Survival-Analyse mit zeitabhängigen Variablen

Anja Schoeps
Universität Heidelberg
Institut für Public Health
Im Neuenheimer Feld 324
69120 Heidelberg
Schoeps@uni-heidelberg.de

Zusammenfassung

In der Survival-Analyse kommt es häufig vor, dass sich Einflussvariablen im Laufe der Beobachtungszeit ändern. Solche Änderungen können und sollten bei der Proportional Hazards Regression mit PROC PHREG einbezogen werden, da sie großen Einfluss auf die Ergebnisse haben.
Ziel dieses Beitrages ist es, die notwendigen Datenstrukturen und die Anwendung mit PROC PHREG zu beschreiben sowie einen kurzen Überblick über die Darstellung von Ergebnissen und Modellgütekriterien zu geben.

Nach einer kurzen Einführung in PROC PHREG werden der Umgang mit zeitabhängigen Variablen und die dafür nötige Struktur der Datentabelle beschrieben. Die eigentliche Kodierung der zeitabhängigen Variablen erfolgt dann innerhalb von PROC PHREG. Hierzu werden zum Beispiel If-Statements, Else-if- oder Select-when-Statements wie im Datenschritt verwendet.
Der zweite Teil dieser Arbeit beschreibt die Ausgabedetails von PROC PHREG.
Schließlich werden Tipps zur geschickten Anwendung von PROC PHREG gegeben.

Schlüsselwörter: PROC PHREG, PROC LIFETEST, PROC SURVEYSELECT, kategoriale Variablen, kontinuierliche Variablen, zeitabhängige Variablen, Modellgüte

1 Einleitung

Bei der Analyse von Survivaldaten kommt es häufig vor, dass sich einige der Einflussvariablen im Laufe der Beobachtungszeit ändern. Um verfälschte Ergebnisse zu vermeiden, ist es notwendig, solche Änderungen bei der Proportional Hazards Regression mit PROC PHREG einzubeziehen. Für diese Analysen sind bestimmte Strukturen in der Datentabelle sowie eine Kodierung der zeitabhängigen Variablen innerhalb von PROC PHREG erforderlich. Ziel dieses Beitrages ist es, diese notwendigen Datenstrukturen und den Umgang mit PROC PHREG mit zeitabhängigen Variablen näher zu beschreiben.

2 Beispieldaten

Die Beispieldaten stammen aus einer Studie zu Einflussfaktoren auf das Überleben nach Kehlkopfkrebsdiagnose, die im Rhein-Neckar-Odenwald-Kreis durchgeführt wurde und ca. 600 Patienten umfasst. Als Zeitvariable wurde in dieser Studie die Zeit von Beobachtungsbeginn bis Beobachtungsende in Tagen betrachtet. Der Endpunkt von Interesse war Mortalität. Im Folgenden werden beispielhaft 3 der Einflussfaktoren in der Analyse beschrieben, um den Umgang mit unterschiedlichen Variablentypen zu veranschaulichen (Tabelle 1).

Tabelle 1: Im Beispiel verwendete Variablen

Name	*Variablentyp*	*Bedeutung*	*Ausprägungen*
AdvancedT	Binär	T Stadium des Primärtumors bei Diagnose	0=Stadium T1, T2 oder unbekannt 1=Stadium T3 oder T4
Rezidiv	Diskret/ Kontinuierlich	Anzahl der Rezidive (inklusive Metastasen)	0 bis 4 Rezidive
Tumlok	Kategorial	Lokalisation des Primärtumors innerhalb des Kehlkopfes	„gl“=glottisch (inkl. transglottisch) „su“=supraglottisch „un“=unbekannt

3 Grundlagen

3.1 Survival-Analyse

Man unterscheidet zwischen deskriptiver und analytischer Survivalanalyse: Die deskriptive Analyse beschäftigt sich mit Überlebenstafeln und Überlebenskurven, die die Überlebenswahrscheinlichkeit in Abhängigkeit von der Zeit seit Beobachtungsbeginn darstellen. Diese Überlebenswahrscheinlichkeiten können zusätzlich nach *genau* einer kategorialen Einflussvariablen stratifiziert werden. Unterschiede in den Überlebenswahrscheinlichkeiten lassen sich mit dem Log-Rank Test analysieren. Diese Form der Survival-Analyse kann in SAS mit PROC LIFETEST durchgeführt werden.
In der Regressionsanalyse wird der simultane Einfluss einer Vielzahl von Variablen auf das Überleben untersucht. Für die Regressionsanalyse von Survivaldaten wird in den meisten Fällen die *Cox Proportional Hazards* Regression verwendet, die sich in SAS mit PROC PHREG (Proportional Hazards Regression) durchführen lässt.

