A. Ziegler[1]
S. Lange[2]
R. Bender[3]

Überlebenszeitanalyse: Die Cox-Regression

– Artikel Nr. 16 der Statistik-Serie in der DMW –

Survival analysis: Cox regress

Hintergrund

Das Cox-Modell (4) ist die populärste Regressionsmethode zur Analyse von Überlebensdaten. Es wird auch als proportionales Hazard Modell (engl.: proportional hazards model) bezeichnet. Ganz analog zu anderen Regressionsverfahren, wie der klassischen multiplen linearen Regression (3) oder der logistischen Regression (2) wird das Cox-Modell eingesetzt, wenn gleichzeitig der Effekt mehrerer Einflussgrößen auf eine Zielvariable untersucht werden soll. Bei der Zielvariablen handelt es sich in diesem Fall um zensierte Überlebenszeiten (7). Bei den zu untersuchenden Effekten kann es sich um einfache Zwei-Gruppenvergleiche handeln, wie z.B. in klinischen Therapiestudien. Denn hier ist häufig von Bedeutung, die Relevanz bzw. das Ausmaß des Therapieeffekts hinsichtlich des Überlebens von Patienten unter gleichzeitiger Berücksichtigung weiterer relevanter Einflussfaktoren im Rahmen einer multiplen Überlebenszeitanalyse zu untersuchen. Außerdem kann bei randomisierten Therapiestudien in der Regel mit präziseren Schätzungen durch die Adjustierung nach prognostisch relevanten Variablen gerechnet werden.

Ganz allgemein liefert das Cox-Modell eine Schätzung des Therapieeffekts auf die Überlebenszeit, adjustiert für die anderen Einflussgrößen des Regressionsmodells. Das Modell erlaubt es, den Hazard – salopp gesprochen das unmittelbare Risiko – für eine Person im Hinblick auf den Tod oder ein anderes interessierendes Ereignis zu schätzen. Hierfür müssen aber gleichzeitig die Werte für alle Einflussvariablen dieser Person gegeben sein.

Dabei ist für die Praxis von Bedeutung, dass beim Cox-Modell keine bestimmte Verteilung für die Überlebenszeiten benötigt wird. Stattdessen wird angenommen – und das ist eine sehr wichtige Voraussetzung –, dass die Effekte verschiedener Variablen auf das Überleben über die Zeit konstant sind. Das heißt insbesondere, dass diese Effekte auf einer bestimmten Skala additiv sind. Die Methode selbst ist zu komplex, als dass sie hier im Detail vorgestellt werden sollte. Die Berechnung selbst ist heutzutage problemlos mit Standardstatistikprogrammen möglich. Im Internet ist sogar eine Javascript-Anwendung verfügbar, die die Berechnung des Cox-Modells erlaubt (http://members.aol.com/johnp71/prophaz.html).

Daher ist das Ziel dieses Abschnitts, die Ergebnisse des Cox-Modells näher zu erläutern und deren Interpretation näher zu beschreiben.

Hazard-Funktion

Der zentrale Begriff zur Interpretation der Ergebnisse des Cox-Modells ist die Hazard-Funktion. Während in einer Kohortenstudie mit festem Beobachtungszeitraum für alle Probanden und binärem Endpunkt, z.B. Tod ja/nein, das Zielereignis zu einem festen Zeitpunkt bestimmt wird, ist dieses bei Überlebenszeitstudien mit unterschiedlich langen Beobachtungszeiten nicht oder nur mit großem Informationsverlust möglich (7). Und während sich bei der logistischen Regression (2) die Interpretation der Chance (1) oder des Risikos für das Eintreten des Zielereignisses auf das definierte Ende der Nachbeobachtungszeit bezieht, z.B. „das Risiko innerhalb eines Jahres zu versterben“, ist dieser feste Beobachtungszeitraum beim Cox-Modell nicht unmittelbar gegeben.

Institut
[1] Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck
[2] Abteilung für Medizinische Informatik, Biometrie u. Epidemiologie, Ruhr-Universität Bochum
[3] Institut für Medizinische Biometrie, Epidemiologie und Informatik, Johannes-Gutenberg-Universität Mainz

Korrespondenz
Prof. Dr. rer. nat. Andreas Ziegler · Institut für Medizinische Biometrie und Statistik
Universitätsklinikum Schleswig-Holstein, Campus Lübeck
Universität zu Lübeck · Ratzeburger Allee 160 · Haus 4 · 23538 Lübeck · E-Mail: ziegler@imbs.uni-luebeck.de

Bibliografie
DOI: 10.1055/s-2004-836074
Dtsch Med Wochenschr 2004; 129: T1–3 · © Georg Thieme Verlag Stuttgart · New York · ISSN 0012-0472

