ebenso sei es oftmals besser, bei starker Abweichung von monotonen Steigungseigenschaften auf das semiparametrische proportional hazards model zurückzugreifen. Der nicht ganz unberechtigte Einwurf, dass eine sozialwissenschaftliche Theorie kaum Hinweise geben kann, ob eher ein Weibull- oder Gompertz-Makeham Modell angebracht ist vereinfacht uns die Analyse auch nicht.
Im Hinblick auf das Cox-Modell werden zwei weitere Nachteile erwähnt:

1. Die Entscheidung darüber, wie die Hazardrate von der Zeit abhängt, worüber wir oftmals kaum Informationen haben. Desweiteren ist die Wahl des passenden Modells mit der Richtigen Form problematisch, wenn wir eine nichtmonotone Hazardfunktion erwarten.

2. Wichtiger als dieses mag jedoch sein, dass es die angeführten Modelle -nach Allison- nicht erlauben, erklärende Variablen mit aufzunehmen, deren Werte sich über die Zeit verändern.

4.2 Semi-Parametrische Modelle: die Regression nach Cox

Das Cox-Modell ist nach Yamaguchi (1991) die populärste Regressionsmethode zur Analyse von Überlebensdaten. Sie findet besonderen Anklang in der demographischen Forschung, beispielsweise in der Untersuchung von Heirat, Scheidung, Geburt, Migration und Job-Mobilität. Ebenso bescheinigt ihr Allison (1984) grosse Beliebtheit, als Beispiel nennt er die biomedizinische Forschung. Das stetige Cox-Modell wird auch als proportionales Hazard Modell (proportional hazards model) bezeichnet. Es beruht auf der *Partial Likelihood* und nicht auf der *Maximum Likelihood* Methode. Der wichtigste Vorteil dieser Partial Likelihood ist die Möglichkeit, Zeitabhängigkeiten zu modellieren, ohne dass eine Annahme über die Form getroffen werden muss. Ein weiterer Vorteil ist die Fähigkeit des Cox-Modells, stratifizierte Modelle umzusetzen. Stratifizierte Modelle erlauben es uns, eine oder mehrere kategoriale Kovariate zu kontrollieren, die komplizierte Interaktionseffekte mit der Zeit aufweisen können, ohne die Form dieser Interaktionseffekte spezifizieren zu müssen. Das Cox-Modell krankt allerdings auch an mindestens vier Nachteilen.

1. Dieses Modell nutzt nur die Information über die relative Reihenfolge der Verweildauern anstelle der exakten Zeitpunkte der Ereignisse und Zensierungen. Der Informationsverlust ist also möglicherweise äusserst gross. Dieser Verlust an Präzision der Partial Likelihood Parameterschätzer im Vergleich zu den Maximum Likelihood Schätzern verschwindet normalerweise immer mehr, je größer die Stichprobe wird, kann aber problematisch sein, wenn nur eine kleine Stichprobe vorhanden ist.

2. Die Handhabung von Ties ist problematisch. Als Daumenregel sollten nicht mehr als 5% der Fälle Ties sein. Die Partial Likelihood Methode kann Ties nicht exakt handhaben, dies ist rechnerisch unerschwinglich. Deshalb werden sie in Programmen, die zur Berechnung der Cox-Methode verwendet werden, approximiert. Diese Annäherung ist jedoch bei einer grossen Anzahl von Ties bestenfalls fragwürdig. Nach Yamaguchi ist dann die ML-Methode vorzuziehen, insbesondere mit diskreten Zeitmodellen.

3. Die Analyse der Form der Zeitabhängigkeit ist mit der PL-Methode nicht möglich. Ist diese von Interesse, dann ist die Anwendung des Cox-Modells eine fruchtlose Angelegenheit.

4. Die PL-Methode basiert auf schwächeren theoretischen Grundlagen als die ML-Methode. Bei der Modellauswahl sind Vorsichtsmaßnahmen empfohlen.

Trotz dieser Nachteile ist das Cox-Modell ungebrochen beliebt bei der Analyse von Ereignisdaten. Wieder einmal unterscheiden sich die Darstellungen dieser Methode je nach Lehrbuch. Mir erscheint es sinnvoll, hier kurz Allison (1984) und Yamaguchi (1991) zusammenfassend darzulegen, da sich durch die verschiedenen Herangehensweisen möglicherweise ein Gewinn an Durchblick erzielen lässt.

4.2.1 Cox-Modell, Notation nach Allison

Das Cox-Modell -öfter auch proportionales Hazardmodell bezeichnet- ist nach Allison (wobei es sich der Eienfachheit halber bei x_1 und x_2 um zeitkonstante erklärende Variablen handelt) definert als:

$$log(h(t)) = a(t) + b_1x_1 + b_2x_2$$

$a(t)$ kann hierbei jede Funktion der Zeit sein. Weil diese Funktion nicht spezifiziert werden muss, wird dieses Modell als semiparametrisch oder partiell parametrisch bezeichnet. Es wird proportionales Hazardmodell genannt, weil für alle zwei Individuen zu jedem Zeitpunkt folgendes gilt:

$$\frac{h_i(t)}{h_j(t)} = c, \text{ für jeden Zeitpunkt } t$$

c kann dabei von den erklärenden Variablen abhängen, nicht jedoch von der Zeit. Im Gegensatz zu dem Namen ist dies keine entscheidende Eigenschaft des Modells, weil die Konstanz der Hazard-Ratios abhanden kommt, wenn zeitveränderliche unabhängige Variablen eingeführt werden. Es ist natürlich einfacher, solch ein Modell aufzustellen als es zu schätzen. Hier ziegt sich das wichtige an Cox's Modell: die *Partial Likelihood Methode*. Diese Methode beruht auf der Tatsache, dass die Likelihoodfunktion für Daten aus dem proportionalen Hazardsmodell in zwei Teile zerlegt werden kann: Der eine Faktor entält nur die Information über die Koeffizienten b_1 und b_2. Der andere Faktor enthält Informationen über b_1, b_2 und die Funktion $a(t)$. Die Partial Likelihood Methode ignoriert einfach den zweiten Faktor und behandelt den Ersten als ganz normale Likelihoodfunktion. Dieser Faktor hängt nur von der Reihenfolge ab, in der die Ereignisse eintreten, nicht jedoch von dem exakten Zeitpunkt ihres Eintretens. Die daraus resultierenden Schätzer sind asymptotisch unverzerrt und normalverteilt. Sie sind nicht komplett effizient,

