Cox 1972, Regression Models and Life- Tables

Bachelor-Seminar an der
Ludwig-Maximilian Universität,
Sommersemester 2014,

Betreuerin: Andrea Wiencierz

Lorenz Winklmaier, Matrikelnummer: 10618481

30. Mai 2014

Inhaltsverzeichnis

1 Einleitung

Sterbetafeln gehören zu den ältesten Techniken in der Statistik, doch deren Gebrauch ist immer noch weit verbreitet. In diesen demographischen Modellen wird die Abgangsordnung eines Bevölkerungsstandes in Form einer Tabelle dargestellt. Durch Ausfälle wie z.B. den Tod eines Teils dieser Bevölkerung, wird jener Bevölkerungsstand reduziert und so können Ausfallwahrscheinlichkeiten geschätzt werden (vgl. Augustin, T. 2013). Sir David Cox hatte einen sehr engen Bezug auf jenes Thema, da er, ausgebildet mit einem Magisterabschluss und einem Doktortitel, an dem Royal Aircraft Establishment und an der Wool Industries Research Association tätig war, wo Ausfallsicherheit und das Testen von bestimmten Lebensdauern zu den wichtigsten Aufgaben gehörten. Später arbeitete er zudem noch am statistischen Labor in Cambridge und als Professor der Statistik an dem Birbeck College, dem Imperial College, und der Universität von North Carolina. So dehnte er auch die Arbeiten von Kaplan und Meier über diese Form von Statistik weiter aus, indem er Elemente der Regression in die Analyse von Sterbetafeln einbaute, was zu einer bahnbrechenden Methode für zensierte Daten mit Ausfällen wurde.

2 Cox' Ideen zur Modellierung von Hazardraten

David Cox wollte mit einer semiparametrischen Regressionsfunktion in exponentieller Form, Einflüsse von bestimmten Kovariablen auf das Verhältnis von Ausfallraten zwischen Forschungsobjekten erklären. In solchen Modellen gibt es oft das Problem, dass ein Teil der Daten rechtszensiert vorliegt. Dies bedeutet, dass Individuen betrachtet werden, sowie deren Ausfallzeiten.
Kommt es nun vor, dass eine Untersuchungseinheit während der ganzen Studie nicht ausfällt, so nennt man dies eine rechtszensierte Beobachtung, denn man weiß nicht wann dieses Objekt ausfallen wird, sondern nur, dass die Ausfallzeit die beobachtete Zeit überschreitet (vgl. Stein,P. 2007). Dies hat zur Folge, dass man keine Informationen über Regressionsparameter erhält. Cox versuchte dem entgegenzuwirken, indem er Ausfallraten mit einer Hazardfunktion darstellt und diese zu bestimmten Zeitpunkten miteinander vergleicht.
Außerdem lieferte David Cox einen sehr leichten Weg zur Schätzung der Regressionsparameter, da er sich auf bedingte Ausfallwahrscheinlichkeiten konzentrierte. Seine Arbeit basiert auf folgendem Regressionsmodell:
$\lambda(t; z) = \lambda_0(t) \exp(z\beta)$,
wobei $\lambda_0(t)$ die Baseline-Hazard Funktion genannt wird. Sie ist eine beliebige und unbekannte Ausfallfunktion, für den Fall dass alle Kovariablen 0 sind, also wenn es keine Einflüsse auf die Ausfallrate gibt. Zudem richtet sich die Baseline-Hazard Funktion schließlich noch nach der Zeit, die während der Studie vergeht.
z ist ein Vektor, der mehrere Kovariablen $(z_1, ..., z_p)$ enthält;
β ist der Vektor der zu schätzenden Parameter. Anhand dessen können Zusammenhänge der Einflüsse in z mit der Ausfallwahrscheinlichkeit ermittelt werden.
Der Term $\exp(z\beta)$ gibt das relative Ausfallrisiko für ein Individuum wieder. Es ist das Risiko verglichen mit dem Risiko, wenn keine Einflüsse vorhanden sind.
Die Idee des David Cox war es nun, frühere Arbeiten an zum Beispiel Weibull-Regressionsmodellen zu verallgemeinern. Die Abhängigkeit der Gefahr eines Ausfalls von der vergangenen Zeit sollte nicht fehlen. So definierte der Autor eine bedingte

