

SAS im Einsatz am Institut für Klinische Biometrie der MedUni Wien und unser Beitrag zur Weiterentwicklung von SAS 9.4

Michael Schemper

Institut für Klinische Biometrie

Zentrum für Med. Statistik, Informatik und Intelligente Systeme

Medizinische Universität Wien

- **Klinische Biometrie** ist DIE Methodik empirischer klinischer Forschung, beinhaltend:
 - Planung von empirischen Forschungsvorhaben
 - Statistische Analyse von Studiendaten
(Quantifizierung der Größe von Effekten, sowie deren Verallgemeinerbarkeit)
- Das **Institut für Klinische Biometrie** wendet im Rahmen von Kooperationen biostatistische Methoden an und entwickelt diese Methoden weiter.

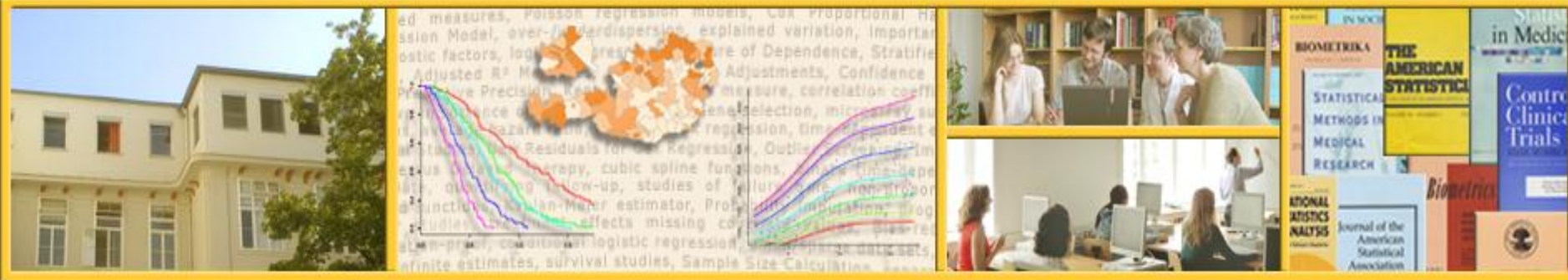
ZENTRUM FÜR MEDIZINISCHE STATISTIK, INFORMATIK UND INTELLIGENTE SYSTEME (CeMSIIS)

Institut für Klinische Biometrie (KB)

» www.meduniwien.ac.at/cemsiis/kb/



HOME ORGANISATION WISSENSCHAFT & FORSCHUNG STUDIUM & LEHRE SERVICES KONTAKT



Willkommen am Institut für Klinische Biometrie

Klinische Biometrie ist die Methodik empirischer klinischer Forschung.

Das seit 1991 bestehende Institut wendet im Rahmen klinischer Kooperationen biostatistische Methoden an und entwickelt diese Methoden weiter.

Für die Anbahnung von klinischen Kooperationen stehen regelmäßig Beratungsstunden zur Verfügung.

SCHNELLINFO

- Kontakt
- Sitemap
- Intranet (Login)
- A K I M
- Randomizer
- Statistische Software
- Statistische Beratung

Wichtigste Gebiete statistisch-methodischer Forschung:

- Lebensdaueranalyse
- Maße der Prognosegenauigkeit und erklärten Variabilität von Regressionsmodellen
- Pönalisierung in Regressionsmodellen
- Analyse kleiner Stichproben (exakte Verfahren)
- Analyse hochdimensionaler (Omics-) Daten

Auf dieser Seite finden Sie von unserem Institut entwickelte Computerprogramme und zugehörige Literaturverweise.

- **PSHREG** - Proportional and non-proportional subdistribution hazards regression
- **t.opt** - Adaptive trimmed t-statistics for microarrays
- **concreg** - Concordance regression
- **FL/logistf** - Logistic regression using Firth's bias reduction: a solution to the problem of separation in logistic regression
- **WCM/coxphw** - Weighted estimation in Cox regression
- **FC/coxphf** - Cox regression using Firth's bias reduction (monotone likelihood)
- **COMPASS** - Comparing parametric survival models
- **RELIMP** - Comparing the importance of prognostic factors in Cox and logistic regression using SAS
- **XLINRANK** - Exact logrank tests for unequal follow-up
- **SURES** - Log-odds and normal deviate residuals for Cox regression
- **R2POI** - Calculating adjusted R^2 measures for Poisson regression models
- **EVLOGIST** - Computing measures of explained variation for logistic regression models
- **SUREV/surev** - Predictive accuracy and explained variation in Cox regression
- **MULTIMP** - Estimation of correlation and explained variation in the presence of censoring assuming an uncensored population
- **RCS** - Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions
- Vienna prediction model for recurrent VTE