da ein Teil der Information (der genaue Zeitpunkt des Eintretens) von dem Verfahren ignoriert bleibt. Dieser Malus an Effizienz ist jedoch normalerweise so gering, dass er nach Allison (1984) nicht der Sorge Wert ist. Wenn die Abhängigkeit des Hazards von der Zeit von Bedeutung ist, lässt sich das Cox-Modell nicht anwenden. Als Beispiel wird das Prinzip der kumulativen Inertia angeführt, die besagt, dass die Wahrscheinlichkeit eines Individuum, seinen Zustand zu ändern abnimmt, je länger es schon in diesem Zustand verharrt. Ist jedoch nur der Effekt der erklärenden Variablen von Bedeutung, und nicht die Abhängigkeit von der Zeit, dann ist das Cox-Modell eine interessante Option.

Zeitveränderliche erklärende Variablen

Das proportionale Hazardmodell kann leicht um erklärende Variablen erweitert werden, die ihre Werte über die Zeit ändern. Hier wird ein Modell aufgeführt, in dem eine der beiden unabhängigen Variablen zeitkonstant ist, die andere zeitveränderlich.

$$log(h(t)) = a(t) + b_1x_1 + b_2x_2(t)$$

Dieses Modell besagt, dass der Hazard zur Zeit t vom Wert der Variable x_2 zum gleichen Zeitpunkt t abhängt. Wenn man Grund zur Annahme hat, dass der Effekt der Variable x_2 zeitverzögert eintritt, kann man dies leicht in die Formel einfliessen lassen.

$$log(h(t)) = a(t) + b_1x_1 + b_2x_2(t - v)$$

Dies ist die generelle Form, dies zu tun. Wenn die Zeit in Monaten gemessen wurde und wir annehmen, dass der Effekt um 3 Monate zeitverzögert wirkt, dann setzen wir für v einfach 3 ein, also:

$$log(h(t)) = a(t) + b_1x_1 + b_2x_2(t - 3)$$

Ein heutzutage obsolet anmutender, jedoch erwähnenswerter Hinweis von Allison soll hier nicht verschwiegen werden: Bei Aufnahme von zeitveränderlichen unabhängigen Variablen in das Modell steigt die Rechenzeit enorm an. Allein die Aufnahme einer zeitveränderlichen Variablen erhöhte die Rechenzeit um den Faktor 10.

4.2.2 Cox-Modell, Notation nach Yamaguchi

Ergänzend dazu ist proportionale Hazardmodell nach Yamaguchi (1991) definiert als:

Cox-Modell

$$h_i(t) = h_0(t) \exp\left\{\sum_k b_k X_{ik}(t)\right\}$$

Die Hazardrate ist definiert als das Produkt einer unspezifizierten *Baseline*-Funktion $h_0(t)$ und einem zweiten Term der den möglichen Einfluss eines Kovariatvektors $X_{ik}(t)$ (für Person i zum Zeitpunkt t und Kovariate k) auf die Hazardrate angibt. Der Effekt der Kovariaten kann proportionale Änderungen der Hazardrate bewirken. Deshalb sollte das Cox-Modell nur verwndet werden, wenn diese Proportionalitätsannahme gerechtfertigt ist. Das Modell nimmt an, dass wenn X_k eine Intervallskalierte Variable ist, sich die Hazardrate mit jeder einheit der intervallskalierten Variable um $\exp\{b_k\}$ vervielfacht, sofern der Effekt der anderen Kovariate kontrolliert ist.
Wenn die Kovariate alle Zeitunabhängig sind, dann ist die Survivorfunktion gegeben durch:

$$S_i(t) = S_0(t)^{\exp\left\{\sum_k b_k X_{ik}\right\}}$$

wobei $S_0(t)$ die Survivorfunktion für die Individuen mit $X_k = 0$ angibt. Sie ist gegeben über:

$$S_0(t) = \exp\left\{-\int h_0(s)ds\right\}$$

Die *log minus log Survivorfunktion* ist gegeben über:

$$\ln - \ln S_i(t)] = \ln\left[-\ln S_0(t)\right] + \sum_k b_k X_{ik}$$

Der erste Teil der Formel auf der rechten Seite ist allen Objekten gemeinsam, der zewite Teil ist nicht Zeitabhängig. Es folgt, dass wenn alle Kovariate Zeitunabhängig sind, die Differenz der log minus log Survivorfunktion unter den Gruppen mit unterschiedlichen Werten auf den Kovariaten, über Zeit konstant werden. Diese Charakteristik kann in einer graphischen Überprüfung der nonproportionalen Effekte für eitunabhängige Kovariate verwendet werden.

Anhang A

Variablen: diskret & stetig

Wichtig für dieses Skript ist die Unterscheidung zwischen diskreten und Stetigen Variablen. Es vermindert meiner Erachtens enorm die Verwirrung, wenn man weiss, warum "manchmal" mit $\sum$ und "manchmal" mit $\int f(x)dx$ gerechnet wird.

A.1 Diskret Variablen

Bei diskreten Variablen handelt es sich um Variablen, deren Ausprägungen endlich oder abzählbar unendlich sind. Uns bekannte diskrete Wahrscheinlichkeitsverteilungen sind die Hypergeometrische Verteilung, die Binomialverteilung oder die Poissonverteilung. In jeder dieser Verteilungen existiert ein Term aus der Kombinatorik, so dass sie schon intuitiv als abzählbar erkannt werden können:

$$Hypergeometrisch = \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}}$$

$$Binomialverteilung = \binom{n}{x} p^x (1-p)^{n-x}$$

$$Poissonverteilung = \frac{\mu^k}{k!} e^{-\mu}$$

In jeder dieser Formeln steht entweder $\binom{n}{k}$ oder $x!$. Diese Werte mögen zwar extrem grosswrden, unendlich sind sie jedoch nicht. Beim Lotto z.B. existieren $\binom{49}{6} = 13.983.816$ möglich Lottoziehungen. Die Ausprägungen die die Variable "Richtige im Lotto" annehmen kann besitzt aber nur die Ausprägungen 1, 2, 3, 4, 5, 6 Richtige und nicht zu vergessen 0 Richtige. -2 Richtige oder 3,5 Richtige sind *nicht* möglich!