Wahrscheinlichkeit, bei welcher der Vektor z durch $z(t)$ ersetzt wird. Für ihn war die gesuchte Wahrscheinlichkeit der Quotient aus dem relativen Risiko für ein Individuum, das zum Zeitpunkt t_i ausfällt und der Summe der relativen Risiken aller Individuen, die zu t_i noch nicht ausgefallen sind. Hier geht man davon aus, dass die alle Ausfallzeiten voneinander verschieden sind:
$t_1 < .. < t_k$;
So ist die Abhängigkeit von der Zeit im Modell, was nötig ist, da Kovariablen zu unterschiedlichen Zeitpunkten unterschiedliche Effekte haben können. Die Likelihood-Funktion sieht wie folgt aus:
$L(\beta) = \prod_{i=1}^{k} [\frac{\exp(z_i(t_i)\beta)}{\sum_{l \in R(t_i)} \exp(z_l(t_i)\beta)}]$,
wobei $z_i(t_i)$ der Kovariablenvektor zur Zeit t_i ist für das i-te Individuum, welches auch zum Zeitpunkt t_i ausfällt.
$R(t_i)$ ist die Menge der Untersuchungseinheiten, die zu t_i noch gefährdet sind, auszufallen, also jene die noch nicht zuvor bereits ausgefallen sind. Auf diese bedingte Wahrscheinlichkeit wird später noch genauer eingegangen.

2.1 Definitionen zu Ausfallraten außerhalb der Regression

2.1.1 Die Hazardrate und Überlebensfunktion

Das Ziel der Cox-Regression ist es, die Einflüsse von Kovariablen auf die Hazardrate zu untersuchen. Stellt man sich jedoch die Frage, wie diese zeitbezogene Ausfallrate genau definiert ist, greift man nicht auf Elemente der Regression zurück, sondern stellt die Rate mithilfe von Wahrscheinlichkeiten dar. Dazu benötigt man die Variable T, welche die Ausfallzeit repräsentiert. Nun kann bereits eine Überlebensfunktion mithilfe dessen aufgestellt werden:
$F(t) = P(T \geq t)$;
Diese Funktion richtet sich nach t, und gibt die Wahrscheinlichkeit wieder, dass die Ausfallzeit T größer oder gleich t ist.
Die Hazardfunktion dagegen gibt die Wahrscheinlichkeit an, dass es zu einem bestimmten Zeitpunkt zu einem Ausfall kommt. Man kann Sterbewahrscheinlichkeiten pro Zeiteinheit angeben. Lässt man nun diese Einheit gegen 0 gehen, so ergibt sich die Wahrscheinlichkeit zu einem Zeitpunkt:
$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t+\Delta t | t \leq T)}{\Delta t}$
Im Zähler steht sozusagen dass die Ausfallzeit gleich t sein soll ($T = t$), es wird

jedoch mit dem Grenzwert für Δt geschrieben, um auf einen Zeitpunkt statt ein Intervall hinzuweisen. Die Bedingung in der Wahrscheinlichkeit bedeutet, dass der Ausfall nicht schon vor t stattgefungen hat.
Im diskreten Fall kann man auf die Bezeichnung mit Grenzwert verzichten, da es nicht durchgehend zu jedem Zeitpunkt zu einer Reduzierung der Untersuchungseinheiten kommen kann, sondern nur in bestimmten diskreten Momenten.
$\lambda(t) = \sum \lambda_{u_j} \delta(t - u_j)$,
mit $\lambda_t = P(T = t|T \geq t)$ und $\delta(x)$, wobei hier das Dirac-Maß gemeint ist. Es hat den Wert 1, falls $x = 0$ bzw. in diesem Fall, wenn $u_j = t$. Ansonsten hat das Maß den Wert 0 (vgl. Schneider,C. 2009). Die Summe geht über alle u_j und dieser Index soll einfach die Zeitspanne der Beobachtung darstellen. Die Glieder der Summe sind gleich 0, außer das Summenglied, welches die wahre Ausfallzeit beinhaltet. An dieser Stelle benötigt man die Sterbewahrscheinlichkeit λ_t mit selber Bedingung wie im diskreten.
Durch diese Hazardfunktionen kann man wiederum die Überlebensfunktion aufstellen. Da $\lambda(t)$ die Ausfallrate wiedergibt, steht $1-\lambda(t)$ für die Überlebenswahrscheinlichkeit zu t. Integriert man jetzt von 0 bis zur Zeit kurz vor t, so ergibt sich auch die Wahrscheinlichkeit, nicht auszufallen für diese gesamte Zeitspanne:
$F(t) = \int_{u=0}^{t-0} (1 - \lambda(u)\, du)$
Das klassische Integral kann man als Flächeninhalt unter der Kurve der Funktion ansehen, den man berechnet als Summe von unendlich vielen (x-Achse wird in minimal kleine Abschnitte unterteilt) Stufen einer Treppenfunktion. Diese einzelnen Flächeninhalte erhält man, indem man den Funktionswert mit der zugehörigen Länge des x-Achsenabschnitts multipliziert und anschließend aufsummiert. Bei der Hazardrate geht es jedoch um Wahrscheinlichkeiten. Sollen 2 Ereignisse gleichzeitig passieren, so multipliziert man die Wahrscheinlichkeiten. Deshalb braucht man für das Integral auch keine Summe, sondern das Produkt der einzelnen Stufen.
$F(t) = \lim \prod_{k=0}^{r-1} (1 - \lambda(\tau_k))(\tau_k + 1 - \tau_k)$,
wobei die x-Achse unterteilt ist durch: $0 = \tau_0 < \tau_1 < ... < \tau_r = t$; Eine Formel über das Produktintegral ist wie folgt definiert:
$\prod_0^t (1 + f(u)du) = \exp(\int_0^t f(u)du)$
(vgl. Slavik,A. 2007), wodurch sich im stetigen Fall folgende Überlebensfunktion ergibt:
$F(t) = \exp(-\int_0^t \lambda(u)du)$
Für den diskreten Fall ergibt sich ein anderes Produktintegral. Da nicht durchgehend, sondern nur zu diskreten Zeiten Ereignisse betrachtet werden, müssen nur in diesen Momenten die Überlebenswahrscheinlichkeiten multipliziert werden:

$F(t) = \prod_{u_j<t}(1 - \lambda_(u_j))$;

2.1.2 Die Kaplan und Meier-Schätzer

Eben wurde behandelt, wie die Hazardrate und Überlebensfunktion genau definiert ist. Kaplan und Meier lieferten zusätzlich noch einen Weg zur Schätzung dieser Komponenten. Dazu definierten sie die Variablen m_i, was für die Zahl der Ausfälle zum Zeitpunkt t_i steht, sowie r_i. Dies gibt die Anzahl der Individuen in der Menge $R(t_i)$. So ergibt sich der Schätzer für die Ausfallrate
$\lambda(t) = \sum_{i=1}^{k} \frac{m_i}{r_i}\delta(t - t_i)$;
Für die Überlebensfunktion geht man vor, wie zuvor und erhält den Schätzer
$F(t) = \int_{u=0}^{t-0}(1 - \lambda(u)du) = \prod_{t_i<t}(1 - \frac{m_i}{r_i})$.

2.2 Regressionsmodelle und Analysemöglichkeiten

Ein Problem bei der Analyse der Cox-Regressionsmodelle stellt die unbekannte Baseline-Hazardfunktion im Prädiktor dar. Falls diese konstant wäre, würde dies die Analyse deutlich einfacher gestalten, was jedoch nie vorkommt. Der Baseline-Hazard ist eine willkürlich gewählte Funktion. Diese Eigenschaft von $\lambda_0(t)$ hat keine schwerwiegenden Auswirkungen auf den Informationsverlust über den zu schätzenden Parameter β, deswegen scheint eine Annäherung an β über Maximum-Likelihoodmethoden als gerechtfertigt. Das eigentliche Problem tritt bei der Analyse der Inferenz auf, weil Informationen über den Parameter erlangt werden sollen unter verschiedenen Annahmen über den Baseline-Hazard, der ebenfalls eine Funktion der Zeit ist. Die Abhängigkeit der Ausfallwahrscheinlichkeit vom Kovariablenvektor z muss nun ausgedrückt werden.

2.2.1 Regressionsbeispiele

In den folgenden Beispielen wird davon ausgegangen, dass keine Bindungen in den Daten vorliegen, das heißt jedes Individuum hat eine eigene Ausfallzeit. Die Basis stellt wiederum folgendes Modell dar:
$\lambda(t; z) = \exp(z\beta)\lambda_0(t)$;
Als erstes Beispiel wird der Zwei-Stichprobenfall betrachtet mit nur einer Kovariable in z. Es ist eine Indikatorvariable für die zwei Stichproben, kann somit die Werte 1 (bei Stichprobe 1) und 0 (bei Stichprobe 2) annehmen. Durch Einsetzen der Werte von z in die Modellgleichung ergibt sich für die Hazardrate bei der 1. Stichprobe

$\lambda(t;z) = \lambda_0(t)\exp(\beta)$
und in Stichprobe 2
$\lambda(t;z) = \lambda_0(t)$.
Im nächsten Beispiel werden erneut 2 Stichproben behandelt und die Variable z_1 bleibt gleich wie jene in der ersten Beispielregression. Zusätzlich wird noch eine zeitabhängige Variable mit in das Modell eingeführt: $z_2 = tz_1$; So erhält man für den Fall $z_1 = 1$ folgende Hazardfunktion:
$\lambda(t;z) = \lambda_0(t)\exp(\beta_1+\beta_2 t)$; Wird β nun geschätzt, so ergeben sich bereits Ergebnisse für die Abhängigkeit der Ausfallrate von den genannten Kovariablen. Also kann man die beiden Stichproben vergleichen bezogen auf die Ausfallrate und noch weitere Variablen in das Modell aufnehmen.

