Simulation von Überlebenszeiten mit Hilfe von SAS®

Ralf Bender
Institut für Qualität und Wirtschaftlichkeit
im Gesundheitswesen (IQWiG)
Dillenburger Straße 27
51105 Köln
Ralf.Bender@iqwig.de

Zusammenfassung

Simulationsstudien stellen ein wichtiges statistisches Hilfsmittel dar, um die Eigenschaften und die Güte statistischer Methoden in bestimmten, klar spezifizierten Datensituationen zu untersuchen, in denen keine analytische Lösung möglich ist. Eine der wichtigsten statistischen Methoden in der biomedizinischen Forschung ist das Cox Proportional Hazards Modell. Zu diesem Modell gibt es zahlreiche Simulationsuntersuchungen, wobei jedoch häufig nur exponentialverteilte Zufallszahlen verwendet werden. Da das Cox Modell über die Hazardfunktion definiert ist, für Simulationsstudien aber die entsprechenden Überlebenszeiten benötigt werden, ist nicht sofort ersichtlich, wie die zu einem spezifizierten Cox Modell passenden Überlebenszeiten erzeugt werden können. In dieser Arbeit wird ein allgemeiner Ansatz zur Erzeugung von Überlebenszeiten für das Cox Modell vorgestellt, welches die wichtigen Verteilungen Exponential-, Weibull- und Gompertz-Verteilung als Spezialfälle enthält. Es wird gezeigt, wie sich der allgemeine Ansatz zur Erzeugung von Überlebenszeiten für Simulationen zum Cox Modell mit Hilfe der Zufallszahlfunktionen von SAS® auf einfache Weise umsetzen lässt.

Schlüsselwörter: Cox Modell, Exponentialverteilung, GompertzVerteilung, Simulation, Überlebenszeiten, Weibull-Verteilung, Zufallszahlen.

1 Einleitung

Das Proportional Hazards Modell von Cox (1972) stellt eines der wichtigsten statistischen Methoden zur Analyse von Überlebenszeiten dar. Um die Eigenschaften dieses Verfahrens zu untersuchen, wurden zahlreiche Simulationsstudien durchgeführt. Die verwendete Methodik wird leider häufig in solchen Simulationsstudien nur unzureichend beschrieben (Burton et al., 2006). Ein spezielles Problem bei Simulationen zum Cox Modell ist es, mit Hilfe von Zufallszahlen künstliche Überlebenszeiten genau so zu erzeugen, dass sie als Realisationen eines vorgegebenen Cox Modells angesehen werden können. In Simulationsstudien zu linearen Regressionsmodellen stellt dies kein Problem dar, da die zu erzeugenden Zielvariablen direkt mit den erklärenden Variablen funktional zusammenhängen. Beim Cox Modell ergibt sich jedoch die Schwierigkeit, dass das Modell über die Hazardfunktion definiert ist, die für Simulationsstudien zu erzeugenden Daten jedoch durch die Überlebenszeiten gegeben sind. Der Zusammenhang zwischen der Hazardfunktion und den zugehörigen Überlebenszeiten eines Cox Modells

ist bei der Exponentialverteilung einfach herzuleiten, so dass diese Verteilung sehr häufig für Simulationsstudien zum Cox Modell verwendet wird. Da es jedoch nicht ausreichend ist, in Simulationsstudien zum Cox Modell nur die einfache Exponentialverteilung zu betrachten, werden Methoden benötigt, um Überlebenszeiten auch für andere Verteilungen zu erzeugen (Bender, Augustin & Blettner, 2005).
Leemis (1987) sowie Bender, Augustin & Blettner (2005) haben einen allgemeinen Ansatz beschrieben, mit dem Überlebenszeiten aus nahezu beliebigen Verteilungen für Simulationsstudien zum Cox Modell sehr einfach erzeugt werden können. Insbesondere lassen sich die Exponential-, die Weibull- und die Gompertz-Verteilung in diesen Ansatz einbetten (Bender, Augustin & Blettner 2005). In dieser Arbeit wird beschrieben, wie sich der allgemeine Ansatz zur Erzeugung von Überlebenszeiten für das Cox Modell mit Hilfe der Zufallszahlfunktionen auf einfache Weise in SAS® umsetzen lässt. Die Methode wird illustriert mit Hilfe einiger beispielhafter Simulationsläufe, in denen die Eignung der Exponential- und der Gompertz-Verteilung zur Erzeugung realistischer Überlebenszeiten auf der Basis einer großen Kohortenstudie untersucht wird.