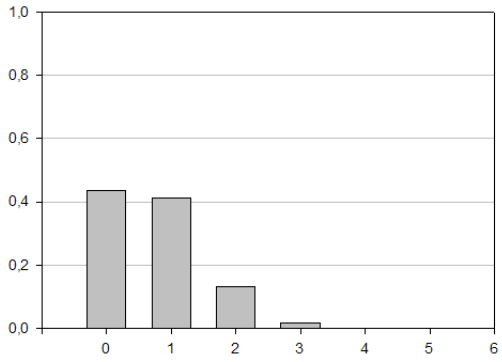

Abbildung A.1: Wahrscheinlichkeit für Gewinn

In Graphik A.1 sehen wir, wie hoch die jeweilige Wahrscheinlichkeit ist, die entsprechende Anzahl an Treffern im Lotto zu erzielen. Es sind zwar Balken in diesem Diagramm zur besseren Übersicht dargestellt, aber wollen wir präziser von "Strichen" reden, da die Ausprägungen keine Intervalle darstellen, sondern Punkte. Wir haben hoffentlich 6 Richtige im Lotto und nicht zwischen 5,9 und 6,1, denn dies ist nicht möglich. Die Wahrscheinlichkeit für 1 und 0 Richtige liegen relativ gleich auf, sie entspricht ungefähr 0,42. Die Wahrscheinlichkeit für 4 bis 6 Richtige ist mit blossem Auge in der Graphik nicht mehr zu erkennen, sie beträgt für 4 Richtige $\approx 0,00096862$ oder anders geschrieben $\approx 9,6862E-04$, für 5 Richtige $\approx 1,845E-05$ und für 6 Richtige $\approx 7,1511E-08$, also $\frac{1}{13.983.816}$. Es ist schon intuitiv logisch, dass die Wahrscheinlichkeit, sofern man denn mitgespielt hat, Eines dieser Ergebnisse zu erhalten, nämlich 0, 1, 2, 3, 4, 5 oder 6 Richtige zu haben eintreten muss. Die kumulierte Wahrscheinlichkeit muss also exakt = 1 betragen.

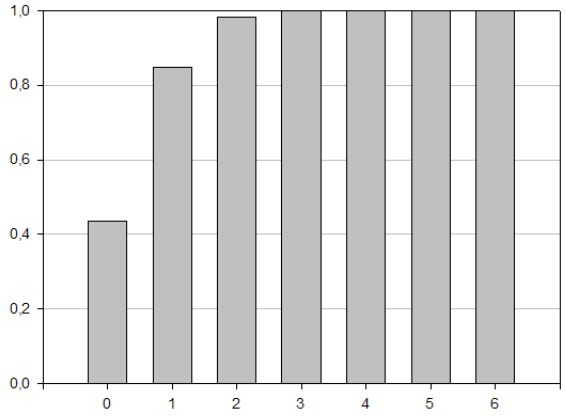

Abbildung A.2: Kumulierte Wahrscheinlichkeit

Wenn wir uns die kumulierte Wahrscheinlichkeit anzeigen lassen sehen wir, dass die Summe 1 ergibt. Auch wenn es in der Graphik A.2 optisch nicht deutlich wird, 1 wird erst mit dem letzten Strich erreicht, vorher liegt der Wert der kumuliertn Wahrscheinlichkeit zwar *sehr* nahe an 1, er ist jedoch noch kleiner als 1.

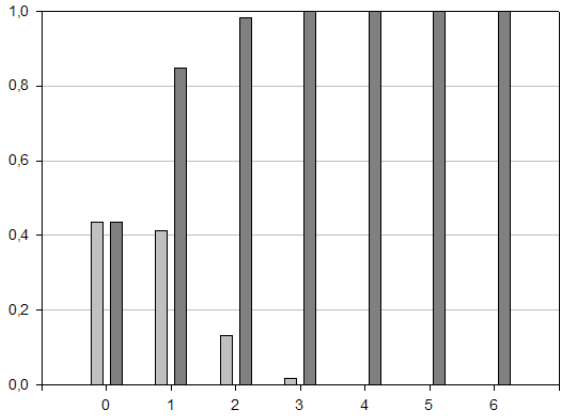

Abbildung A.3: Wahrscheinlichkeit & kumulierte Wahrscheinlichkeit

Für unser hypergeometrisch verteiltes Lotto-Beispiel gilt folgendes,

$$\sum_{i=0}^{6} p(x_i) = 1$$

da wir es mit abzählbaren Ausprägungen zu tun haben, und daraus die Summe bilden können, was wir eben getan haben.