2.2.2 Analyse durch bedingte Wahrscheinlichkeiten im Stetigen

Wie vorher gesagt wurde, stellt die zeitabhängige Baseline-Hazardfunktion ein Problem für die Schätzung der Regressionsparameter dar. David Cox fand einen Weg zur Lösung durch bedingte Wahrscheinlichkeiten. Zunächst wird sich auf den stetigen Fall konzentriert. Zeitintervalle spielen keine Rolle, sondern die Momente t_i, in denen es zu Ausfällen von Individuen kommt. Diese sind für jedes Individuum unterschiedlich. Für Zeitpunkte ohne Ausfälle ist es nicht möglich, Informationen über β zu gewinnen, da $\lambda_0(t)$ gleich 0 sein wird und somit auch die gesamte Regressionsgleichung. Cox kam die Idee, diese Hindernisse zu umgehen durch Aufstellen der Wahrscheinlichkeit, dass von allen zu t_i noch lebenden Individuen ausgerechnet Individuum i stirbt. Er nannte dies ein bedingtes Risiko, da es die Ausfallwahrscheinlichkeit war, bedingt auf die Gesamtheit der Gefährdeten zu t_i $(= R(t_i))$. Sozusagen teilt man die Ausfallrate für Individuum i durch die Summe der restlichen Ausfallraten, wodurch die störende Baseline-Hazardfunktion der Schätzung nicht mehr im Weg steht:

$$\frac{\exp(z_i\beta)}{\sum\limits_{l\in R(t_i)} \exp(z_l\beta)}$$

Anhand dessen kann nun die Likelihood-Funktion aufgestellt werden:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(z_i\beta)}{\sum\limits_{l\in R(t_i)} \exp(z_l\beta)} = \frac{\exp(\sum z_i\beta)}{\prod \sum\limits_{l\in R(t_i)} z_l\beta} \;;$$

Durch Logarithmieren ergibt sich die log-likelihood:

$$l(\beta) = \sum_{i=1}^{k}(z_i\beta) - \sum_{i=1}^{k} \log[\sum_{l\in R(t_i)} \exp(z_l\beta)];$$

Als nächsten Schritt der Maximum-Likelihood Schätzung muss man die Score-Funktion

aufstellen. Diese ergibt sich durch Ableiten der log-likelihood nach einem der β-Parameter.

$S_\xi(\beta) = \frac{\partial l(\beta)}{\partial \beta_\xi} = \sum_{i=1}^{k}(z_{\xi_i} - A_{\xi i}(\beta))$, wobei $A_{\xi i}(\beta) = \frac{\sum\limits_{l \in R(t_i)} z_{\xi l} \exp(z_l \beta)}{\sum \exp(z_l \beta)}$

β ist ein Vektor der Parameter, und ξ steht hier stellvertretend für den Index einer Komponente dieses Vektors, nach welcher man ableitet.

Setzt man die Funktion nun gleich 0, so erhält man eine Score-Gleichung

$S(\beta_{ML}) = 0$, und die Lösung dieser Gleichung ist der gesuchte Schätzer für β (vgl. Küchenhoff,H. 2013). $A_{\xi_i}(\beta)$ ist ein exponentiell gewichteter Durchschnitt der Variable z_ξ für die begrenzte Population $R(t_i)$. Die Eigenschaft eines Durchschnitts wird deutlich wenn man annimmt β sei gleich 0. Das Einsetzen liefert das arithmetische Mittel, sozusagen eine spezielle Form der Gewichtung:

$A_{\xi i}(0) = A_{\xi i} = \frac{\sum z_{\xi_l}}{r(t_i)}$

Der nächste Schritt behandelt die Fisher-Information. Das ist der negative Wert der 2. Ableitung der log-likelihood. Dies asymptotische Varianzmatrix ergibt sich als Inverse der Fisher-Information (vgl. Küchenhoff,H. 2013). Dabei wird nach 2 unterschiedlichen Komponenten vom Vektor β abgeleitet, nämlich ξ und η.

$I_\xi \eta = -\frac{\partial^2 l(\beta)}{\partial \beta_\xi \partial \beta_\eta} = \sum_{i=1}^{k} C_{\xi\eta i}(\beta)$ und die Kovarianz im gewichteten Fall von z_ξ und z_η ist:

$C_{\xi\eta i}(\beta) = [\sum z_{\xi l} z_{\eta l} \exp(z_l \beta) / \sum(\exp(z_l \beta))] - A_{\xi i}(\beta) A_{\eta i}(\beta)$;

Unter anderem konnte Cox zeigen, dass die Beiträge der unterschiedlichen Ausfallzeiten zur Score-Funktion unkorreliert sind, sowie dass $F(\beta)$ ein Schätzer für die Varianz von $S(\beta)$ ist.