3.2 Mathematische Grundlagen von PROC PHREG

Der Fokus bei der Proportional Hazards Regression liegt auf dem Vergleich der Risiken zwischen Beobachtungen mit unterschiedlichen Merkmalsausprägungen; die Berechnung eines absoluten Sterberisikos ist mit diesem Modell nicht möglich. In einem Mo-

dell ohne zeitabhängige Variablen kann die Formel für den Vergleich der Risiken zum Zeitpunkt t zwischen Individuum i und Individuum j mit n Variablen folgendermaßen dargestellt werden:

$$3.1 \qquad \frac{h_i(t)}{h_j(t)} = \exp\left[\beta_1\left(\mathrm{x}_{i1} - x_{j1}\right) + \cdots + \beta_\mathrm{n}\left(\mathrm{x}_{in} - x_{jn}\right)\right]$$

Durch die Division der Risiken zweier Beobachtungen kürzt sich die Baseline Hazard Funktion heraus. Da die verbleibende Gleichung völlig unabhängig von der Zeit ist, bleibt das Verhältnis zwischen den Risiken über die gesamte Beobachtungsdauer gleich (Proportional Hazards).
PROC PHREG liefert für alle Einflussvariablen die Schätzer β, den Standardfehler, Wald Chi-Quadrat, p-Wert und ein berechnetes Hazard Ratio. Das ausgegebene Hazard Ratio für eine Variable $\mathrm{x_k}$ mit dem Schätzer β_k wird von SAS folgendermaßen berechnet: $HR = \exp(\beta_\mathrm{k})$.
PROC PHREG unterscheidet dabei nicht zwischen binären, kategorialen und kontinuierlichen Variablen. Bei der Interpretation wird immer die Risikoerhöhung bei Erhöhung der unabhängigen Variablen um eine Einheit berechnet. Sollen kategoriale Variablen in das Modell einbezogen werden, werden die einzelnen Kategorien in Dummy-Variablen umkodiert (siehe Kap. 3.3). Bei einer Kodierung in 0 und 1 gilt: $(\mathrm{x}_{ik} - x_{jk}) = 1$. Daher können die ausgegebenen Werte des Hazard Ratio für binäre oder kategoriale Variablen als das Risiko für Beobachtungen mit der Ausprägung von Interesse verglichen mit der Baseline-Ausprägung interpretiert werden.
Handelt es sich um diskrete oder kontinuierliche Variablen, wird das Hazard Ratio demzufolge für zwei Beobachtungen berechnet, deren Werte der Variable sich genau um den Wert 1 unterscheiden. Betrachtet man beispielsweise eine Variable, die das Alter eines Individuums in Jahren angibt, gibt das Hazard Ratio das Verhältnis der Risiken zweier Individuen an, die einen Altersunterschied von genau einem Jahr aufweisen.
Eine komplexe Beschreibung des Proportional Hazards Modell findet sich bei Allison, 1995 [1].

3.3 PROC PHREG Code

Der SAS Code für die Proportional Hazards Regression muss mindestens ein Model-Statement enthalten.

```
proc phreg data=regression;
class tumlok (ref="gl");
model time * verstorb(0) = advancedt rezidiv tumlok;
run;
```

