

EIN AUSWERTUNGSTOOL FÜR DIE EXPLORATIVE ANALYSE VON ÜBERLEBENSZEITEN

Hans-Peter Altenburg und Thomas Floren

Universität Heidelberg, Fakultät für Klinische Medizin Mannheim
Medizinische Statistik, Biomathematik und Informationsverarbeitung
Theodor-Kutzer-Ufer, D68135 Mannheim,
E-mail: km05@rumms.uni-mannheim.de

Zusammenfassung

Viele Studien der klinischen und experimentellen Forschung haben als Zielkriterium die Überlebenszeit oder ganz allgemein die Zeit bis ein vordefiniertes Ereignis eintritt. Ein Problem ist dabei, daß das Merkmal Zeitdauer nicht exakt beobachtet wird, sondern die Daten zensiert sind. Das mathematische Hauptzielkriterium bei solchen Studien ist oft die Überlebensfunktion, die im einfachsten Fall unzensierter Daten sich als $1 - F(t)$ schreiben läßt, wenn $F(t)$ die Verteilungsfunktion der Zielgröße bezeichnet. Der Vortrag beschreibt die Entwicklung eines explorativen Auswertungstools für die Visualisierung und Analyse von Studien mit Zielgröße Überlebenszeit nach verschiedenen Zielsetzungen. Die dabei verwendeten Methoden lassen sich in zwei große Klassen einteilen: *parametrische Methoden* und *nichtparametrische Methoden*. Parametrische Methoden lassen sich mit Hilfe der SAS-Prozedur LIFEREG analysieren. Nichtparametrische Analysen kann man mit Hilfe der Prozeduren LIFETEST (Produkt-Limit-Schätzer der Überlebensfunktion) oder PHREG (Proportional-Hazard / Cox-Regressions-Modell) durchführen, wobei die letztere Prozedur auch die Realisierung komplexerer Analyseverfahren und Strategien zur Modellbildung bzw. Selektion relevanter prognostischer Faktoren erlaubt. Ziel des Projektes war, einem mit SAS unerfahrenen Benutzer zu helfen, Studienergebnisse zu visualisieren, notwendige Vergleiche durchzuführen, prognostische Faktoren zu ermitteln, und dabei eine graphische Benutzeroberfläche für die Handhabung der in einer Analyse erforderlichen SAS-Prozeduren zur Verfügung zu stellen. Die Realisierung dieser Ziele erfolgt mit Hilfe des SAS-Produktes SAS/AF.

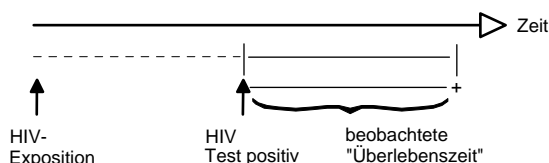
Einleitung

Überlebenszeit-Analyse ist die vereinfachte Bezeichnung für eine Reihe populärer statistischer Verfahren zur Analyse positiv-wertiger Zufallsgrößen. Meist wird dabei eine Zielgröße betrachtet, die die Zeit bis zum Eintritt eines vordefinierten Ereignisses betrachtet. Vordefinierte Ereignisse sind z.B.: Tod eines Individuums, Auftreten einer Erkrankung, Remission einer Tumorerkrankung, Progression, Gesundung, Rückfall, Entlassung aus dem Krankenhaus, Auftreten von Metastasen, usw. In der Regel wird nur **ein** solches **Ereignis** untersucht. Bei mehreren Ereignissen spricht man von konkurrierenden Risiken, was aber hier nicht weiter behandelt werden soll.

Beispiel: Wir betrachten eine Studie mit Leukämie-Patienten. Die Hauptzielvariable sei das Merkmal Tumorremission, und es soll untersucht werden, wie lange die Patienten im Zustand der Remission verbleiben. In der Regel beträgt die Beobachtungsdauer mehrere Wochen nach Behandlungsbeginn.