Zusätzlich können natürlich auch Signifikanztests über die Parameter oder Teilmengen davon durchgeführt werden. Ein einfaches Beispiel hierfür wäre das Testen der globalen Nullhypothese, das heißt $H_0 : \beta = 0$. Ein Score-Test wird in diesem Fall benötigt, welcher überprüft ob die Kovariablen einen Einfluss auf die Hazardrate haben. Die zugehörige Teststatistik

$S(0)^T I(0)^{-1} S(0)$

ist asymptotisch χ^2-verteilt mit p Freiheitsgraden (vgl. Kauermann,G. 2013) und

$S_\xi(0) = \sum_{i=1}^{k}(z_{\xi i} - A_{\xi i})$, sowie

$I_{\xi\eta}(0) = \sum_{i=1}^{k} C_{\xi\eta i}$, wobei $C_{\xi\eta i} = C_{\xi\eta i}(0)$; Es ist wiederum die Kovarianz von z_ξ und z_η, nur in diesem Fall in nicht-gewichteter Form, da $\beta = 0$ getestet wird.

2.2.3 Analyse in diskreter Zeit

Der Unterschied der Cox-Regression im Stetigen zum Diskreten ist, dass in Letzterem die Möglichkeit besteht, dass Bindungen innerhalb der Daten vorliegen. Das bedeutet, dass mehrere Individuen zu einem Zeitpunkt ausfallen können, was im Stetigen unmöglich ist. Im diskreten Fall gilt also: $m_i \geq 1$. Die Analyse und Schätzung in der Regression verläuft analog zu jener im Stetigen, nur dass diese Vielzahl der Todesfälle berücksichtigt werden muss. Deshalb wird die Formel für die bedingte Wahrscheinlichkeit umgeändert zu

$$\frac{\exp(s_i\beta)}{\sum\limits_{l \in R(t_i, m_i)} \exp(s_l\beta)} \; ;$$

z_i wurde hier durch s_i ersetzt. Das ist die Summe bezogen auf z der Individuen, die zum Zeitpunkt t_i ausfallen. Im Nenner der Formel ist die Summe über allen Individuen gemeint, welche zur Zeit t_i noch leben oder sterben. Mit der gleichen Vorgehensweise wie bei 2.2.2 ergeben sich die log-likelihood

$$l(\beta) = \sum_{i=1}^{k} (s_i\beta) - \sum_{i=1}^{k} \log[\sum_{l \in R(t_i, m_i)} \exp(s_l\beta)],$$

sowie die Score-Funktion und Fisher-Information für den Fall $\beta = 0$, falls wiederum die globale Nullhypothese getestet werden soll:

$$S_\xi(0) = \sum_{i=1}^{k} (s_{\xi i} - m_i A_{\xi i}) \text{ und}$$

$$I_{\xi\eta}(0) = \sum_{i=1}^{k} \frac{m_i(r_i - m_i)}{r_i - 1} C_{\xi\eta i};$$

Diese Formel können auch im stetigen Fall verwendet werden, weil $s_i = z_i$ und $m_i = 1$ somit gilt und man identische Formeln erhält.

2.3 Der 2-Stichproben-Fall als Illustrationsbeispiel

Hypothesentests kann man schließlich auch für den Fall, dass zwei Stichproben miteinander verglichen werden, aufstellen. Die Nullhypothese lautet:

$H_0 : \lambda_1(t; z) = \lambda_2(t; z) = \lambda_0(t)$;

Wiederum testet man sozusagen, ob $\beta = 0$, denn dadurch erhält der Faktor $\exp(z\beta)$ in beiden Stichproben den Wert 1 und die zwei Hazardraten sind gleich.

Um dies zu veranschaulichen, wird nun das erste Beispiel aus 2.2.1 verwendet, wo es nur eine Kovariable gibt, welche als Indikatorvariable die Zugehörigkeit zu einer Stichprobe angibt. Wenn $z = 1$, so ist das Individuum Teil der ersten Stichprobe, bei $z = 0$ der zweiten Stichprobe. Zum Überprüfen wird erneut ein Score-Test benötigt. Da dieser nun zwei Stichproben vergleichen soll, ändert sich die Teststatistik von 2.2.2 zu