A.2 Stetige Variablen

Bei stetigen Variablen liegt der Fall anders, hier haben wir es mit überabzählbar unendlich vielen Eigenschaften oder Ausprägungen zu tun. Wir können -anders als bei den Lottoergebnissen- jede Ausprägung noch genauer Messen. Eine Person kann theoretisch beliebig genau gewogen, in der Grösse vermessen oder ihr Alter bestimmt werden. So ist es beispielsweise möglich eine Person nicht "grob" auf 1.75m - 1.76m in ihrer Grösse zu messen, sondern anzugeben, ob sie 1.75m, 1.754m, 1.7548m oder 1.75482m gross ist. zwischen jeden beliebigen zwei Messwerten liegen unendlich viele andere. Wir haben hier also nicht nur 7 Striche vorleigen wie in unserem Lotto-Beispiel. Nicht einmal 100 Striche. Auch 1.000, 5.000 oder 523.495.685.932 Striche genügen nicht. Da die Anzahl der Ausprägungen gegen unendlich geht, liegen die Striche unendlich dicht beieinander. Und damit sind wir sehr nahe an einer wichtigen Schlussfolgerung. Wonach sehen unendlich viele Striche unendlich nahe beieinander aus? Erinnern wir uns, an unsere ersten ausmal-Versuche in der Grundschule oder dem Kindergarten. Richtig. Sie sehen aus wie ein Fläche. Flächen berechnet man in der Mathematik über Integrale. Also heisst dies für uns, wir rechnen nicht

$$\sum_{i=1}^{\infty} p(x) = 1$$

denn dies würde unendlich lange dauern, sondern

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

In den nachfolgenden Graphiken sehen wir, das wir, wie sich aus einer Ansammlung von Strichen eine Fläche entwickelt. Es ist der Auflösung des PC-Bildschirms geschuldet, dass schon bei einer relativ "ungenauer Messung" bestehend aus 0.01er Schritten (Bild unten rechts) die Ansammlung der Striche als Fläche erscheint.

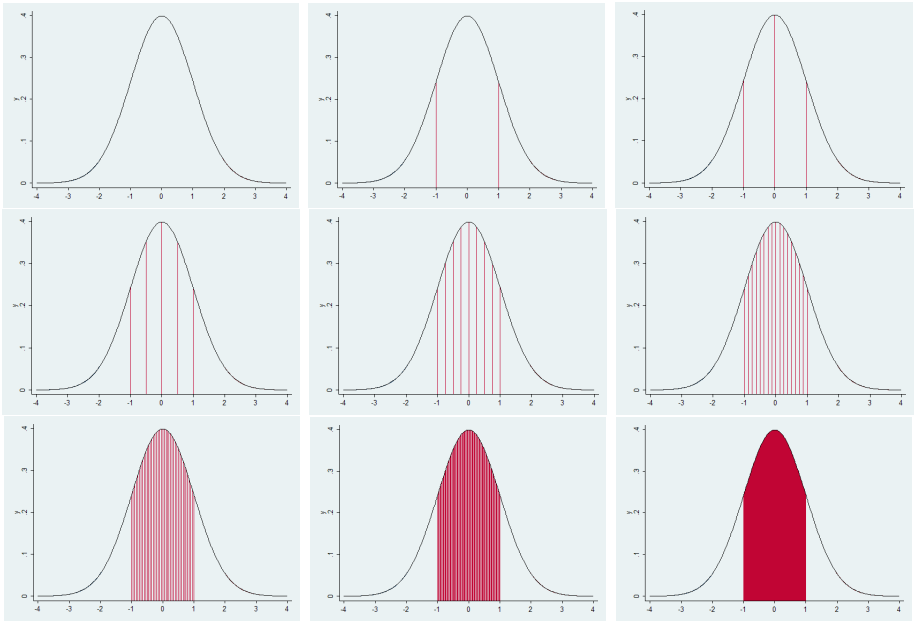

Abbildung A.4: Abnehmender Abstand zwischen den Messungen

Anhang B

Dichtefunktion & Verteilungsfunktion

Eine Dichtefunktion, auch als Wahrscheinlichkeitsdichte oder Wahrscheinlichkeitsdichtefunktion bekannt, dient dazu die Wahrscheinlichkeitsverteilungen einer Variablen zu beschreiben. Die Wahrscheinlichkeiten für die einzelnen Ausprägungen einer stetigen Zufallsvariablen können im Gegensatz zum diskreten Fall nicht angegeben werden, da die Wahrscheinlichkeiten für jede einzelne Ausprägung 0 sind, da die Intervalle gegen Null gehen, und damit die Wahrscheinlichkeit, in ein bestimmtes Intervall zu fallen ebenfalls gegen Null gehen. Es lassen sich nur Wahrscheinlichkeiten dafür angeben, dass die Werte innerhalb eines Intervalls um den interessierenden Wert x liegen. Die Wahrscheinlichkeit, dass die Zufallsvariable Werte zwischen a und b annimmt, entspricht dem Integral der Funktion. Es gilt

$$P(a < x < b) = \int_a^b f(x)dx = \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx$$

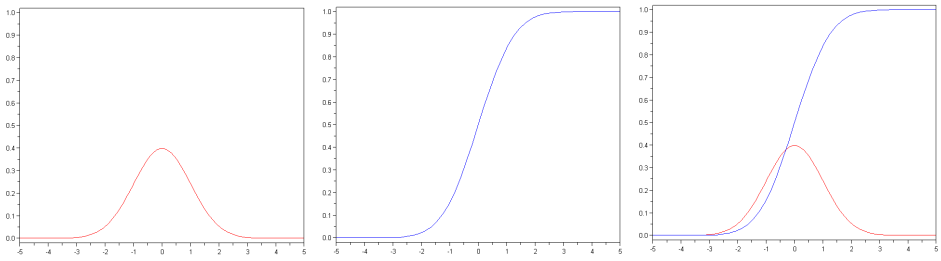

Abbildung B.1: Dichtefunktion und Verteilungsfunktion der Standardnormalverteilung

Anhang C

Grundlagen der Analysis

Ein paar Grundlagen in Analysis scheinen mir auserordentlich nützlich, um die Zusammenhänge in diesem Skript besser nachvollziehen zu können.

C.1 Ausgangsfunktion $f(x)$

Funktionen die wir kennen, die kennen wir üblicherweise urch ihre normale Funktion, die ich in diesem Zusammenhang "Ausgangsfunktion" nennen möchte. Die Parabel der Funktion $f(x) = x^2$ zeigt sich nur in ihrer Ausgangsfunktion.