2 Simulation von Überlebenszeiten

2.1 Allgemeiner Ansatz

Das Cox Modell wird üblicherweise über die Hazardfunktion definiert durch

$$h(t|x) = h_0(t)\exp(\beta' x) \tag{1}$$

wobei t die Zeit, x den Vektor der betrachteten Kovariablen, β den Vektor der Regressionskoeffizienten und $h_0(t)$ die im Modell unspezifizierte Baseline Hazardfunktion darstellt. Die Überlebensfunktion des Cox Modells (1) ist gegeben durch

$$S(t|x) = \exp[-H_0(t)\exp(\beta' x)] \tag{2}$$

wobei $H_0(t)$ die kumulative Baseline Hazardfunktion ist (Kalbfleisch & Prentice, 2002). Hieraus lässt sich die Verteilungsfunktion des Cox Modells ableiten, die gegeben ist durch

$$F(t|x) = 1 - \exp[-H_0(t)\exp(\beta' x)] \tag{3}$$

Sei Y eine Zufallsvariable mit Verteilungsfunktion F, so gilt allgemein, dass $U=F(Y)$ eine auf [0,1] gleichverteilte Zufallsvariable ist (Mood, Graybill & Boes, 1974), d.h. $U \sim \mathrm{U}[0,1]$. Ebenso gilt $1-U \sim \mathrm{U}[0,1]$. Sei nun T die zugehörige Überlebenszeit zum Cox Modell (1), so folgt bei Anwendung dieser Regeln auf (3), dass

$$U = \exp[-H_0(T)\exp(\beta' x)] \sim \mathrm{U}[0,1] \tag{4}$$

Unter der Voraussetzung einer überall positiven Baseline Hazardfunktion kann $H_0(t)$ invertiert werden und die Überlebenszeit T des Cox Modells (1) ergibt sich als

$$T = H_0^{-1}[-\log(U)\exp(-\beta' x)] \tag{5}$$

wobei U eine Zufallsvariable ist mit $U \sim \mathrm{U}[0,1]$.

Mit Hilfe der allgemeinen Formel (5) lassen sich nun für beliebige Verteilungen mit positiver Hazardfunktion Überlebenszeiten zu spezifizierten Cox Modellen erzeugen. Man benötigt lediglich gleichverteilte Zufallszahlen sowie die inverse kumulative Baseline Hazardfunktion.

2.2 Verwendung bekannter Verteilungen

Von den bekannten Verteilungen für Überlebenszeiten sind nur die Exponential-, die Weibull- und die Gompertz-Verteilung mit der Annahme proportionaler Hazards kompatibel (Lee & Go, 1997). Eine Übersicht über diese drei Verteilungen mit einer detaillierten Formelsammlung geben Bender, Augustin & Blettner (2005). Die Anwendung von (5) für diese drei Verteilungen führt zu den Formeln

Exponentialverteilung: $$T = -\frac{\log(U)}{\lambda \exp(\beta' x)} \quad , \quad h(t \mid x) = \lambda \exp(\beta' x) \tag{6}$$

Weibull-Verteilung: $$T = \left(-\frac{\log(U)}{\lambda \exp(\beta' x)} \right)^{\frac{1}{\nu}} , \quad h(t \mid x) = \lambda \exp(\beta' x)\, \nu\, t^{\nu-1} \tag{7}$$