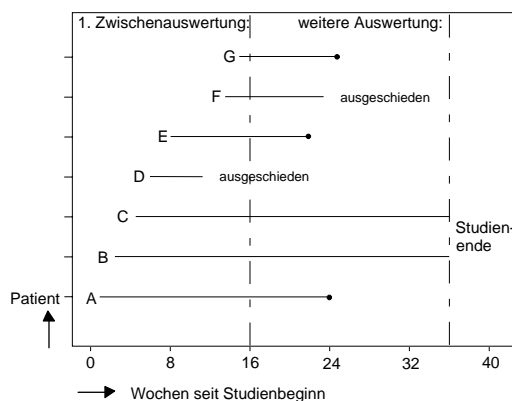
Ein mehr oder weniger großes Problem stellt die Tatsache dar, daß die beobachteten Daten *zensiert* sind, d.h. die exakte (Überlebens-) Zeit ist nicht immer genau bekannt, sondern nach links oder rechts "abgeschnitten". Die Gründe für eine Zensierung sind sehr vielschichtig, etwa, die Studie endet bevor das definierte Ereignis bei dem Probanden eingetreten ist, der Proband ist verloren gegangen oder der Proband ist aus der Studie ausgeschieden (z.B. weil er verstorben ist aber "Tod" nicht als die eigentliche Zielgröße vorgesehen war bzw. eine "Adverse Drug Reaction" eingetreten ist). Man unterscheidet *Links-* oder *Rechts-Zensierung*, wobei wir uns im folgenden auf rechtszensierte Daten beschränken wollen.

Beispiel für eine Linkszensierung:



Beispiele für eine Rechtszensierung:

Beobachtbare Überlebenszeiten in einer Follow-up Studie



Mathematische Terminologie

Aus mathematischer Sicht wird eine Zufallsvariable T mit Wertebereich $T > 0$ betrachtet. Mit t bezeichnet man eine spezielle Realisierung dieser Zufallsvariable. $F(t) = P(T \leq t)$ sei die zugehörige Verteilungsfunktion und $f(t)$ die Dichtefunktion. Die *Überlebensfunktion* ist dann gegeben durch $S(t) = P(T > t) = 1 - F(t)$. Neben der Überlebensfunktion wird oft auch die *Hazardrate* oder *Hazardfunktion* (eine Art bedingte Fehlerrate) zur Analyse von Überlebenszeitstudien herangezogen. Sie ist gegeben durch $h(t) = \lim_{\Delta t \rightarrow 0} P(t \leq T \leq t + \Delta t \mid T \geq t) / \Delta t$, oder $h(t) = -S'(t)/S(t) = -f(t)/(1-F(t))$, was zu der wichtigen Beziehung $S(t) = \exp(\int_0^t h(u) du)$ führt. Im Fall einer konkreten Stichprobe sind all diese Größen durch ihre entsprechenden empirischen Größen zu ersetzen. Die Zensierung der beobachteten Daten kompliziert die mathematische Behandlung dieser Darstellung.

Ziele einer Überlebenszeitanalyse

Die wichtigsten und am häufigsten vorkommenden Ziele einer Überlebenszeitanalyse sind:

- (1) Schätzen und Interpretieren von Überlebens- und / oder Hazardfunktion auf Grund der Daten einer konkreten Studie mit zensierten Daten,
- (2) Vergleichen von Überlebens- und / oder Hazardfunktionen bei verschiedenen Stratifizierungsgruppen, (z.B. verschiedene Behandlungsgruppen) und
- (3) Nachweis / Aufstellen von Beziehungen zwischen erklärenden Variablen (Kovariablen) und der Zielgröße (Überlebenszeit), z.B. Anwendung des Proportional Hazard / Cox-Modells.