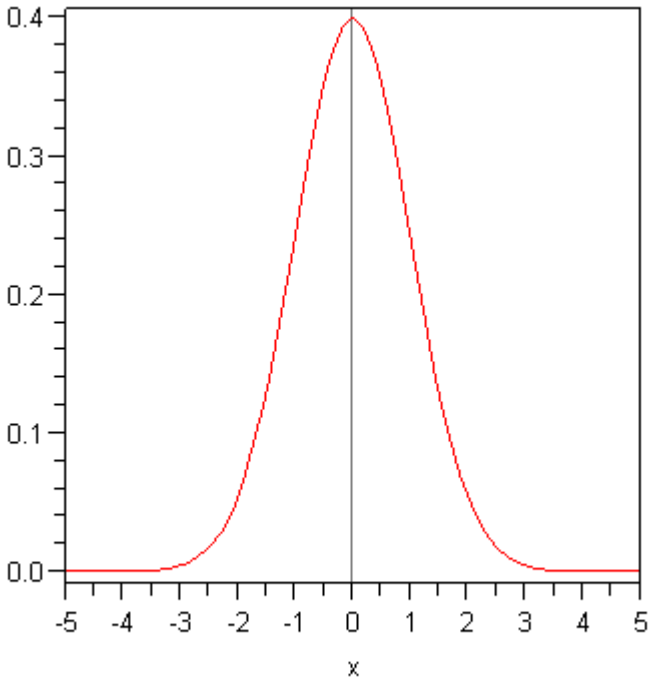

Abbildung C.1: Ausgangsfunktion der Standardnormalverteilung

In der vorangehenden Graphik C.1 sehen wir die uns bekannte Standardnormalverteilung (im folgenden: SNV). Ihr typischer brachte ihr den Namen Glockenfunktion oder Gaussche Glocke ein. Doch auch hier zeigt sich der Verlauf nur in der Ausgangsfunktion $f(x)$. Wie sich die Form, und die Interpretation verändert, wenn man aus dieser Funktion die Ableitung $f'(x)$ oder die Stammfunktion $F(x)$ bildet sehen wir nun:

C.2 Stammfunktion $F(x)$

Die Stammfunktion wird über intergrieren der Ausgangsfunktion gewonnen. Das Integral einer Funktion beschreibt den Flächeninhalt zwischen Kurve der Funktion und der $x - Achse$. Der Flächeninhalt unter der gesamten Kurve wird über

$$F(x) = \int_{-\infty}^{\infty} f(x)dx$$

beschrieben. Allerdings ist es allgemein notwendig sich die Teilstücke zwischen den Nullstellen der Ausgangsfunktion gesondert anzuschauen. Dies ist

hier jedoch nicht erforderlich, da die Ausgangsfunktion der SNV über keine Nullstellen verfügt. In Graphik C.2 sehen wir einen bekannten Sachverhalt. Die Stammfunktion nimmt für $x = 0$ den Wert 0.5 an. Dies ist der Wert, den der Flächeninhalt zwischen Kurve und x-Achse annimmt. Da der Flächeninhalt der Gesamten SNV (von $-\infty$ bis ∞) gleich 1 ist, und die SNV symmetrisch zum Ursprung ist, war dieser Wert erwartet. Der Wert der Stammfunktion gibt den Flächeninhalt von $-\infty$ bis zu dem Punkt an, der auf der x-Achse abgelesen wird.

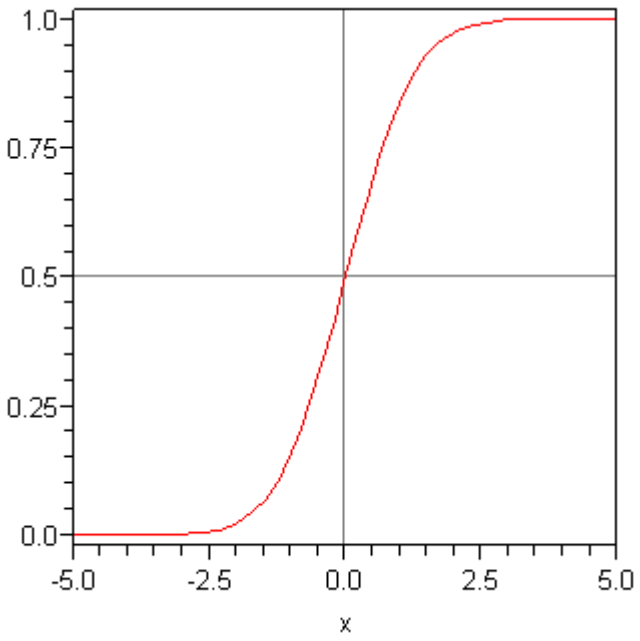

Abbildung C.2: Ausgangsfunktion der Standardnormalverteilung

Es gilt also für die Standardnormalverteilung:

$$F(0) = \int_{-\infty}^{0} f(x)dx = 0.5$$

Andere bekannte Werte für die Stammfunktion der SNV sind:

$$F(1.645) = \int_{-\infty}^{1.645} f(x)dx \approx 0.95$$

$$F(1.96) = \int_{-\infty}^{1.96} f(x)dx \approx 0.975$$

$$F(2.326) = \int_{-\infty}^{2.326} f(x)dx \approx 0.99$$

Hier lüftet sich vielleicht etwas der kryptische Schleier der Z-Werte, den mit denen hatten wir es gerade zu tun.