Aus diesem Grund wird das Konzept des Hazards bzw. der Hazard-Funktion benötigt. Damit wird die Wahrscheinlichkeit pro Zeiteinheit bezeichnet, dass eine Person innerhalb eines kleinen Zeitintervalls das Zielereignis (z.B. Tod) erfährt, wenn sie denn bis zum Beginn dieses Zeitintervalls überlebt hat. Sie kann daher als das Risiko pro Zeiteinheit für das Sterben zur Zeit t interpretiert werden. Die Hazard-Funktion wird üblicherweise mit $h(t)$ bezeichnet und kann durch

$$h(t) = \frac{\text{Anzahl der Personen mit Zielereignis im Intervall, das bei } t \text{ beginnt}}{(\text{Anzahl an Personen, die bis } t \text{ überleben}) \times \text{Intervallbreite}}$$

beschrieben werden.

Die Hazard-Funktion im Cox-Modell

Ganz analog zum multiplen linearen Regressionsmodell und dem logistischen Regressionsmodell ist auch im Cox-Modell das Ziel die gleichzeitige Schätzung des Einflusses verschiedener Variablen. Dabei wird im Cox-Modell die Hazard-Funktion in Abhängigkeit der Einflussvariablen betrachtet. Das Cox-Modell ist gegeben durch:

$$h(t) = h_0(t) \times \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m),$$

wobei exp wie bei der logistischen Regression die Eulersche „e"-Funktion bezeichnet, $X_1, \ldots, X_m$ die Werte der Einflussvariablen sind, und $\beta_1, \ldots, \beta_m$ die zu schätzenden Regressionskoeffizienten der Einflussvariablen bezeichnen. Letztere geben wie in anderen Regressionsmodellen die Stärke der Bedeutung der jeweiligen Einflussvariablen an. Verändert sich der Wert einer Einflussvariablen, d.h. das Risikoprofil einer Person, dann bestimmen die Koeffizienten β die erwartete Veränderung des Hazards bezogen auf die Veränderung der Einflussvariablen um eine Einheit.

Die Größe $h_0(t)$ heißt Baseline-Hazard und gibt den Hazard für das Sterben, also das Eintreten des Ereignisses, in t an, wenn alle Einflussvariablen gleich null sind. Damit hat der Baseline-Hazard ähnliche Bedeutung wie das Absolutglied, der Achsenabschnitt (engl. „intercept"), in der linearen Regression bzw. der logistischen Regression.

Das Modell kann ähnlich wie das logistische Regressionsmodell als lineare Gleichung dargestellt werden; nur muss hier nicht die logistische Funktion sondern der Logarithmus betrachtet werden:

$$\ln h(t) = \ln h_0(t) + \beta_1 X1 + \beta_2 X_2 + \ldots + \beta_m X_m$$

Nimmt man an, dass der Quotient der Hazard-Funktionen zweier Gruppen A und B – also im Wesentlichen der Quotient von Risiken – über die Zeit hinweg konstant ist, erlaubt dieses eine von der Zeit unabhängige und damit eindeutige Definition des Hazard Ratios:

$$\text{HR} = h_A(t)/h_B(t) = const.$$

Diese Proportionalität ist die zentrale Annahme im Cox-Modell, was gleichzeitig eine Stärke und eine Schwäche dieser Methode ist. Denn die Proportionalität bedeutet z.B. in einer Therapiestudie, dass der Therapieeffekt, wenn er denn tatsächlich besteht, über die Zeit hinweg konstant ist. Entsprechend wäre eine Therapie A gleichmäßig besser als eine Therapie B. Nun wird klar, warum die Annahme proportionaler Hazards auch kritisch betrachtet werden muss. Denn schließlich könnten bei einer neuen Behandlungsform Vorteile gegenüber der Standardtherapie ggf. nur früh oder nur spät in der Zeit, also eben nicht gleichmäßig über die Zeit, auftreten.

Entsprechend muss der Proportionalität besondere Bedeutung in der Überprüfung der Modellannahmen beigemessen werden. Und so könnte die Annahme eines konstanten relativen Risikos nur für einen gewissen Beobachtungszeitraum gelten, für den dann das Cox-Modell auch nur verwendet werden darf; in besonders ungünstigen Fällen müssen andere statistische Verfahren eingesetzt werden.