Im Model-Statement müssen eine Überlebenszeitvariable (*time*), eine Zensurvariable (*verstorb*) sowie mindestens eine Einflussvariable (bspw. *advancedt*) enthalten sein. Die Überlebenszeitvariable gibt die Zeit von Beobachtungsbeginn bis Beobachtungsende an,

während die Zensurvariable angibt, ob ein bestimmtes Ereignis von Interesse (hier: Tod des Patienten) eingetreten ist. Hinter der Zensurvariablen wird in Klammern der Wert angegeben, den die Zensurvariable bei Nicht-Eintreten des Ereignisses annimmt (Zensurwert).
Das Class-Statement (seit SAS Version 9.2) erleichtert die Kodierung von Dummy-Variablen für kategoriale Variablen - ein Aufwand, der sonst per Hand durchgeführt werden müsste. Trotz der roten Darstellung des Class-Statements im Enhanced Editor funktioniert dieses Statement problemlos. Im Beispiel handelt es sich bei der Variablen *Tumlok* um eine kategoriale Variable für die als Referenz die Ausprägung „gl“ für glottische Tumoren verwendet werden soll. Alternativ ist es natürlich auch möglich, die Dummy-Variablen von Hand zu kodieren und danach die einzelnen Variablen (mit Ausnahme der Referenzkategorie) in das Model-Statement einzubeziehen.

3.4 Grundlagen der zeitabhängigen Variablen

Variablen, deren Werte sich im Laufe der Beobachtungszeit verändern, werden als zeitabhängige Variablen bezeichnet. Die Variable *Rezidiv* in diesem Beispiel ist eine zeitabhängige Variable, da sich die Anzahl der Rezidive über die Beobachtungszeit erhöht (Abbildung 1). Die schwarzen Pfeile geben den Zeitpunkt an, an dem das jeweils letzte Rezidiv (einschließlich Metastasen) der Patienten aus der Gruppe mit drei oder vier Rezidiven diagnostiziert wurde. Beim Vergleich der Überlebenskurve dieser Patienten mit der Überlebenskurve der Patienten mit ein oder zwei Rezidiven fällt auf, dass sich der Verlauf der Kurven innerhalb der ersten sieben Jahre seit Beobachtungsbeginn kaum unterscheidet. Berücksichtigt man aber den Diagnosezeitpunkt des jeweils letzten Rezidivs (Pfeile), erkennt man, dass die Patienten kurz nach der Diagnose des jeweils letzten Rezidivs verstarben. Da diese Patienten natürlich nicht bereits ab dem Zeitpunkt der Erstdiagnose drei oder vier Rezidive aufwiesen, gehörten sie bis zur Diagnose des dritten Rezidivs in die Gruppe der Patienten mit ein oder zwei Rezidiven. Gleiches gilt für die Patienten in der Gruppe mit ein oder zwei Rezidiven, die bis zur Diagnose des ersten Rezidivs zur Gruppe der Patienten mit null Rezidiven gehörten. Dieser Sachverhalt wird in der Survival-Analyse mit zeitabhängigen Variablen berücksichtigt.