Die hierbei verwendeten Verfahren werden in zwei große Klassen eingeteilt: *parametrische* und *nichtparametrische* Verfahren. Im ersten Hauptmenu des Auswertungstools wird deshalb außer nach den allgemeinen Angaben zu Datei, Variablen, Zensierung, und Stratifizierung auch nach der gewünschten Analyseart gefragt. Das folgende Bild (Abb.1) zeigt dieses Hauptmenu. Es wird dabei vorausgesetzt, daß der Benutzer seine Daten bereits ins SAS-System in einer SAS-Datenmenge abgespeichert hat.

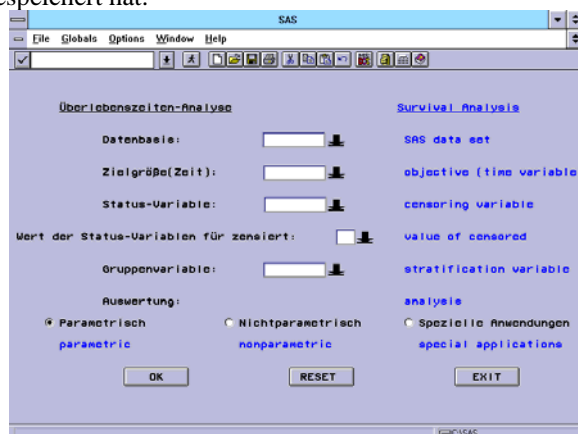


Abb.1.: Hauptmenü

Bei den *parametrischen Methoden* erfolgt im wesentlichen die Anpassung der beobachteten Daten an ein konkretes parametrisches Modell, wobei bei den o.g. Zielen dann die entsprechenden Parameter und / oder Überlebens- bzw. Hazardfunktionen verglichen sowie der Einfluß von Kovariablen auf die Überlebenszeit beurteilt werden kann. Ist T_0 die Zeit ohne den Einfluß von Kovariablen X_1, \dots, X_n , so wird die Überlebensfunktion durch $P(T > t \mid X_1, \dots, X_n) =$

$P(T_0 > \exp(-\sum_i \beta_i X_i) t)$ modelliert. Drei wichtige und in vielen Anwendungen modellierte Verteilungen für die Baseline-Verteilung seien hier als Beispiel genannt:

	Exponentialverteilung	Weibullverteilung	Logistische Verteilung
Parameter:	λ	λ, c	a, k
Verteilungsfunktion:	$F(t) = 1 - \exp(-\lambda t)$	$F(t) = 1 - \exp(-(\lambda t)^c)$	$F(t) = 1 - 1/[1 + \exp((t-a)/k)]$
Dichtefunktion:	$f(t) = \lambda \exp(-\lambda t)$	$f(t) = c\lambda^c t^{c-1} \exp(-(\lambda t)^c)$	$f(t) = \exp((t-a)/k) / [k(1 + \exp((t-a)/k))^2]$
Überlebensfunktion:	$S(t) = \exp(-\lambda t)$	$S(t) = \exp(-(\lambda t)^c)$	$S(t) = 1/[1 + \exp((t-a)/k)]$
Hazardfunktion:	$h(t) = \lambda$	$h(t) = c\lambda t^{c-1}$	$h(t) = \exp((t-a)/k) / [k(1 + \exp((t-a)/k))]$

Das SAS-System hat in der Prozedur LIFEREG die parametrische Analyse von Überlebenszeiten realisiert, wobei neben den o.g. Verteilungen noch weitere (wie z.B. die Normal- oder logarithmische Normalverteilung, Gamma- und log-logistische Verteilung) angepaßt werden können. Nachdem man in obigem Hauptmenu die entsprechenden Angaben gemacht hat, liefert der Auswertungstool ein weiteres Menu in dem weitere Optionen für die parametrische Analyse angekreuzt werden können, z.B. der zu modellierende Verteilungstyp, maximale Iterationszahl, Konfidenzintervalle, ins Modell einbezogene Kovariablen sowie ganz bestimmte Ausprägungen dieser Kovariablen für eine konkrete Vorhersage. Über ein implementiertes SAS-Macro erfolgt danach die Ausgabe des zugehörigen Schaubildes.