Beispiel

Das Cox-Modell wird unter Verwendung von Daten zur Stilldauer der Mutter in Abhängigkeit einer Reihe von Einflussvariablen illustriert. Die amerikanischen Daten von 927 erstgeborenen Kindern, deren Mütter stillen, sind im Internet frei verfügbar und wurden schon mehrfach als Beispiel in der Literatur verwendet (5). Ziel der Analyse ist es, ein Modell zu entwickeln, mit dem eine Vorhersage der Zeit bis zum Abstillen möglich ist. Aus einem Fragebogen wurden die folgenden potenziellen Variablen extrahiert: Ethnizität der Mutter (Kaukasier, sonstige), Indikator für den Wohlstand bei Geburt des Kindes, Rauchverhalten der Mutter bei Geburt des Kindes (jeweils ja/nein), Jahr der Geburt ('78 – '86), Dauer der Ausbildung der Mutter (Jahre der Beschulung).

Die Ergebnisse des Cox-Regressionsmodells sind in **Tab.1** dargestellt. Das Hazard-Ratio von 1.28 für die Variable Rauchen zeigt, dass das Risiko für das Abstillen bei Müttern, die zum Zeitpunkt der Geburt rauchen, höher ist; es ist 28% höher als bei den entsprechenden Nichtraucherinnen. Wie bei der logistischen Regression gilt dieses mit dem Zusatz, dass gleichzeitig für die anderen Variablen des Modells adjustiert wurde (2). Dieser methodisch feine aber bedeutsame Zusatz wird im Folgenden der Einfachheit halber weggelassen. Das Konfidenzintervall gibt an, dass mit 95%iger Sicherheit der Bereich zwischen 1.1 und 1.5 das tatsächliche Hazard Ratio für den Vergleich der Risiken für das Abstillen zwischen Müttern, die zum Zeitpunkt der Geburt rauchen mit denen, die zum Zeitpunkt der Geburt nicht rauchen überdeckt. Analog zeigt **Tab.1**, dass weiße Frauen ein geringeres Risiko für das Abstillen als Frauen anderer Ethnizitäten haben.

Doch wie sind nun die stetigen Variablen Geburtsjahr und Dauer der Ausbildung zu interpretieren? Hier gibt das in **Tab.1** angegebene Hazard-Ratio die Erhöhung bzw. Verringerung des Risikos an, wenn die Geburt 1 Jahr später bzw. die Ausbildung 1 Jahr länger ist. So zeigt der Wert 0.94 für die Variable „Dauer der Ausbildung", dass eine Frau mit einer Ausbildung von – sagen

Tab. **1** Ergebnisse des Cox-Modells für die Zeit bis zum Abstillen.

Variable	*Hazard Ratio*	*95% Konfidenzintervall für das Hazard-Ratio*			*p-Wert*
Wohlstand (hoch versus niedrig)	1.21	1.01	–	1.45	0.0417
Rauchen (ja versus nein)	1.28	1.10	–	1.49	0.0017
Geburtsjahr (in Jahren)	1.07	1.03	–	1.11	0.0001
Dauer der Ausbildung (in Jahren)	0.94	0.90	–	0.98	0.0022
Ethnizität (Kaukasier versus nicht Kaukasier)	0.80	0.68	–	0.93	0.0033

wir – 10 Jahren ein 0.94-faches und damit ein geringeres Risiko für das Abstillen hat im Vergleich zu einer Frau, die 9 Jahre zur Schule gegangen ist. Bei der Angabe von Hazard Ratios muss daher immer die Einheit der betrachteten Variablen mit angegeben werden.

Prinzipiell lassen sich auch die Risikounterschiede zwischen Frauen mit verschiedenen Gruppen von Einflussvariablen berechnen. Doch übersteigt dieses das Ziel der vorliegenden Arbeit. Hier wird der interessierte Leser zum Selbststudium auf das Buch von Kleinbaum (6) verwiesen.

kurzgefasst: **Mit Hilfe des Cox-Modells lässt sich der Einfluss von erklärenden Variablen auf eine Überlebenszeit untersuchen. Aus den Regressionskoeffizienten lassen sich adjustierte Hazard Ratios als Maß für die Stärke des Zusammenhangs berechnen.**

Literatur

[1] Bender R, Lange S. Die Vierfeldertafel. Dtsch Med Wochenschr 2001; 126: T36–8
[2] Bender R, Ziegler A, Lange S. Logistische Regression. Dtsch Med Wochenschr 2002; 127: T11–3
[3] Bender R, Ziegler A, Lange S. Multiple Regression. Dtsch Med Wochenschr 2002; 127: T8–11
[4] Cox DR. Regression models and life tables. J Roy Stat Soc B 1972; 34: 187–220
[5] Klein JP, Moeschberger ML. Survival analysis. Techniques for censored and truncated data. New York: Springer, 1997
[6] Kleinbaum DG. Survival analysis: A self-learning text. New York: Springer, 1996
[7] Ziegler A, Lange S, Bender R. Überlebenszeitanalyse: Eigenschaften und Kaplan-Meier Methode. Dtsch Med Wochenschr 2002; 127: T14–16