$S(0)/\sqrt{I(0)}$;
Außerdem wird die Nullhypothese anhand einer Standardnormalverteilung betrachtet. Diesen Test kann man sich vorstellen, als würde man bei jedem Zeitpunkt eines Ausfalls eine Kontingenztabelle aufstellen mit den Variablen Sample (Sample 1 und Sample 2 sind die Ausprägungen) und Status (mit ausgefallen und überlebt). Durch das Kombinieren der Informationen der einzelnen Tabellen kann so überprüft werden, ob ein Unterschied zwischen den Stichproben besteht.
Setzt man die entsprechenden Werte von z in die für die Teststatistik benötigten Terme ein, ergibt sich für dieses Beispiel
$S_\xi(0) = n_1 - \sum_{i=1}^{k} m_i A_i$;
n_1 gibt die Zahl der Ausfälle in der ersten Stichprobe für alle Zeitpunkte an, da $n_1 = \sum_{i=1}^{k} s_i$.
Für die Fisher-Information kann C nun, da man nur eine z-Variable hat, wie folgt geschrieben werden: $C_i(\beta) = \sum z_l \exp(z_l\beta)/\sum \exp(z_l\beta) - A_i(\beta)A_i(\beta) = A_i(1 - A_i)$;
Insgesamt erhält man für die Fisher-Information
$I(0) = \sum_{i=1}^{k} \frac{m_i(r_i - m_i)}{r_i - 1} A_i(1 - A_i)$;
Dadurch kann für dieses Beispiel die Nullhypothese getestet werden.

2.4 Die Cox-Regression mit R

Die Cox-Regression kann schließlich auch mit der Programmiersprache R durchgeführt werden. Dazu gibt es das package survival, wovon man für die Überlebenszeitanalyse passende Datensätze einlesen kann. Im Folgenden wird der Datensatz cancer betrachtet. Hier sind Krebspatienten aufgelistet. Für die folgende Regression ist die Variable time wichtig, welche die Überlebenszeit in Tagen angibt. Status gibt an, ob die Ausfallzeit des Individuum zensiert ist, das heißt ob es die Beobachtungszeit überlebt hat oder nicht. Der Wert 1 bedeutet zensiert, 2 bedeutet gestorben.
Zudem werden noch die Variablen age (in Jahren) und sex betrachtet. Hier bedeutet die Ausprägung 1 männlich und 2 weiblich.
Für die Cox-Regression in R muss man sich ein Survivalobjekt erstellen, welches die Zielvariable repräsentiert. Dies geschieht in R durch die Surv()-Funktion. Es ist eine Funktion von der überlebten Zeit und vom Zensurstatus. Diese Objekte werden an den Datensatz angehängt und von R folgendermaßen dargestellt:

```
> head(data$SurvObj)
[1]  306   455  1010+  210   883  1022+
```

Ist ein + an der Ausprägung, so bedeutet dies, dass das Individuum nicht gestorben ist, sondern dass dessen Ausfallzeit zensiert ist. Mithilfe der Überlebenszeit und des Zensurstatus stellt diese Funktion die Hazardrate als Zielvariable dar.
Die entsprechende Regression erfolgt mit der coxph()-Funktion. Wie gesagt, wird das Survival-Objekt durch die erklärenden Variablen age und sex beeinflusst (vgl. Fox,J. 2002). Der zugehörige R-Output berechnet die Schätzer für die Parameter anschließend:

```
> cox1 <- coxph(formula = SurvObj ~ age + sex , data = data)
> summary(cox1)
Call:
coxph(formula = SurvObj ~ age + sex, data = data)

  n= 228, number of events= 165

        coef exp(coef)  se(coef)      z Pr(>|z|)
age  0.017045  1.017191  0.009223  1.848  0.06459 .
sex -0.513219  0.598566  0.167458 -3.065  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    exp(coef) exp(-coef) lower .95 upper .95
age    1.0172     0.9831    0.9990    1.0357
sex    0.5986     1.6707    0.4311    0.8311

Concordance= 0.603  (se = 0.026 )
Rsquare= 0.06   (max possible= 0.999 )
Likelihood ratio test= 14.12  on 2 df,   p=0.0008574
Wald test            = 13.47  on 2 df,   p=0.001187
Score (logrank) test = 13.72  on 2 df,   p=0.001048
```