SAS im Einsatz am Institut für Klinische Biometrie:

1. im Rahmen medizinischer Studien:

- lange Vorgeschichte seit 1980
- SAS ist Basis der Biostatistik für alle klinischen Projekte
- SAS bedeutet für uns hohe Effizienz und Verlässlichkeit bei den statistischen Analysen für unsere klinischen Kooperationspartner
- von Anbeginn hat SAS in der Verständlichkeit der Manuale sowie der weiterführenden SAS-Literatur ganz große Maßstäbe gesetzt
- meist kommen bei uns die Prozeduren und Data-Management Funktionen des Base SAS zur Anwendung, sowie die Prozeduren von SAS/STAT, SAS/GRAPH und SAS/IML

SAS im Einsatz am Institut für Klinische Biometrie (2):

2. im Rahmen eigener methodischer Forschung:

- für fast alle neu entwickelten Verfahren haben wir SAS Makros entwickelt und bieten diese zum kostenlosen Download an.
- für wiederkehrende Fragestellungen, für die SAS noch keine Verfahren anbietet, verwenden wir ebenfalls eigene SAS Makros
- Im Bereich der Programmierung komplexer statistischer Berechnungsverfahren bzw. Algorithmen haben wir früher FORTRAN (mit NAG bzw. IMSL Subroutines) verwendet, nun vermehrt R.

SAS im Einsatz am Institut für Klinische Biometrie (3):

FACIT

Die Bedeutung von SAS für unser Institut ist über Jahrzehnte konstant geblieben, und dies wird sehr wahrscheinlich auch für das nächste Jahrzehnt gelten.

Aus dem Inhaltsverzeichnis zum SAS/STAT Manual der Version 9.4:

5972 ♦ Chapter 72: The PHREG Procedure

	Proportional Subdistribution Hazards Model for Competing-Risks Data	6054
	Hazard Ratios	6058
	Proportional Rates/Mean Models for Recurrent Events	6060
	Newton-Raphson Method	6062
→	Firth's Modification for Maximum Likelihood Estimation	6062
	Robust Sandwich Variance Estimate	6064
	Testing the Global Null Hypothesis	6064
	Type 3 Tests	6065
	Confidence Limits for a Hazard Ratio	6066
	Using the TEST Statement to Test Linear Hypotheses	6068
	Analysis of Multivariate Failure Time Data	6069
	Model Fit Statistics	6078
→	Schemper-Henderson Predictive Measure	6079
	Residuals	6080
	Diagnostics Based on Weighted Residuals	6082
	Influence of Observations on Overall Fit of the Model	6083
	Survivor Function Estimators	6084

PROC PHREG Statement

PROC PHREG < *options* > ;

The PROC PHREG statement invokes the PHREG procedure. Table 72.1 summarizes the *options* available in the PROC PHREG statement.

Table 72.1 PROC PHREG Statement Options

Option	Description
ALPHA=	Specifies the level of significance
ATRISK	Displays a table that contains the number of units and the corresponding number of events in the risk sets
COVM	Uses the model-based covariance matrix in the analysis
COVOUT	Adds the estimated covariance matrix to the OUTEST= data set
COVSANDWICH	Requests the robust sandwich estimate for the covariance matrix
DATA=	Names the SAS data set to be analyzed
→ EV	Requests the Schemper-Henderson predictive measures
INEST=	Names the SAS data set that contains initial estimates
MULTIPASS	Recompiles the risk sets

Schemper-Henderson Predictive Measure

Measures of predictive accuracy of regression models quantify the extent to which covariates determine an individual outcome. Schemper and Henderson's (2000) proposed predictive accuracy measure is defined as the difference between individual processes and the fitted survivor function.