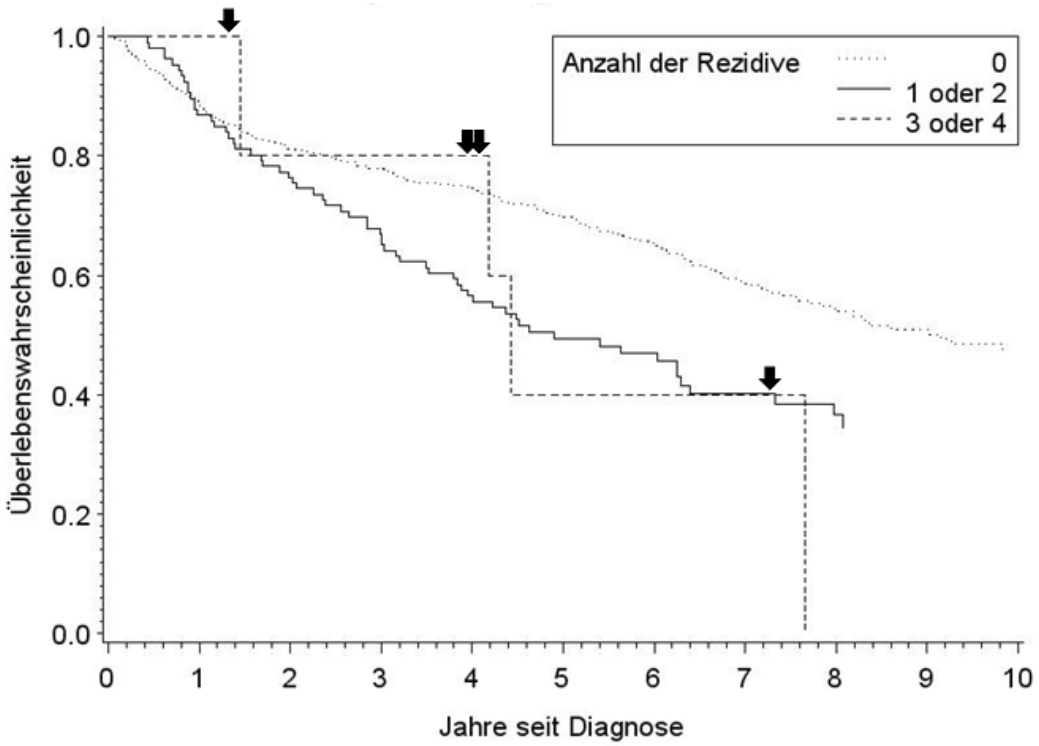

Abbildung 1: Überlebenswahrscheinlichkeit nach Kehlkopfkrebsdiagnose in Abhängigkeit von der Zeit seit Diagnose, stratifiziert nach Anzahl der Rezidive. Pfeile markieren den Diagnosezeitpunkt des jeweils letzten Rezidivs der Patienten, die mindestens 3 Rezidive (einschl. Metastasen) entwickelten.

4 PROC PHREG mit zeitabhängigen Variablen

Für die Implementierung von zeitabhängigen Variablen in PROC PHREG werden neben den unterschiedlichen Werten der Variablen über die Beobachtungszeit auch Informationen zu den entsprechenden Zeitpunkten der Änderung benötigt. Das grundsätzliche Modell 3.1 erweitert sich dabei zu folgender Gleichung, falls die m-te Variable abhängig von der Zeit (t) ist.

$$4.1 \quad \frac{h_i(t)}{h_j(t)} = \exp\left[\beta_1\left(\mathrm{x}_{i1} - x_{j1}\right) + \cdots + \beta_m\left(\mathrm{x}_{im}(\mathrm{t}) - x_{jm}(\mathrm{t})\right) + \cdots + \beta_\mathrm{n}\left(\mathrm{x}_{in} - x_{jn}\right)\right]$$

4.1 Datenstruktur

Grundsätzlich gibt es zwei Möglichkeiten zur Konstruktion der Datenstruktur: Bei der ersten Methode wird jeweils eine Spalte pro möglichem Änderungszeitpunkt einer zeitabhängigen Variable benötigt, die den aktuellen Wert der Variable enthält. Da diese Struktur jedoch bei beliebigen Änderungszeitpunkten zu einer unüberschaubaren Anzahl von Spalten führen würde, ist diese Methode nur bei regelmäßigen Erhebungen der Variablen sinnvoll, wie beispielsweise einer monatlichen Messung des Blutdrucks.
Die zweite Methode eignet sich auch für unregelmäßige oder unvorhersehbare Änderungszeitpunkte. Hierbei berechnet sich die maximal benötigte Spaltenzahl nach der

maximalen Anzahl von Änderungen in den Werten der zeitabhängigen Variablen. Da die Änderungszeitpunkte a priori nicht festgelegt sind, wird für jede Änderung der Variablen zusätzlich eine Spalte benötigt, die den Zeitpunkt der Änderung angibt. Ändert sich der Wert der Variablen also viermal, werden insgesamt neun Spalten für die Variable benötigt: Eine Spalte für den Ausgangswert der Variablen, vier Spalten für die jeweils neuen Werte der Variablen und vier Spalten, die die Zeitpunkte angeben, an denen sich diese Werte ändern (vgl. Tabelle 2).
Im Beispiel beträgt die maximale Anzahl der Änderungen in der zeitabhängigen Variablen *Rezidiv* vier, da in dieser Studie nicht mehr als vier Rezidive bei einem Patienten diagnostiziert wurden. Die Datentabelle enthält also vier Spalten für die jeweiligen neuen Werte in der Variablen *Rezidiv* (rez1-rez4) und vier Spalten für die dazugehörigen Zeitpunkte der Änderung (reztime1-reztime4, Tabelle 2). Auf eine Spalte für den Ausgangswert konnte verzichtet werden, da dieser für alle Patienten 0 ist.