Für das Alter zeigt sich ein leicht signifikanter Zusammenhang und bei dem Geschlecht ein sehr signifikantes Ergebnis. Die Koeffizienten werden durch die Spalte $\exp(coef)$ interpretiert.
Geht man davon aus, dass das Geschlecht und t im Folgenden konstant bleibt und der Hazard verglichen wird für ein bestimmtes Alter und ein um eins höheres Alter, so ergibt sich:

$$\frac{\lambda(t; z_1 + 1)}{\lambda(t; z_1)} = \frac{\lambda_0(t)\exp(z\beta)\exp(\beta_1)}{\lambda_0(t)\exp(z\beta)} = \exp(\beta_1);$$
Das bedeutet, dass wenn das Alter (z_1) um 1 Jahr höher ist, so ist die Ausfallwahrscheinlichkeit um den Faktor $\exp(\beta_1)$ ($= 1.0172$) auch höher.
Analog wird bei der Interpretation der zweiten Kovariable vorgegangen:
$$\frac{\lambda(t; z_2 + 1)}{\lambda(t; z_2)} = \frac{\lambda_0(t)\exp(z\beta)\exp(\beta_2)}{\lambda_0(t)\exp(z\beta)} = \exp(\beta_2);$$
In diesem Fall kann man aussagen, dass die Hazardrate bei Frauen um den Faktor $\exp(\beta_2)$ ($= 0.5986$) niedriger ist, als bei Männern.

2.5 Physikalische Interpretation

Das besprochene Regressionsmodell ist vorhergesehen als Repräsentation des Verhaltens von Ausfallzeiten. Dabei kamen Diskussionen über dessen physikalische Bedeutung auf anhand der Variable s. Sie gibt eine bestimmte Art von Belastung, wie zum Beispiel Druck oder Temperatur, in einer Überlebenszeitstudie an und eine Standardbelastung entspricht $s = 1$. Betrachtet wird, wie der physikalische Prozess des Ausfallens an den unterschiedlichen Levels von s ausfällt.
Ein einfaches Modell dazu ergibt sich, falls man davon ausgeht, dass der Mechanismus des Sterbens zu den unterschiedlichen Ausprägungen von s identisch ist, jedoch die abgelaufene Zeit von der Variable abhängig ist. Hierfür multipliziert man eine Funktion $g()$ der Belastung mit der Zeit und baut dies in Überlebensfunktion ein: $F(t, s) = F(g(s)t, 1)$, wobei $g(1) = 1$ gelten muss.
Der dazugehörige Hazard zu einem Zeitpunkt t und während einer Belastung s lautet:
$g(s)\lambda_0(g(s)t)$.
Die Baseline-Hazardfunktion ist hier die Ausfallrate wenn $s = 1$.
Definiert man nun $g(s) = s^\beta$ und $z = log(s)$, so erhält man eine ähnliche Hazardfunktion wie zuvor:
$\exp(z\beta)\lambda_0(\exp(z\beta)t)$, da $e^z = s$ wegen $z = log(s)$ und $g(e^z) = \exp(z\beta)$.
Dieses Modell wird bei einem kumulativen Anstieg der Belastung verwendet, zum Beispiel wenn s mit der Zeit immer größer wird.
Bei physikalischen Prozessen, in denen es zu keiner Anhäufung der Belastung kommt, sondern wo eine bestimmte Reizschwelle durch diese Belastung zunächst überwunden werden muss, gibt es Änderungen. Man muss berücksichtigen, dass in diesem Fall die Ausfallraten abhängig von s sind. Dies erfolgt beispielsweise wenn sich Ausfälle auf die Überschreitung von Aktivierungsenergie(die Energie, die man aufwenden muss, dass eine Reaktion stattfindet) beziehen durch Temperaturbelastungen. So

wird das vorherige Modell gebraucht, wobei der Wert der Baseline-Hazardfunktion 1 beträgt:
Unterschiedlich zu den ersten beiden Fällen wäre ein Modell, in welchem der Prozess des Alterns unabhängig, aber die Rate der Anstiege der Belastung abhängig von s ist. Zusätzlich ist die bedingte Wahrscheinlichkeit das Produkt von einem zeitabhängigen Term $\lambda_0(t)$, der sich auf den Vorgang des Alterns bezieht, sowie einem Term $h(s)$, der von einer nicht kumulativen Belastung abhängt. Somit ist die Ausfallwahrscheinlichkeit
$h(s)\lambda_0(t)$.
Gilt wieder, dass $h(s) = s^\beta$ und $s = e^z$, so ergibt sich das Ausgangsmodell der Cox-Regression:
$\exp(z\beta)\lambda_0(t)$.

2.6 Das Vorgehen im bivariaten Fall

David Cox untersuchte ebenfalls Sterbetafeln mit bivariaten Daten. Dies kann man sich vorstellen, als gäbe es für jedes Individuum zwei Sorten von Ausfällen, die auch zu verschiedenen Zeiten stattfinden können. Es geht sozusagen um die Ausfallzeiten zweier unterschiedlicher, aber zusammengehöriger Bestandteile einer Untersuchungseinheit. Dabei ist die Zensur bei keinem, einem oder beiden Bestandteilen möglich. Die gemeinsame Hazardfunktion wird dargestellt durch die Wahrscheinlichkeit, dass ein Bestandteil zum Zeitpunkt t ausfällt, bedingt darauf, dass die andere Komponente (nicht) schon vor t ausgefallen ist. Durch die Kombination der passenden Ausfallwahrscheinlichkeiten zu bestimmten Zeiten und Überlebenswahrscheinlichkeiten für bestimmte Intervalle kann eine bivariate Wahrscheinlichkeitsverteilungs-Dichtefunktion für die benötigte Zeitspanne aufgestellt werden.