For the i th individual ($1 \leq i \leq n$), let l_i , X_i , Δ_i , and \mathbf{Z}_i be the left-truncation time, observed time, event indicator (1 for death and 0 for censored), and covariate vector, respectively. If there is no delay entry, $l_i = 0$. Let $t_{(1)} < \dots < t_{(m)}$ be m distinct event times with d_j deaths at $t_{(j)}$. The survival process $Y_i(t)$ for the i th individual is

$$Y_i(t) = \begin{cases} 1 & l_i \leq t < X_i \\ 0 & t \geq X_i \text{ and } \Delta_i = 1 \\ \text{undefined} & t \geq X_i \text{ and } \Delta_i = 0 \end{cases}$$

Let $\hat{S}(t)$ be the Kaplan-Meier estimate of the survivor function (assuming no covariates). Let $\hat{S}(t|\mathbf{Z})$ be the fitted survivor function with covariates \mathbf{Z} , and if you specify TIES=EFRON, then $\hat{S}(t|\mathbf{Z})$ is computed by the Efron method; otherwise, the Breslow estimate is used.

The predictive accuracy is defined as the difference between individual survival processes $Y_i(t)$ and the fitted survivor functions with ($\hat{S}(t|\mathbf{Z}_i)$) or without ($\hat{S}(t)$) covariates between 0 and τ , the largest observed time. For each death time $t_{(j)}$, define a mean absolute distance between the $Y_i(t)$ and the $\hat{S}(t)$ as

$$\begin{aligned} \hat{M}(t_{(j)}) = & \frac{1}{n_j} \sum_{i=1}^n I(l_i \leq t_{(j)}) \left\{ I(X_i > t_{(j)} \geq l_i)(1 - \hat{S}(t_{(j)})) + \Delta_i I(X_i \leq t_{(j)})\hat{S}(t_{(j)}) \right. \\ & \left. + (1 - \Delta_i)I(X_i \leq t_{(j)}) \left[(1 - \hat{S}(t_{(j)}))\frac{\hat{S}(t_{(j)})}{\hat{S}(X_i)} + \hat{S}(t_{(j)}) \left(1 - \frac{\hat{S}(t_{(j)})}{\hat{S}(X_i)} \right) \right] \right\} \end{aligned}$$

where $n_j = \sum_{i=1}^n I(l_i \leq t_{(j)})$. Let $\hat{M}(t_{(j)}|\mathbf{Z})$ be defined similarly to $\hat{M}(t_{(j)})$, but with $\hat{S}(t_{(j)})$ replaced by $\hat{S}(t_{(j)}|\mathbf{Z}_i)$ and $\hat{S}(X_i)$ replaced by $\hat{S}(X_i|\mathbf{Z}_i)$. Let $\hat{G}(t)$ be the Kaplan-Meier estimate of the censoring or potential follow-up distribution, and let

$$w = \sum_{j=1}^m \frac{d_j}{\hat{G}(t_{(j)})}$$

The overall estimator of the predictive accuracy with (\hat{D}_z) and without (\hat{D}) covariates are weighted averages of $\hat{M}(t_{(j)}|\mathbf{Z})$ and $\hat{M}(t_{(j)})$, respectively, given by

$$\begin{aligned}\hat{D}_z &= \frac{1}{w} \sum_{j=1}^m \frac{d_j}{\hat{G}(t_{(j)})} \hat{M}(t_{(j)}|\mathbf{Z}) \\ \hat{D} &= \frac{1}{w} \sum_{j=1}^m \frac{d_j}{\hat{G}(t_{(j)})} \hat{M}(t_{(j)})\end{aligned}$$