C.3 Erste Ableitung $f'(x)$

Die erste Ableitung der Ausgangsfunktion hat eine ebenso interessante Interpretation. Die erste Ableitung beschreibt die Steigung der Tangente, die die Kurve in dem Punkt berührt, der auf der x-Achse abgelesen wird. Wir können in Graphik C.3 shen, in welchen Bereichen die Ausgangsfunktion der SNV steigt (positiver y-Wert = positive Steigunt), fällt (negativer y-Wert = negative Steiung, die Funktion fällt also der x-Achse entgegen) und wo die Funktion keine Steigung besitzt. Dies ist hier bei $x = 0$ der Fall. Über die Struktur der Ableitungen kann man eine Menge über die Charakteristik der Funktion erfahren. Z.B. können wir errechnen, wo die Funktion eine Extremstelle besitzt, indem wir die erste Ableitung $f'(x) = 0$ setzen und diese Gleichung lösen. Um jedoch zu entscheiden, ob es sich um ein Minimum oder Maximum handelt, reicht die Aussage $f''(x) \neq 0$ nicht, sie gibt nur an, dass es sich um eine Extremstelle handelt. Also müssen wir die zweite Ableitung genauer untersuchen. Ist $f''(x) < 0$ handelt es sich um ein Maximum, ist $f''(x) > 0$ handelt es sich um ein Minimum. Wir erinnern uns wahrscheinlich dunkel an die Begrifflichkeiten von notwendiger und hinreichender Bedingung. Für Wendestellen müssen wir untersuchen, ob und wo $f''(x) = 0$ gilt. Ist in diesem Punkt $f'''(x) \neq 0$, so haben wir es mit einem oder mehreren Wendepunkten zu tun. Bei der SNV haben wir 2 Wendepunkt vorliegen, bei 1 und -1. Generell gilt für Normalverteilungen: Wendepunkte bei $\pm 1 \cdot \sigma$. Wir können in Graphik C.3 gut erkennen, dass die Ausgangskurve ihr Steigungsverhalten ändert.

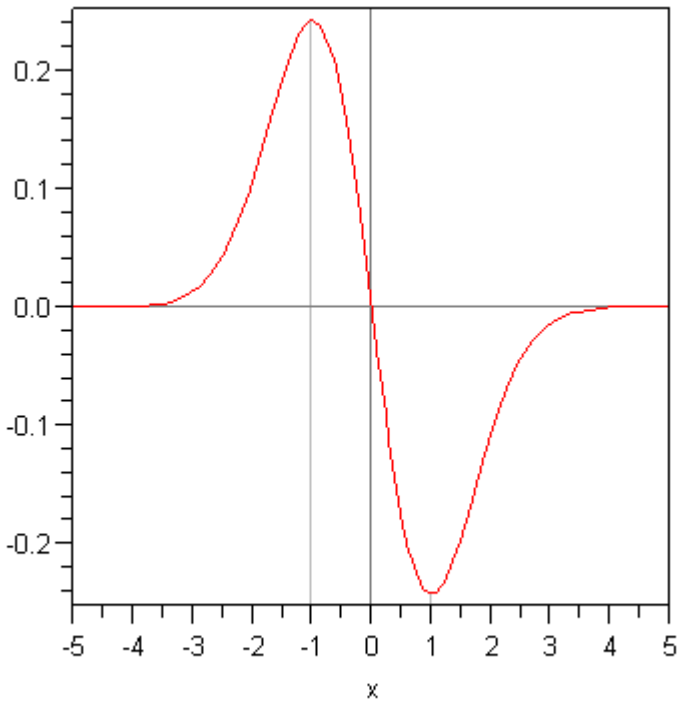

Abbildung C.3: Erste Ableitung der Standardnormalverteilung

Erst steigt die Kurve (der ersten Ableitung) an, dies bedeutet eine grösser werdende Steigung in der Ausgangsfunktion (!) also eine Linkskurve. Am Punkt -1 verharrt die Ableitung kurz und fällt dann, was eine Rechtskurve für die Ausgangsfunktion bedeutet, bis zum Punkt $+1$. Dort verharrt die Kurve der Ableitung auch infinitesimal kurz und beginnt dann wieder zu steigen, was einer Linkskurve für die Ausgangsfunktion gleichkommt.

C.4 Beispiel einiger Funktionen

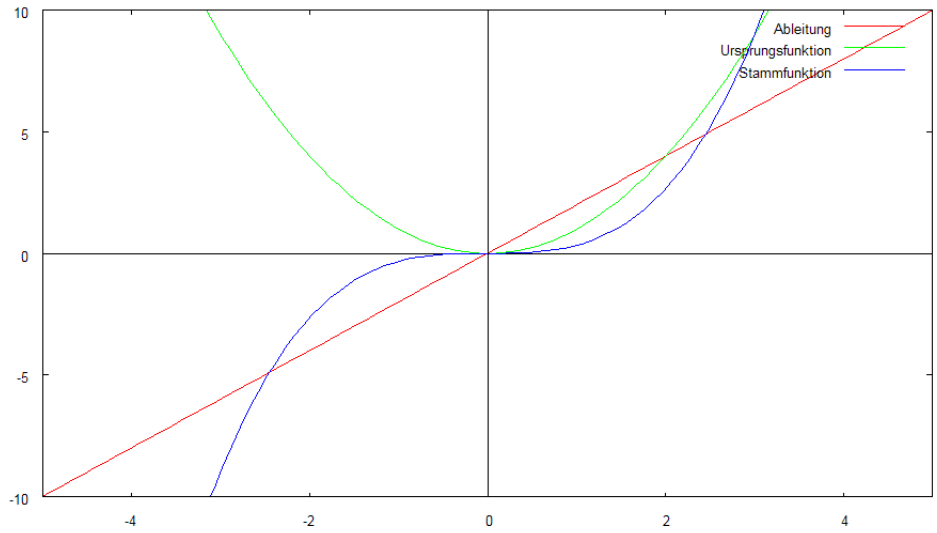

Abbildung C.4: Erste Ableitung und Stammfunktion von $f(x) = x^2$

Wobei:

$$f'(x) = 2x$$

$$f(x) = x^2$$

$$F(x) = \frac{1}{3}x^3$$

Tabelle C.1: Ableitungs- & Integrationsregeln

$f''(x)$	$f'(x)$	$f(x)$	$F(x)$
0	0	a	ax
$n(n-1)x^{n-2}$	nx^{n-1}	x^n	$\frac{1}{n+1}x^{n+1}$
$-\frac{1}{4x\sqrt{x}}$	$\frac{1}{2\sqrt{x}}$	$\sqrt{x}$	$\frac{2}{3}\sqrt{x^3}$
$-\sin x$	$\cos x$	$\sin x$	$-\cos x$
$-\cos x$	$-\sin x$	$\cos x$	$\sin x$
e^x	e^x	e^x	e^x
$a^x(\ln a)^2$	$a^x \ln a$	a^x	$\frac{a^x}{\ln a}$
$-\frac{1}{x^2}$	$\frac{1}{x}$	$\ln x$	$x \ln x - x$
$-\frac{1}{x^2 \ln a}$	$\frac{1}{x \ln a}$	$\log_a x$	$\frac{1}{\ln a}(x \ln x - x)$
$\frac{2}{x^3}$	$-\frac{1}{x^2}$	$\frac{1}{x}$	$\ln \lvert x \rvert$
$\frac{3}{4\sqrt{x^5}}$	$-\frac{1}{2\sqrt{x^3}}$	$\frac{1}{\sqrt{x}}$	$2\sqrt{x}$