Gompertz-Verteilung: $$T = \frac{1}{\alpha} \log\left[1 - \frac{\alpha \log(U)}{\lambda \exp(\beta' x)} \right] , \quad h(t \mid x) = \lambda \exp(\beta' x) \exp(\alpha t) \tag{8}$$

Bei allen drei Verteilungen erkennt man an den Hazardfunktionen, dass ein Cox Modell mit einer Baseline Hazard, die zu einer der drei Verteilungen mit Parameter λ gehört, zu der gleichen Verteilung der Überlebenszeit des Cox Modells führt, allerdings mit einem Parameter $\lambda^*(x) = \lambda \exp(\beta' x)$, der somit eine Funktion der Kovariablen darstellt.

2.3 Andere Verteilungen

Mit Hilfe der allgemeinen Formel (5) lassen sich auch beliebige Verteilungen zur Erzeugung von Überlebenszeiten für das Cox Modell verwenden. Die einzige Voraussetzung ist, dass die inverse kumulative Baseline Hazardfunktion zumindest numerisch bestimmbar sein muss. Die Anwendung einer Baseline Hazard, die nicht aus der Exponential-, Weibull- oder Gompertz-Verteilung kommt, führt allerdings zu bislang nicht bekannten Überlebenszeitverteilungen. Beispiele hierfür werden von Bender, Augustin & Blettner (2005) beschrieben.

2.4 Umsetzung in SAS®

Bei bekannter inverser kumulativer Baseline Hazardfunktion lässt sich Formel (5) direkt in SAS® in einem Data Step umsetzen, da SAS® mit der Funktion RANUNI über einen Generator für gleichverteilte Zufallszahlen verfügt (SAS Institute Inc., 2004a). Speziell für die Exponentialverteilung kann auch die Funktion RANEXP verwendet werden (SAS Institute Inc., 2004a). Seit Version 9 steht außerdem die neue Funktion RAND zur Verfügung, mit der unter anderem exponential- und Weibull-verteilte Zufallszahlen erzeugt werden können (SAS Institute Inc., 2004a). Eine weitere Möglichkeit stellt die

Verwendung der Call Routine RANDGEN in SAS/IML® dar (SAS Institute Inc., 2004b). Die Funktionsweise entspricht derjenigen der Funktion RAND innerhalb eines Data Steps (SAS Institute Inc., 2004b).
Im Folgenden werden die möglichen SAS® Codes zur Erzeugung von Zufallszahlen gemäß der Exponential-, der Weibull- und der Gompertz-Verteilung im Rahmen eines Data Steps zusammengestellt. Hierbei ist `X` eine im Datensatz vorhandene Kovariable (z.B. erzeugt mit Hilfe der Funktion RANNOR), `beta` der Wert des Regressionskoeffizienten und `lamda`, `alpha`, `nue`, `a` und `b` die Parameter der entsprechenden Verteilungen. Für `seed` ist eine nicht-negative ganze Zahl $<2^{31}-1$ einzusetzen. Bei Verwendung der Null wird die aktuelle Uhrzeit als Startwert verwendet (SAS Institute Inc., 2004a).

2.4.1 Exponentialverteilung

a) Verwendung der Funktion RANUNI:

```
T = -log(RANUNI(seed))/(lamda*exp(beta*X));
```

b) Verwendung der Funktion RANEXP:

```
T = RANEXP(seed)/(lamda*exp(beta*X));
```

c) Verwendung der Funktion RAND:

```
T = RAND('EXPONENTIAL')/(lamda*exp(beta*X));
```

2.4.2 Weibull-Verteilung

a) Verwendung der Funktion RANUNI:

```
T = (-log(RANUNI(seed))/(lamda*exp(beta*X)))**(1/nue);
```

b) Verwendung der Funktion RAND:

Bei Verwendung der Funktion RAND ist zu beachten, dass in SAS® die Weibull-Verteilung anders parametrisiert ist als in Formel (7). Es müssen also zunächst die für RAND benötigten Paramter `a` und `b` aus `lamda` und `nue` berechnet werden.

```
a = nue;
b = 1/((lamda*exp(beta*X))**(1/a));
T = RAND('WEIBULL',a,b);
```

2.4.3 Gompertz-Verteilung

Zur Erzeugung von Gompertz-verteilten Überlebenszeiten gibt es in SAS® keine direkte Möglichkeit. Mit Hilfe der Funktion RANUNI kann aber der allgemeinen Ansatz (5) auch für die Gompertz-Verteilung verwendet werden.

```
U = RANUNI(seed);
T = (1/alpha)*log(1-((alpha*log(U))/(lamda*exp(beta*X))));
```

2.4.4 Zensierung

Wenn wie oben beschrieben exponential-, Weibull- oder Gompertz-verteilte Zufallszahlen erzeugt werden, so muss in einem nächsten Schritt noch eine Zensierung eingeführt werden, um die typische Datenlage klinischer oder epidemiologischer Studien mit Überlebenszeiten abzubilden. Eine Zensierung kann entweder über eine Modellierung des Zensierungsprozesses (Burton et al., 2006) oder über die Bestimmung oder Simulation von variierenden Eintrittsdaten und der Festlegung eines fixen Zeitpunkts für das Beobachtungsende eingeführt werden. Letzteres ist insbesondere adäquat wenn man Überlebenszeiten für eine ganz bestimmte Studiensituation mit bereits vorliegenden Eintrittsdaten simulieren möchte. Im Folgenden wird ein möglicher SAS® Code für diese Situation vorgestellt.

Sei `ENTRY` die vorliegende Variable für das (variierende) Eintrittsdatum jeder Beobachtungseinheit, `FUDATE` das (fixe) Datum des Beobachtungsendes, `T` die mit einer der oben beschriebenen Methoden erzeugten (nicht zensierten) Variable der Überlebenszeiten (mit der Einheit Jahr), dann kann man auf folgende Weise eine Zensierung erzeugen. Die neue Variable `ST` enthält hierbei die zensierten Überlebenszeiten (mit der Einheit Jahr) und der binäre Indikator `EVENT` enthält die Information, ob das interessierende Ereignis eingetreten ist.

```
DATA SIM;    ...
  EVENT=0;   END=FUDATE;
  DAYS=T*365.25;
  if ENTRY+DAYS<FUDATE then do;
    EVENT=1; END=ENTRY+DAYS;
  end;
  ST = (END-ENTRY)/365.25;
```

3 Beispiel

Die deutsche Uranbergarbeiterstudie ist eine große Kohortenstudie zur Untersuchung des Zusammenhangs zwischen Radonbelastung und Krebsmortalität (Kreuzer et al., 2002). Diese Studie enthält Daten von ca. 60000 ehemaligen Bergarbeitern der früheren SAG/SDAG WISMUT. Die Exposition durch Radon wurde mit Hilfe einer Job-Expositions-Matrix (JEM) berechnet. Da neben der Radonbelastung auch weitere Risikofaktoren wie Rauchen, Arsen und Stäube berücksichtigt werden sollten, kommt für die Datenanalyse insbesondere das Cox Modell in Frage. Aufgrund der Eigenschaften der JEM spielen bei der Datenauswertung die Effekte von Messfehlern in der Exposition eine Rolle, die mit Hilfe einer Simulationsstudie untersucht wurden (Bender & Blettner, 2002). Da diese Simulationen speziell die Datensituation der deutschen Uranbergarbeiterstudie wiedergeben sollten, mussten die Überlebenszeiten so simuliert werden, dass sich eine für deutsche Bergarbeiter realistische Mortalität ergibt, d.h. es sollten a) – wie in der Kohortenstudie tatsächlich beobachtet – zwischen 10000 und 20000 Ereignisse vorkommen und b) sich realistische Altersverteilungen ergeben.