Bei den *nichtparametrischen Methoden* werden i.w. drei Verfahren ermöglicht:

- Life-Table-Methode
- Kaplan-Meier-Produkt-Limit-Schätzer
- Cox-Proportional-Hazard-Modell

Die *Life-Table-* oder *Sterbetafel-Methode* zählt zu den klassischen Verfahren. Man kann mit ihr nur ungruppierte Beobachtungen analysieren, und sie spielt in den meisten klinischen Anwendungen kaum eine Rolle. Ungünstig ist bei ihr, daß alle beobachteten Zeiten in ein Intervallraster gezwungen werden, wobei in den meisten Fällen sicher Informationen verloren gehen.

Besser ist meist der *Kaplan-Meier-Schätzer (Product-Limit-Verfahren)*, welcher eine rekursive Bestimmung der Überlebensfunktion erlaubt. Der Wert der Überlebensfunktion zum Zeitpunkt $t_{(j)}$ ist gegeben durch: $S(t_{(j)}) = S(t_{(j-1)}) \times P(T > t_{(j)} | T > t_{(j-1)})$ oder $S(t) = \prod_{i < t} (1 - (d_i / n_i))$, wobei d_i die Anzahl Verstorbener zum Zeitpunkt t_i und n_i die Anzahl Lebender zur Zeit t_i^- (d.h. bis unmittelbar vor dem Zeitpunkt t_i) ist. Die Hazardrate zum Zeitpunkt t_i ist $\lambda_i = d_i / n_i$. Der Vorteil des Kaplan-Meier-Schätzers liegt darin, daß alle in den Beobachtungen vorhandenen Informationen bei der Auswertung berücksichtigt werden. Allerdings erlaubt es nicht die Einbeziehung von Kovariablen wie beim parametrischen Modell, sondern nur die stratifizierte deskriptive Beschreibung der beobachteten Überlebenszeiten. Wenn in einer Stichprobe keine zensierten Beobachtungen vorkommen, ist die Kaplan-Meier-Kurve nichts anderes als eine normale empirische Überlebensfunktion.

Das *Cox-Proportional-Hazard-Modell* dagegen berücksichtigt analog zum parametrischen Regressionsmodell auch Kovariablen. Der Modellansatz für die Einbeziehung der Kovariablen erfolgt über die Hazardfunktion. Anstelle der Standardhazardfunktion $h(t) = -S'(t)/S(t)$ lautet die Proportional-Hazard-Funktion: $h(t, x) = h_0(t) \exp(\sum_i \beta_i x_i) = h_0(t) \exp(x' \beta)$, wobei $h_0(t)$ die Baseline-Hazardfunktion bezeichnet, die diejenige Hazardfunktion darstellt, die bei einem Individuum oder einer Gruppe mit allen Kovariablen null vorkommt, und x_i sind Realisierungen der Kovariablen X_i . Die β_i sind wieder die den Kovariablen X_i zugeordneten Regressionskoeffizienten. Alle Kovariablen werden als zeitunabhängig angesehen. Die Überlebensfunktion ist gegeben durch $S(t, x) = [S_0(t)]^{\exp(x' \beta)}$ mit $S_0(t) = \exp(-\int_0^t h_0(u) du)$. Die Vorteile des Cox-Modelles liegen einmal darin, daß es bei Vorliegen eines korrekten parametrischen Modelles dieses auch schon gut approximiert und, wenn man im Zweifel ist, es bereits ganz gute Schätzergebnisse liefert. Die Effekte der Kovariablen lassen sich hierbei auch ohne die Kenntnis der Baselinefunktion schätzen. Außerdem erlaubt es die Adjustierung von Überlebenskurven.