Tabelle 2: Erforderliche Datenstruktur für die Kodierung der zeitabhängigen Variablen *Rezidiv*

Obs	*Time*	*Verstorb*	*rez1*	*rez2*	*rez3*	*rez4*	*reztime1*	*reztime2*	*reztime3*	*reztime4*
1	462	1	0	0	0	0	462	462	462	462
22	3867	0	1	1	1	1	2284	2284	2284	2284
42	2677	1	1	2	2	2	1855	2434	2434	2434
142	528	1	1	2	3	3	303	487	518	518
15	1528	1	1	2	3	4	579	945	976	1492

Für Patient 1 in diesem Beispiel bleibt der Wert für die Variable *Rezidiv* über die gesamte Beobachtungszeit null, da kein Rezidiv diagnostiziert wurde. Die Zeitpunkte der Änderungen wurden mit der gesamten Überlebenszeit gleichgesetzt. Betrachtet man nun Patient 42, bei dem insgesamt zwei Rezidive diagnostiziert wurden, sieht man, dass das erste Rezidiv nach 1855 Tagen diagnostiziert wurde, das zweite Rezidiv nach 2434 Tagen und dass der Patient nach insgesamt 2677 Tagen verstarb. Die verbleibenden Spalten rez3 und rez4 sowie reztime3 und reztime4 wurden jeweils mit den Werten aus rez2 und reztime2 aufgefüllt.

4.2 Programmcode

Der Code für PROC PHREG wird für die Survival-Analyse mit zeitabhängigen Variablen durch die Kodierung der zeitabhängigen Variablen innerhalb des Codes ergänzt. Der Rest bleibt unverändert, jedoch wird nun die zeitabhängige Variable *Rezidiv* erst innerhalb der Prozedur kodiert:

```
proc phreg data=regression;
class tumlok;
model time * verstorb(0) = advancedt rezidiv tumlok;
     if time>reztime4 then rezidiv=rez4;
     else if time>reztime3 then rezidiv=rez3;
     else if time>reztime2 then rezidiv=rez2;
```

```
        else if time>reztime1 then rezidiv=rez1;
        else rezidiv=0;
run;
```

Nach dem Model-Statement folgen nun einige Zeilen If-Statements, die innerhalb der Prozedur genau wie innerhalb eines Datenschrittes gebraucht werden können. Die Prozedur ist daher nicht auf die Kodierung mit If-Statements begrenzt, es kann auch mit Arrays, mit Select-When-Statements oder mit %include zum Einbeziehen extern abgespeicherter Datenschritte gearbeitet werden.
Im ersten If-Statement sieht man, dass die Variable *Rezidiv* gleich dem Wert aus der Spalte rez4 gesetzt wird, wenn die Überlebenszeit größer ist als die angegebene Zeitspanne in der Spalte reztime4. Gleiches gilt für das zweite If-Statement, das die Überlebenszeit mit dem Wert in Spalte reztime3 vergleicht. Trifft keines der If-Statements zu, wird die zeitabhängige Variable *Rezidiv* gleich null gesetzt, da zu diesem Zeitpunkt noch kein Rezidiv diagnostiziert wurde (vgl. Tabelle 2).
Diese Kodierung erscheint zunächst verwunderlich, da der Diagnosezeitpunkt der Rezidive logischerweise immer vor dem Todeszeitpunkt der Patienten liegen muss. Zum Verständnis muss man sich das zu Grunde liegende Modell (Partial Likelihood) klar machen, welches die Wahrscheinlichkeit des Versterbens eines Patienten in Relation setzt zu der Wahrscheinlichkeit des Versterbens der anderen Patienten, die zu diesem Zeitpunkt noch unter Risiko stehen (am Leben sind). Verstirbt also der i-te Patient zum Zeitpunkt (t+Δt), ist diese Likelihood:

$$4.2 \qquad L_i = \frac{\mathrm{h}_i(\mathrm{t}+\Delta\mathrm{t})}{\mathrm{h}_i(\mathrm{t}+\Delta\mathrm{t}) + \cdots + \mathrm{h}_n(\mathrm{t}+\Delta\mathrm{t})},$$

wobei im Nenner lediglich die Hazards für diejenigen Personen aufgeführt sind, die zum Zeitpunkt t+Δt noch unter Risiko stehen.
Es werden also in den If-Statements die Überlebenszeiten der gerade verstorbenen Patienten mit den Diagnosezeitpunkten der Rezidive der noch lebenden Patienten verglichen.