Nützlich ist oftmals das Umschreiben bestimmter Ausdrücke wie z.B.:

$$\sqrt[n]{x} = x^{\frac{1}{n}}$$

$$\frac{1}{x^m} = x^{-m}$$

$$\sqrt[n]{x^m} = x^{\frac{m}{n}}$$

$$\frac{1}{\sqrt[n]{x^m}} = x^{-\frac{m}{n}}$$

C.4.1 Beispiel: Integration von $\frac{3}{4\sqrt{x^5}}$

Wir wollen

$$f(x) = \frac{3}{4\sqrt{x^5}}$$

ableiten. Dafür schreiben wir um in:

$$f(x) = \frac{3}{4}x^{-\frac{5}{2}}$$

Nach $f(x) = x^n \rightarrow F(x) = \frac{1}{n+1}x^{n+1}$ erhalten wir:

$$F(x) = \frac{3}{4} \cdot \frac{1}{-\frac{5}{2}+1}x^{-\frac{5}{2}+1} = \frac{3}{4} \cdot \frac{1}{-\frac{3}{2}}x^{-\frac{3}{2}} = -\frac{3 \cdot 2}{4 \cdot 3}x^{-\frac{3}{2}}$$

Vereinfachen und Kürzen führt uns auf das Ergebnis

$$F(x) = -\frac{1}{2}\frac{1}{\sqrt{x^3}} = -\frac{1}{2\sqrt{x^3}}$$

C.4.2 Beispiel: Ableitung von $\frac{2}{3}\sqrt{x^3}$

Es gilt $f(x) = x^n \rightarrow f'(x) = nx^{n-1}$. Wir schreiben unsere Formel erst einmal in diese Form um:

$$f(x) = \frac{2}{3}\sqrt{x^3} = \frac{2}{3}x^{\frac{3}{2}}$$

und wenden nun die angegebene Forschrift an:

$$F(x) = \frac{2}{3} \cdot \frac{3}{2}x^{\frac{3}{2}-1} = \frac{2 \cdot 3}{3 \cdot 2}x^{\frac{1}{2}} = x^{\frac{1}{2}}$$

Schreiben wir nun noch um erhalten wir

$$F(x) = \sqrt{x}$$

als Ergebnis

Auf gebrochenrationale Funktionen wird an dieser Stelle nicht eingegangen, dort sind die Ableitungen nicht notwendigerweise schwerer, aber aufwendiger, da dort beispielsweise mit der Produkt-, Ketten- und/oder Quotientenregelregel gearbeitet werden muss. Ebenso bleibt die behandlung mehrdimensionaler Funktionen unbeleuchtet, auch wenn sie in der Statistik prinzipiell bedeutend sind (Beispielsweise in der Herleitung der Regression).

Anhang D

Herleitung der logistischen Regressionsgleichung

Der Einfachheit und Übersichtlichkeit halber verkürzen wir die Schreibweise von

$$\sum_{i=1}^{n} b_i x_i \quad \text{auf} \quad bx_i$$

also auf den bivariaten Fall und

$$p(x) \quad \text{auf} \quad p$$

Wenn wir eine Wahrscheinlichkeit durch lineare Regression vorhersagen wollen treffen wir auf Probleme: Die Wahrscheinlichkeit ist auf das Intervall von 0 bis 1 festgelegt. Sie können nicht negativ oder grösser 1 werden, so wie es die rechte Seite der Formel kann.

$$p = a + bx_i$$

Um dieses Problem zu lösen betrachtet man die Odds, also den Quotienten aus zwei Wahrscheinlichkeiten.

$$\frac{p}{1-p} = a + bx_i$$

Der Odd der Wahrscheinlichkeit zu "Überleben" für $p(x) = 0.75$ beträgt $\frac{p(x)}{1-(x)} = \frac{0.75}{0.25} = 3$. Also ist die Wahrscheinlichkeit zu überleben 3 mal höher als nicht zu überleben. Das ist schon besser. aber immer noch nicht OK, den die Odd-Ratios können nich negativ werden, also besitzen sie einen Wertebereich zwischen 0 und $+\infty$. Durch logarithmieren (üblicherweise mit dem

logarithmus naturalis ln) erreichen wir einen Wertebereich zwischen $-\infty$ und $+\infty$.

$$\ln \frac{p}{1-p} = a + bx_i$$

Wenn wir die Gleichung nun nach p auflösen wollen gehen wir folgendermaßen vor:

$$e^{\ln \frac{p}{1-p}} = e^{a+bx_i}$$

Da gilt $e^{\ln x} = \ln e^x = x$, es sich also um die Umkehrfunktion handelt gilt folgendes:

$$\frac{p}{1-p} = e^{a+bx_i}$$

Multiplikation mit $1 - p(x)$

$$p = e^{a+bx_i}(1-p)$$

Ausmultiplizieren

$$p = e^{a+bx_i} - pe^{a+bx_i}$$

Addition, um pe^{a+bx_i} auf die linke Seite zu bringen:

$$p + pe^{a+bx_i} = e^{a+bx_i}$$

Ausklammern von p

$$p\left(1 + e^{a+bx_i}\right) = e^{a+bx_i}$$

Dividieren durch $\left(1 + e^{a+bx_i}\right)$

$$p = \frac{e^{a+bx_i}}{1 + e^{a+bx_i}}$$

Hier ist in manchen Lehrbüchern Schluss, wir haben die Formel der logistischen Regression erreicht. Doch kann man noch weiter vereinfachen: Klammern wir unter dem Bruchstrich e^{a+bx_i} aus.