Unter dem Menüpunkt *spezielle Anwendungen* im Hauptmenu (Abb.1) sind zwei Spezialfälle realisiert, die in einigen Anwendungen anstelle der o.g. Standardverfahren verwendet werden können. Es sind dies einmal die simultanen Konfidenzbänder für Kaplan-Meier-Kurven bestimmt nach der Hall-Wellner-Methode oder nach Efron's Bootstrap-Verfahren. Es wurden dabei die SAS-Macros aus dem Buch von Harris and Albert [1] verwendet. Die zweite spezielle Anwendung ist die Bestimmung von Überlebenskurven bei intervall-zensierten Beobachtungen. Intervall-zensierte Beobachtungen sind gar nicht so selten zu beobachten, da man ja oft die Probanden nur in bestimmten Zeitintervallen beobachtet, wie etwa die quartalsweise Nachbeobachtung von Tumorpatienten. Hierfür wurde ein SAS-Macro auf der Basis eines Beispiels im SAS/IML Ergänzungshandbuch Version 6.11 [7] erstellt.

Realisierung im SAS-System

Für die aufgelisteten statistischen Verfahren wurden folgende SAS-Prozeduren aus SAS/STAT verwendet:

- parametrisch die SAS-Prozedur LIFEREG, die die Berücksichtigung folgender parametrischen Baseline-Verteilungsfunktionen erlaubt:

Weibull-	Normal-	logistische	Gamma-
Exponential-	logarithm. Normal-	log-logistische	

- nicht parametrisch die SAS-Prozeduren LIFETEST und PHREG. Mit Hilfe der Prozedur LIFETEST können Life-Tables und die Kaplan-Meier-Schätzer bestimmt, mit PHREG das Cox'sche Proportional-Hazard-Modell zur Einbeziehung von Kovariablen verwendet werden.
- eigene SAS-MACROS werden benutzt zur Bildung von Teilgruppen, für die graphische Ausgabe und das Schätzen ganz bestimmter Variablen- und / oder Parameterkonstellationen bei der Einbeziehung von Kovariablen.

Die Analyse erfolgt menügesteuert, wobei alle Aktionen, wie z.B. Auswahl von SAS-Datenmenge, Variablen, Art der zu bestimmenden Kurve, Teilgruppenauswahl, Ergebnisfenster bei Gruppenvergleichen, etc. durch Anklicken mit der Maus und ohne die Angabe von Programmstatements erreicht werden. Folgende SAS-Produkte wurden verwendet: SAS/AF, FRAME, SCL. Die folgenden Abbildungen zeigen eine Auswahl der möglichen Fenster.