Anhand von beispielhaften Simulationsläufen wird illustriert, dass mit Hilfe der Gompertz-Verteilung realistische Überlebenszeiten für die Situation der deutschen Uranbergarbeiterstudie erzeugt werden können, nicht jedoch mit Hilfe der Exponentialverteilung. Für die Simulationen wurden die SAS® Codes aus den Abschnitten 2.4.1 bzw. 2.4.3 verwendet. Der Einfachheit halber betrachten wir nur das Eintrittsalter als Kovariable mit Regressionskoeffizient $\beta=0.15$, was einem Hazard Ratio von HR=1.16 pro Altersjahr entspricht. In Tabelle 1-3 befinden sich deskriptive Daten zu drei Simulationsläufen, zwei mit Verwendung der Exponential- ($\lambda=0.001$ bzw. $\lambda=0.0002$) und einer mit Verwendung der Gompertz-Verteilung ($\lambda=0.7\times10^{-7}$, $\alpha=0.2138$).

Tabelle 1: Simulierte Überlebenszeiten mit Hilfe der Exponentialverteilung ($\lambda=0.001$)

Variable	Anzahl	Minimum	Maximum	Mittelwert	SD	Summe
Eintrittsalter	59813	12.45	73.80	24.30	8.3814	1453712
Ereignis	59813	0	1	0.9785	0.1452	**58525**
T	59813	0.000012	79.38	4.16	5.8008	249028
ST	59813	0.000012	50.39	4.00	5.1942	239349
Todesalter	59813	**14.29**	**95.72**	**28.47**	7.8945	1702740

Tabelle 2: Simulierte Überlebenszeiten mit Hilfe der Exponentialverteilung ($\lambda=0.0002$)

Variable	Anzahl	Minimum	Maximum	Mittelwert	SD	Summe
Eintrittsalter	59813	12.45	73.80	24.30	8.3814	1453712
Ereignis	59813	0	1	0.3041	0.4600	**18187**
T	59813	0.000069	3845.11	208.04	288.94	12443234
ST	59813	0.000069	53.00	28.09	15.5819	1679937
Todesalter	59813	**14.26**	**3861.64**	**232.34**	285.46	13896946

Tabelle 3: Simulierte Überlebenszeiten mit Hilfe der Gompertz-Verteilung ($\lambda=0.7\times10^{-7}$, $\alpha=0.2138$)

Variable	Anzahl	Minimum	Maximum	Mittelwert	SD	Summe
Eintrittsalter	59813	12.45	73.80	24.30	8.3814	1453712
Ereignis	59813	0	1	0.2341	0.4235	**14005**
T	59813	0.95	69.30	50.07	8.4223	2994631
ST	59813	0.95	53.00	34.83	11.9168	2083173
Todesalter	59813	**29.04**	**98.46**	**74.37**	6.4943	4448344

Man erkennt an den Tabellen 1-3, dass es mit Hilfe der Exponentialverteilung hier nicht möglich ist, realistische Überlebenszeiten zu erzeugen. Entweder ist die Zahl der Ereignisse zu hoch (Tab. 1: 58525 Ereignisse) oder die Altersverteilung ist unrealistisch (Tab. 2: mittleres Todesalter 232 Jahre). Realistische Überlebenszeiten gelingen jedoch mit Hilfe der Gompertz-Verteilung (Tab. 3: 14005 Ereignisse, mittleres Todesalter 74 Jahre), so dass in der Simulationsstudie zur deutschen Uranbergarbeiterstudie die Gompertz-Verteilung verwendet wurde (Bender & Blettner, 2002).

Ausführliche Ergebnisse der Simulationsuntersuchungen bezüglich des Effekts von Messfehlern auf die Parameterschätzung im Cox Modell sind bei Bender & Blettner (2002), Bender, Augustin & Blettner (2005) sowie Küchenhoff, Bender & Langner (2007) zu finden.