- Bei der Berechnung der Likelihood für das Versterben von Patient 142 ist der Wert der Variablen *Rezidiv* für diesen Patienten mit 3 entwickelten Rezidiven zum Zeitpunkt seines Versterbens nach 528 Tagen drei. Da jedoch bei keinem der anderen Patienten aus Tabelle 2 innerhalb von 528 Tagen ein Rezidiv diagnostiziert wurde, beträgt der Wert für die Variable *Rezidiv* für alle anderen Patienten null.
- Patient 42 verstarb nach 2677 Tagen. Nach dieser Zeitspanne betragen die Werte der Variablen *Rezidiv* für Patient 22 eins und für Patient 42 selbst zwei. Die Patienten 1, 142 und 15 werden in diesem Fall vernachlässigt, da sie zu diesem Zeitpunkt nicht mehr unter Risiko stehen.

Bei häufigen Änderungen in einer zeitabhängigen Variablen kann es sinnvoll sein, statt der einzelnen If-Statements alternativ Array-Statements zu verwenden, um Zeilen im SAS Code einzusparen und den Code lesbar zu halten:

```
proc phreg data=regression;
class tumlok;
model time * verstorb(0) = advancedt rezidiv tumlok;
     rezidiv=0;
     array reztime(*) reztime1-reztime4;
     array rez(*) rez1-rez4;
     do i=1 to 4;
     if time>reztime[i] then rezidiv=rez[i];
     end;
run;
```

5 Lesen des SAS Outputs

5.1 Ergebnistabelle

Tabelle 3 zeigt einen berechneten Wert von β=1,01 für den Schätzer der zeitabhängigen Variablen *Rezidiv*. Damit erhöht sich das Mortalitätsrisiko bei Erhöhung um ein Rezidiv um exp(1,01)=2,75. Das Mortalitätsrisiko eines Patienten mit einem Rezidiv ist also knapp dreimal höher als das Mortalitätsrisiko eines Patienten mit null Rezidiven. Gleiches gilt für das Risikoverhältnis von Patienten mit zwei Rezidiven zu Patienten mit einem Rezidiv. Möchte man beispielsweise das Risiko für Patienten mit zwei Rezidiven im Vergleich zu Patienten mit null Rezidiven ermitteln, gilt:

$$HR = \exp\left(\beta(x_i - x_j)\right) = \exp\big(\beta(2-0)\big) = \exp(2{,}02) = 7{,}54.$$

Tabelle 3: Ergebnistabelle aus dem SAS Output für das beispielhafte Regressionsmodell

Parameter		*DF*	*Parameter Estimate*	*Standard Error*	*Chi-Square*	*Pr > ChiSq*	*Hazard Ratio*
AdvancedT		1	0.95662	0.13681	48.8916	<.0001	2.603
Rezidiv		1	1.01072	0.09216	120.2828	<.0001	2.748
Tumlok	*su*	1	0.30689	0.14396	4.5441	0.0330	1.359
Tumlok	*un*	1	0.64611	0.20717	9.7269	0.0018	1.908

Weitere Angaben in der Ergebnistabelle umfassen den Standardfehler, Chi-Quadrat-Wert und den p-Wert. Mit der Option *rl* oder *risklimits*= im Model-Statement können gleichzeitig auch Konfidenzintervalle mit ausgegeben werden:

```
model time * verstorb(0) = advancedt rezidiv tumlok / rl;
```

Vergleicht man das Hazard Ratio für die zeitabhängige Variable *Rezidiv* (2,748) mit dem Hazard Ratio für die Variable *Rezidiv* aus dem Modell ohne Zeitabhängigkeit (1,482, Ergebnisse nicht abgebildet), erkennt man den großen Einfluss der zeitabhängi-

gen Analyse auf die Ergebnisse. In diesem Beispiel wird der Effekt der zeitabhängigen Variable *Rezidiv*, die sich erst über den Beobachtungszeitraum entwickelt, in der Analyse mit Vernachlässigung der Zeitabhängigkeit nach unten verfälscht, also unterschätzt.