$$p = \frac{e^{a+bx_i}}{e^{a+bx_i}(\frac{1}{e^{a+bx_i}} + 1)}$$

Umschreiben, da gilt $\frac{1}{a} = a^{-1}$

$$p = \frac{e^{a+bx_i}}{e^{a+bx_i}(e^{-(a+bx_i)}+1)}$$

Finales Kürzen

$$p(x) = \frac{1}{e^{-(a+bx_i)}+1}$$

Literaturverzeichnis

Allison, Paul .D.: Event History Analysis - Regression for Longitudinal Event Data. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-046. Beverly Hills London und Neu Dehli: Sage Publications, 1984

Blossfeld, Hans-Peter / **Rohwer**, Götz: Techniques of Event History Modeling - New Approaches to Causal Analysis. Mahwah, New Jersey und London: Lawrence Erlbaum Associates, 2002

Diekmann, Andreas / **Mitter**, Peter: Methoden zur Analyse von Zeitverläufen - Anwendung stochastischer Prozesse bei der analyse von Ereignisdaten. Stuttgart: B. G. Teubner, 1984

Krempel, Lothar: Soziale Interaktionen: Einstellungen, Biographieb, Situationen und Beziehungsnetzwerke - Dynamische Ereignisanalyse. Sozialwissenschaften, Band 2. Bochum: Schallwig, 1987

Vermunt, Jeroen K.: Log-Linear Models for Event Histories. Advanced techniques in the social sciences, Vol. 8. Thousand Oakes, London und Neu Dehli: Sage Publications, 1997

Wagenpfeil, Stefan: Dynamische Modelle zur Ereignisanalyse. München: Herbert Utz Verlag, 1996

Yamaguchi, Kazuo : Event History Analysis. Applied Social Research Method Series, Vol. 28. Newbury Park, London und Neu Dehli: Sage Publications, 1991

Internet

Arias, Elizabeth: United States Life Tables, 2003. National Vital Statistics Report - Volume 54, Number 14

`http://www.cdc.gov/nchs/data/nvsr/nvsr54/nvsr54_14.pdf`

Braun, Norman / **Engelhardt**, Henriette: Difussion Processes and Event History Analysis. Max-Planck-Institut für demographische Forschung: MPIDR Woking Paper WP 2002-007

`http://www.demogr.mpg.de/papers/working/wp-2002-007.pdf`

Golsch, Katrin: Ereignisanalyse in Stata 9.

`http://eswf.uni-koeln.de/mitarbeiter/golsch/ereignisanalyse.pdf`

Ludwig-Mayerhofer, Wolfgang: ILMES - Internet-Lexikon der Methoden der empirischen Sozialforschung

`http://www.lrz-muenchen.de/~wlm/ilmes.htm`

Pötter, Ulrich / **Rohwer**, Götz: Introduction to Event History Analysis.

`http://www.stat.ruhr-uni-bochum.de/pub/eha/eha_txt.ps`

Steele, Fiona: Event History Analysis. ESRC National Centre for Research Methods NCRM Methods Review Papers, NCRM/004

`http://www.ncrm.ac.uk/publications/methodsreview/MethodsReviewPaperNCRM-004.pdf`

Vermunt, Jeroen K. / **Moors**, Guy: Event History Analysis. Department of Methodology and Statistics, Tilburg University.

`http://spitswww.uvt.nl/~vermunt/esbs2005b.pdf`

Vermunt, Jeroen K.: Log-linear event history analysis: a general approach with missing data, latent variables, and unobserved heterogeneity.

`http://spitswww.uvt.nl/~vermunt/thesis.pdf`

Wu, Lawrence L.: Event History Models for Life Course Analysis. CDE Working Paper No. 2001-17

`http://www.ssc.wisc.edu/cde/cdewp/2001-17.pdf`

Ziegler, Andreas et al.: Überlebenszeitanalyse: Eigenschaften und Kaplan-Meier Methode - Artikel Nr. 15 der Statistik-Serie in der DMW.

`http://www.thieme-connect.com/ejournals/pdf/dmw/doi/10.1055/s-2002-32819.pdf`

Abbildungsverzeichnis

2.1 f(x) & F(x) . . . 11
2.2 Eintrittswahrscheinlichkeit Ereignis . . . 12
2.3 Flächeninhalt unter Kurve = Integral: survival function & distribution function . . . 13
2.4 Teilweise linkszensiert . . . 15
2.5 Rechtszensierung . . . 16
2.6 Keine Zensierung . . . 16

3.1 Einteilung in diskrete Intervalle . . . 22
3.2 Vergleich mehrerer Gruppen . . . 24
3.3 Beispiel: Kaplan-Meier Kurve . . . 26

4.1 Dichte- & Survivor-Funktion im Exponential Hazard Rate Modell . . . 32
4.2 Survivorfunktion (variabel) und Hazardrate (konstant) . . . 32
4.3 Beispiel einer Piecewise-Funktion . . . 35
4.4 Hazardrate im Weibull-Modell . . . 36
4.5 Hazardrate im Gompertz-Makeham-Modell . . . 37
4.6 Hazardrate im Log-Logistischen Modell . . . 39
4.7 hazard rate im Sickle-Modell . . . 43

A.1 Wahrscheinlichkeit für Gewinn . . . 50
A.2 Kumulierte Wahrscheinlichkeit . . . 51
A.3 Wahrscheinlichkeit & kumulierte Wahrscheinlichkeit . . . 51
A.4 Abnehmender Abstand zwischen den Messungen . . . 53

B.1 Dichtefunktion und Verteilungsfunktion der Standardnormalverteilung . . . 54

C.1 Ausgangsfunktion der Standardnormalverteilung . . . 56
C.2 Ausgangsfunktion der Standardnormalverteilung . . . 57
C.3 Erste Ableitung der Standardnormalverteilung . . . 59

C.4 Erste Ableitung und Stammfunktion von $f(x) = x^2$ 60