4 Schlussfolgerung

In Simulationsstudien zum Cox Modell genügt es nicht, zur Erzeugung von Überlebenszeiten die einfache Exponentialverteilung mit konstanter Hazardfunktion zu betrachten. Um spezielle Datensituationen in realistischer Weise abzubilden, werden flexiblere Verteilungen wie die Weibull- oder die Gompertz-Verteilung benötigt. Die beim Cox Modell übliche Trennung von Baseline Hazard und Kovariablen kann in Nicht-Standardsituationen, wie z.B. beim Auftreten von Messfehlern, im Allgemeinen nicht aufrechterhalten werden (Prentice, 1982). Aus diesem Grund können die Ergebnisse von Simulationsstudien zum Cox Modell durchaus von der gewählten Verteilung abhängen (Bender, Augustin & Blettner, 2005). Der allgemeine Ansatz (5) ermöglicht die Verwendung beliebiger Verteilungen mit positiver Hazardfunktion zur Erzeugung von Überlebenszeiten und stellt damit die Voraussetzung dar, die Eigenschaften des Cox Modells mit Hilfe von Simulationen in umfassender Weise zu untersuchen (Leemis, 1987; Bender, Augustin & Blettner, 2005). Mit Hilfe der in SAS® implementierten Zufallszahlfunktionen lässt sich der allgemeine Ansatz zur Erzeugung von Überlebenszeiten für das Cox Modell in SAS® im Rahmen eines Data Steps oder innerhalb von SAS/IML® direkt und bequem mit wenigen Programmzeilen umsetzen.

Literatur

[1] R. Bender, T. Augustin, M. Blettner: Generating survival times to simulate Cox proportional hazards models. *Stat. Med.* 2005; 24: 1713-1723.

[2] R. Bender, M. Blettner: Diskussion der Messfehlerproblematik durch Verwendung einer Job-Exposure-Matrix (JEM). In: Geschäftsstelle der Strahlenschutzkommission beim Bundesamt für Strahlenschutz (Hrsg): *Stand der Forschung zu den "Deutschen Uranbergarbeiterstudien", 1. Fachgespräch am 7./8. Mai 2001 in St. Augustin*. Urban und Fischer, München, 2002: 97-105.

[3] A. Burton, D.G. Altman, P. Royston, R.L. Holder: The design of simulation studies in medical statistics. *Stat. Med.* 2006; 25: 4279-4292.

[4] D.R. Cox: Regression models and life tables (with discussion). *J. R. Stat. Soc. B* 1972; 34: 187-220.

[5] J.D. Kalbfleisch, R.L. Prentice: *The Statistical Analysis of Failure Time Data, 2nd Ed*. Wiley, Hoboken, NJ, 2002.

[6] M. Kreuzer, A. Brachner, F. Lehmann, K. Martignoni, H.E. Wichmann, B. Grosche: Characteristics of the German Uranium Miners Cohort Study. *Health Phys.* 2002; 83, 26-34.

[7] H. Küchenhoff, R. Bender, I. Langner: Effect of Berkson measurement error on parameter estimates in Cox regression models. *Lifetime Data Anal.* 2007; 13: 261-272

[8] E.T. Lee, O.T. Go: Survival analysis in public health research. *Annu. Rev. Public Health* 1997; 18: 105-134.

[9] L.M. Leemis: Variate generation for accelerated life and proportional hazards models. *Oper. Res.* 1987; 35: 892-894.

[10] A.M. Mood, F.A. Graybill, D.C. Boes: *Introduction to the Theory of Statistics*. McGraw-Hill, New York, 1974.

[11] R.L. Prentice: Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 1982; 69: 331-342.

[12] SAS Institute Inc.: *SAS® 9.1 Language Reference: Dictionary*. SAS Institute Inc., Cary, NC, 2004a.

[13] SAS Institute Inc.: *SAS/IML® 9.1 User's Guide*. SAS Institute Inc., Cary, NC, 2004b.