5.2 Informationen zum Modell

In der ausgegebenen Tabelle mit den Modellinformationen ist in der letzten Zeile das *Ties Handling* angegeben (Tabelle 4). Bindungen (*ties*) bezeichnen identische Überlebenszeiten von verschiedenen Beobachtungen, was zu Schwierigkeiten bei der Berechnung der Partial Likelihood führt. Die Standardeinstellung für derartige Bindungen verwendet die *Breslow*-Methode. Genaue Werte werden mit der *Exact*- oder der *Discrete*-Methode errechnet, wobei die *Exact*-Methode sich für kontinuierlich gemessene Überlebenszeiten eignet. Die *Discrete*-Methode wird für identische Überlebenszeiten angewendet, bei denen angenommen wird, dass sich die Überlebenszeiten nur aufgrund von Messungenauigkeiten entsprechen. Da es sich bei der *Exact*- und der *Discrete*-Methode um genaue Methoden handelt, wird gewöhnlich mehr Rechenzeit benötigt, was bei einer großen Anzahl von Beobachtungen und/oder Bindungen zu einem relativ hohen Zeitaufwand führen kann. Das Ties Handling kann als Option im Model-Statement wie folgt spezifiziert werden:

```
model time * verstorb(0) = advancedt rezidiv tumlok / ties=exact;
```

Tabelle 4: Modellinformationen für das Beispielmodell

Model Information	
Data Set	WORK.REGRESSION
Dependent Variable	Time
Censoring Variable	Verstorb
Censoring Value(s)	0
Ties Handling	BRESLOW

5.3 Class Informationen

Wird in PROC PHREG eine Class-Variable spezifiziert, wird im Output die Kodierungsmatrix für die Dummy-Variablen angegeben (Tabelle 5). Im Beispiel ist die Ausprägung „gl“ mit der Kodierung 0-0 die Referenzkategorie.

Tabelle 5: Matrix für Variablen aus dem Class-Statement

Class Level Information			
Class	*Value*	*Design Variables*	
Tumlok	*gl*	0	0
	su	1	0
	un	0	1

5.4 Modellgütekriterien

In der ersten Zeile von Tabelle 6 sind die -2LogLikelihood-Werte für die Modelle mit und ohne Einflussvariablen angegeben, wobei eine große Differenz dieser beiden Werte zeigt, dass das Modell gut zu den Daten passt. Zusätzlich werden das Akaike Informationskriterium (AIC) sowie das Schwarz-Bayes Kriterium (SBC) aufgelistet. Das Akaike Informationskriterium wird anhand folgender Gleichung berechnet:

$$AIC = -2LogL + 2 \cdot p\,,$$

wobei p die Anzahl der Variablen im Modell darstellt. Da kategoriale Variablen in Dummy-Variablen umkodiert werden, erhöht sich p in Abhängigkeit von der Anzahl der Kategorien der Variablen. Im Beispiel nimmt p den Wert 4 an (1(*AdvancedT*) + 1(*Rezidiv*) + 2(*Tumlok*)), was zu einem AIC von 3066,4 führt. Das Akaike Informationskriterium wird zum Vergleich zwischen Modellen mit verschiedenen Kombinationen von Einflussvariablen herangezogen. Beim Vergleich zwischen Modellen sollte dasjenige mit dem minimalen Wert des AIC ausgewählt werden [2].

Tabelle 6: Modellgütekriterien für das Beispielmodell

Model Fit Statistics		
Criterion	*Without Covariates*	*With Covariates*
-2 LOG L	3201.576	3058.399
AIC	3201.576	3066.399
SBC	3201.576	3080.778

6 Tipps für PROC PHREG

Bei der Anwendung von PROC PHREG mit zeitabhängigen Variablen bei großen Datenmengen werden teilweise lange Rechenzeiten benötigt. Es gibt jedoch einige Möglichkeiten, diese Rechenzeiten zu reduzieren:

- Ties Handling: Die Standardeinstellung (*ties=breslow*) benötigt im Allgemeinen weniger Rechenzeit als die exakten Methoden (*exact, discrete*). Die endgültige Auswahl des Ties Handling ist von den zu analysierenden Daten abhängig. Allerdings kann für einen schnelleren Überblick über die Daten die Breslow-Methode nützlich sein.
- If-Statements: In der Proportional Hazards Regression wird die Partial Likelihood des Eintritts des Ereignisses von Interesse bei jedem einzelnen Event berechnet. Da die Werte der zeitabhängigen Variablen sich an jedem dieser Zeitpunkte unterscheiden können, werden die Werte dieser Variablen bei jedem Event neu berechnet, was zu einer sehr langen Rechenzeit führen kann. Daher empfiehlt es sich, anstelle von If-if-Statements If-else if- oder Select-when-Statements zu verwenden. Außerdem sollten die häufigsten Ausprägungen zuerst aufgelistet werden, da somit ein schneller Ausstieg aus der If- bzw. Select-Bedingung erfolgt. Trifft also beispielsweise das erste If-Statement auf 60% der Beobachtungen zu, müssen die darauffolgenden Else-if-Statements für diese gar nicht mehr durchlaufen werden. Gleiches gilt bei der Wahl von Select-when-Statements.
- Proc Surveyselect: Da bei sehr großen Datenmengen auch die vorher genannten Tipps die Rechenzeit nicht auf ein angenehmes Maß reduzieren können (vor allem wenn eine Vielzahl von Modellen miteinander verglichen werden soll), gibt es die Möglichkeit, mit der Prozedur Surveyselect eine Zufallsstichprobe aus der Datenmenge zu ziehen:

  ```
  proc surveyselect data=regression method=srs n=2000
  out=regression2; run;
  ```

  In diesem Beispiel wird aus einer Datenmenge von knapp 25.000 Beobachtungen eine Stichprobe von 2.000 Beobachtungen gezogen, um verschiedene Kombinationen von Einflussvariablen zu testen.

Meistens sollen die Ergebnisse aus der Survival-Analyse veröffentlicht und in Tabellen dargestellt werden. Da SAS aber keine publikationsfertigen Ergebnistabellen liefert, ist es nötig, einige einfache aber effektive Nachbearbeitungen unter Verwendung von SAS ODS vorzunehmen, um aus der ausgegebenen Ergebnistabelle eine publikationsfertige Tabelle zu erstellen (vgl. Tabellen 7a und 7b). Eine Anleitung zum Erstellen solcher publikationsfertiger Ergebnistabellen findet sich bei Ramroth, 2008 [3].

Tabelle 7 a): Ergebnisstabelle aus dem SAS Output inklusive Konfidenzintervalle

Parameter		*DF*	*Parameter Estimate*	*Standard Error*	*Chi-Square*	*Pr > ChiSq*	*Hazard Ratio*	*95% Hazard Ratio Confidence Limits*		*Label*
AdvancedT		1	0.95702	0.13682	48.9251	<.0001	2.604	1.991	3.405	
Rezidiv		1	1.01087	0.09215	120.3250	<.0001	2.748	2.294	3.292	
Tumlok	*su*	1	0.30664	0.14397	4.5366	0.0332	1.359	1.025	1.802	Tumlok su
Tumlok	*un*	1	0.64578	0.20717	9.7165	0.0018	1.907	1.271	2.863	Tumlok un

Tabelle 7 b): Publikationsfertige Ergebnistabelle

Charakteristik	*Kategorie*	*Hazard Ratio*	*95% Konfidenzintervall*
Fortg. T Stadium		2.6	(2.0, 3.4)
Rezidiv		2.7	(2.3, 3.3)
Tumorlokalisation	glottisch	1	
Tumorlokalisation	supraglottisch	1.4	(1.0, 1.8)
Tumorlokalisation	unbekannt	1.9	(1.3, 2.9)

Literatur

[1] Allison Paul D: Survival analysis using the SAS system. A practical guide; SAS Institute: Cary, NC, 1995

[2] Kleinbaum David G. & Klein Mitchel: Survival Analysis. A Self-Learning Text; Springer: USA, 2005 (2.Ed)

[3] Ramroth H: Publikationsfertige Kombination von Häufigkeiten und Risiko-Kennwerten aus Ergebnissen von klinisch-epidemiologischen Studien; Proceedings der 12. KSFE; Aachen, 2008.