3 Der Einfluss von Cox' wegweisenden Ideen

Sir David Cox hat es geschafft, durch seine Arbeit und Ideen, die zu dieser Zeit vorherrschenden Erkenntnisse und Methoden über Sterbetafeln deutlich auszubauen. Da er die fehlende Abhängigkeit des Risikos von der Zeit kritisierte und zeitabhängige Terme $z_i(t_i)$ verwendete, konnte er die Regression für Ausfallstatistiken ermöglichen, da solche Terme die Flexibilität der Regression erhöhen. Zusätzlich erleichterte der Autor die damit verbundene Parameterschätzung und Analyse, indem er bedingte Wahrscheinlichkeiten nutzte für die Aufstellung der Score-Funktionen. Dabei wurde der Begriff "bedingte Wahrscheinlichkeit"viel kritisiert und diskutiert. Autoren beklagten, dass die daraus resultierende Likelihood-Funktion keine sei, die sich auf eine bedingte Verteilung bezieht, weswegen es nicht zu gewöhnlichen Ergebnissen bei der Schätzeffizienz kommt. Cox klärte dies mit seiner nachfolgenden Definition und Umbenennung zu einer partial-likelihood-Funktion und zeigte, dass der Verlust an Effizienz sehr gering ist.
Seine Arbeiten regten die methodologische Arbeit über verwandte Bereiche an. Die Cox-Regression wird bis heute sehr oft verwendet, um Behandlungseffekte und Krankheitsrisiken in klinischen Proben zu erklären. Neben der Epidemiologie, zeigten sich die Ideen des David Cox auch in anderen Forschungsbereichen wie Industrie, Technologie, Bevölkerungsstatistik oder in der Wirtschaft als sehr hilfreich.
Der stärkste und am längsten andauernde Einfluss war jener auf die Statistiker, vor allem in der Forschung. Probleme der rechts-zensierten Regression mit Ausfallzeiten konnte David Cox durch das Vergleichen von Sterbezeiten zu bestimmten Zeitpunkten umgehen, wodurch seine Arbeit eine sehr große Bedeutung für Analysemöglichkeiten, zum Beispiel in der (Bio-)Statistischen Forschung hat.

Literaturverzeichnis

[Augustin, T. 2013] Skript der Vorlesung von Prof. Dr. Augustin: Wirtschafts- und Sozialstatistik im WS 2013/14 an der Ludwig-Maximilian Universität, München. http://www.statistik.lmu.de/institut/ag/agmg/lehre/2013_WiSe/Wiso/WiSo_folien_kap_5.3_20140112.pdf

[Fox, J. 2002] John Fox: Cox Proportional-Hazards Regression for Survival Data. http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf

[Kauermann, G. 2013] Skript der Vorlesung von Prof. Dr. Kauermann: Wahrscheinlichkeitstheorie und Inferenz II. http://www.stat.uni-muenchen.de/institut/lehrstuhl/wisoz/lehre/stat4_ss13/download/Statistik_IV.pdf

[Küchenhoff, H. 2013] Skript der Vorlesung von Prof.Dr. Küchenhoff: Generalisierte Regression im WS 2013/14 an der Ludwig-Maximilian Universität, München. http://www.stablab.stat.uni-muenchen.de/sites/files/GRM13-4.pdf

[Schneider, C. 2009] Skript der Vorlesung von PD. Dr. Dr. Schneider: Maß- und Integrationstheorie. http://www.stat.uni-muenchen.de/~christin/Skript_Teil1.pdf

[Slavik, A. 2007] Antonin Slavik: Product Integration, its history and applications. http://www.karlin.mff.cuni.cz/~prusv/ncmm/notes/download/product_integration.pdf

[Stein, P. 2007] Skript der Vorlesung von Prof. Dr. Petra Stein: Ereignisanalyse, an der Universität Duisburg-Essen.

https://www.uni-due.de/imperia/md/content/soziologie/stein/ereignisanalyse.pdf

Erklärung zur Urheberschaft

Hiermit versichere ich, dass ich die vorliegende Seminararbeit selbstständig aus den Vorlagen von Andrea Wiencierz angefertigt habe.

München, den 30. Mai 2014

(Lorenz Winklmaier)