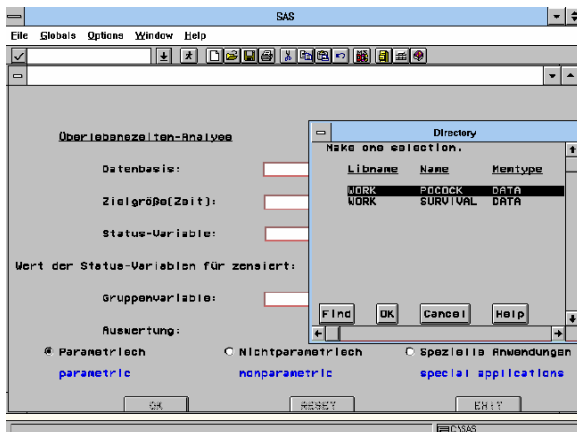


Abb.2: Auswahl der Datenmenge

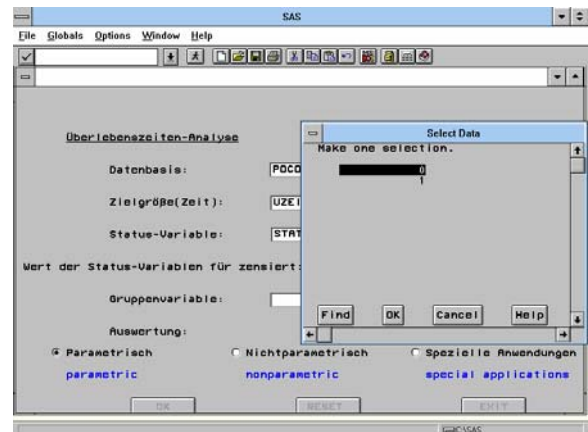


Abb.3: Eingabe der Verschlüsselung für die zensierten Ausprägungen

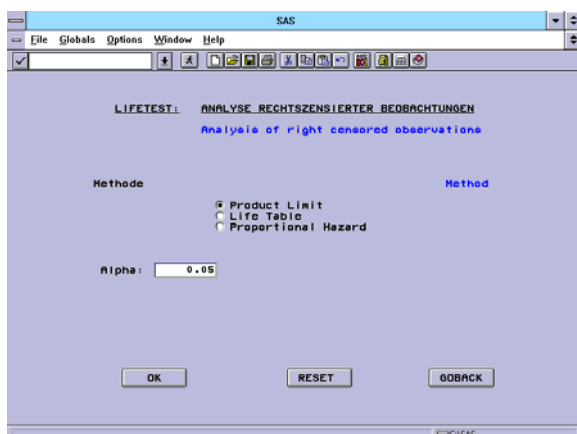


Abb.4: Auswahl der nichtparametrischen Auswertungsmethode

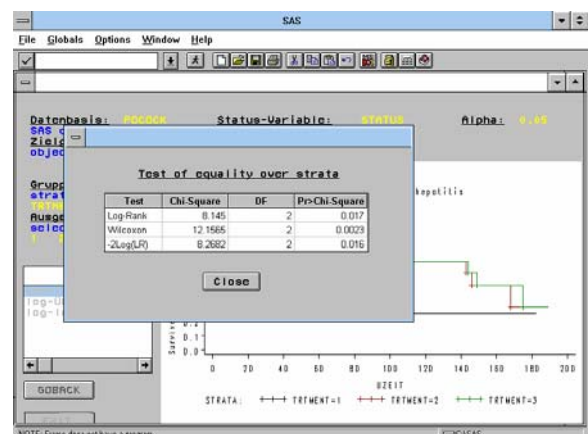


Abb. 5: Anzeigen der Teststatistik

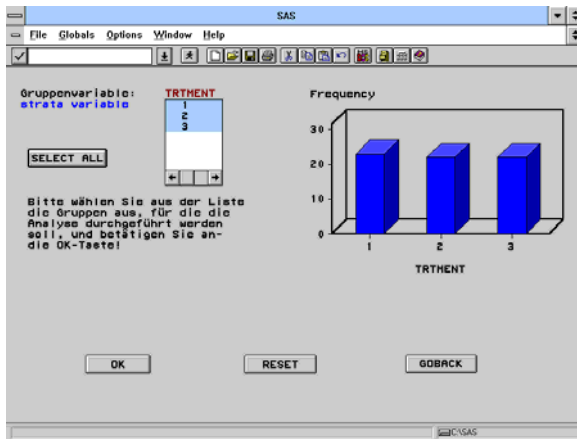


Abb.6:

Auswahl aller Gruppen und graphische Darstellung der zugehörigen Kaplan-Meier-Kurven

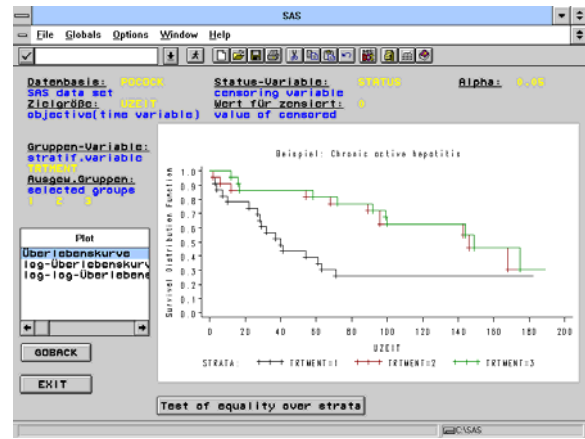


Abb.7:

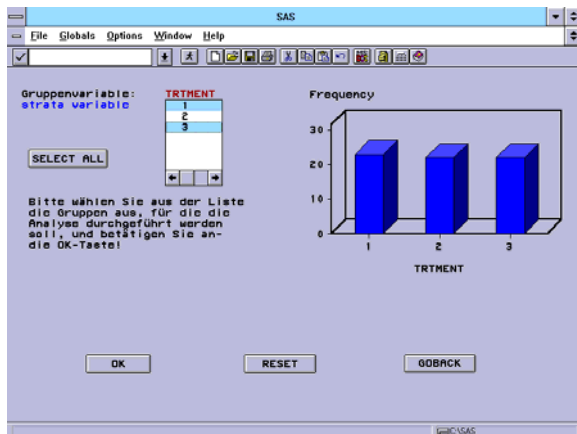


Abb.8:

Teilauswahl von nur zwei Behandlungsgruppen und Darstellung der entsprechenden Überlebenskurven

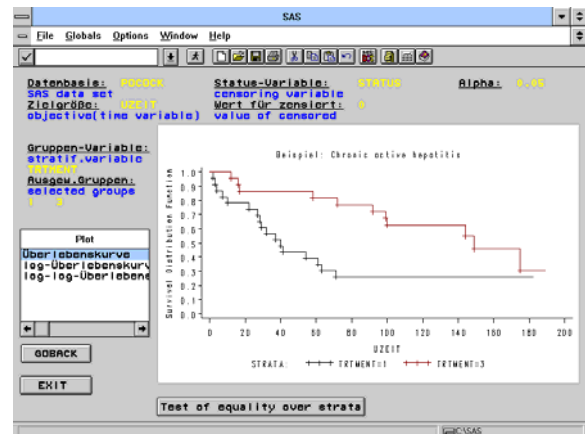


Abb.9:

Bei der Erstellung des Anwendungstool sind uns folgende Punkte **positiv** aufgefallen:

- die grafische Oberfläche zur FRAME-Entwicklung ist schnell erlernbar, komfortabel und einfach bedienbar,
- die Objektorientierung der SCL erlaubt eine schnelle Programmentwicklung für die zuvor erstellten Frames (Abbildung der Schnittstellenelemente, Eigenschaften / Verhalten → Objekte),
- schnelle Compilierung und Austestung
- Compilierungsfehler werden im Messagefenster angegeben (mit Zeilen-Nr. und Hinweis bzw. Erklärung),
- Laufzeitfehler werden im LOG-Fenster angegeben,
- es existieren eine Reihe Benutzerschnittstellenelementen in SAS 6.11 wie z.B. Buttons, Listboxes, etc.

Etwas gewöhnungsbedürftig war die Tatsache, daß beim Schliessen der FRAMES und Programmfenster die Inhalte ohne Nachfrage abgespeichert werden. Dies war besonders dann störend, wenn ein veraltetes Fenster (mit gleichem Namen) noch geöffnet war, da dann das alte Fenster wieder abgespeichert wurde und alle durchgeführten Änderungen unberücksichtigt blieben. Die Online-Hilfe und Handbücher waren manchmal unübersichtlich organisiert. Spezielle Probleme erforderten dann langes Suchen und Nachschlagen an verschiedenen Stellen und / oder in verschiedenen Handbüchern.

Literatur

- 1) Cox, D.R. and Oakes, D.: *Analysis of Survival Data*. Chapman and Hall 1984
- 2) Harris, E.K. and Albert, A.: *Survivorship Analysis for Clinical Studies*. M. Dekker, New York 1991
- 3) Kleinbaum, D.G.: *Survival Analysis - A Self-Learning Text*. Springer, New York 1996
- 4) Marubini, E. and Valsecchi, M.G.: *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley, Chichester 1995
- 5) McNeil, D.: *Epidemiological Research Methods*. Wiley, New York 1996
- 6) Miller, R.G.: *Survival Analysis*. Wiley, New York, 1981
- 7) SAS/IML Software Changes and Enhancements Through Release 6.11. SAS Institute 1996