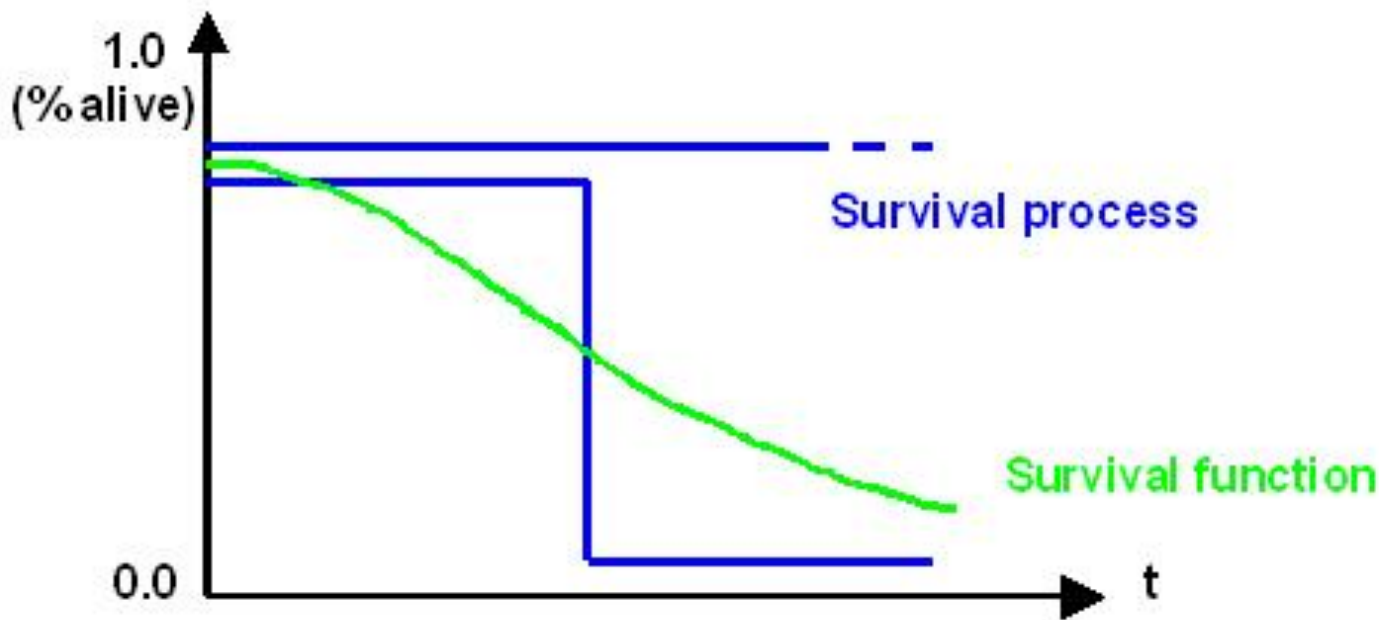
The explained variation by the Cox regression is

$$V = 100 \left(1 - \frac{\hat{D}_z}{\hat{D}} \right) \%$$

Because the predictive accuracy measures \hat{D}_z and \hat{D} are based on differences between individual survival processes and fitted survivor functions, a smaller value indicates a better prediction. For this reason, \hat{D}_z and \hat{D} are also referred to as predictive inaccuracy measures.

Survival Process versus Survival Function

„beobachtet“ vs. „erwartet“:



References

- Heinze, G. (1999), *The Application of Firth's Procedure to Cox and Logistic Regression*, Technical Report 10/1999, updated in January 2001, Section of Clinical Biometrics, Department of Medical Computer Sciences, University of Vienna.
- Heinze, G. and Schemper, M. (2001), "A Solution to the Problem of Monotone Likelihood in Cox Regression," *Biometrics*, 51, 114–119.
- Hosmer, D. W., Jr. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley & Sons.
- Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001), *Bayesian Survival Analysis*, New York: Springer-Verlag.
- Kalbfleisch, J. D. and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley & Sons.
- Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981), "On the Regression Analysis of Multivariate Failure Time Data," *Biometrika*, 68, 373–379.
- Reid, N. and Crèpeau, H. (1985), "Influence Functions for Proportional Hazards Regression," *Biometrika*, 72, 1–9.
- Ripatti, S. and Palmgren, J. (2000), "Estimation of Multivariate Frailty Models Using Penalized Partial Likelihood," *Biometrics*, 56, 1016–1022.
- Sargent, D. J. (1998), "A General Framework for Random Effects Survival Analysis in the Cox Proportional Hazards Setting," *Biometrics*, 54, 1486–1497.
- Schemper, M. and Henderson, R. (2000), "Predictive Accuracy and Explained Variation in Cox Regression," *Biometrics*, 56, 249–255.
- Schoenfeld, D. A. (1982), "Partial Residuals for the Proportional Hazards Regression Model," *Biometrika*, 69, 239–241.

Einführendes Beispiel

Lungenkrebs Daten (Kalbfleisch&Prentice,1980)

n = 137; 9 zensierte Lebensdauern

Prognosefaktoren	Datentyp	Verteilung
<i>Behandlung</i>	dichotom	n_1, n_2 : 69, 68
<i>Alter</i>	quantitativ	Quartile: 51, 62, 66
<i>Histologie</i>	qualitativ	$n_1 - n_4$: 35, 48, 27, 27
<i>Performance Status</i>	quantitativ	Quartile: 40, 60, 75

Lungenkrebs Daten (Kalbfleisch&Prentice,1980)

Analyse partieller Effekte (multiples Cox Modell)

Prognosefaktoren	Rel. Risiko	p-Wert	EV
<i>Behandlung</i>	1.3	0.15	1.0%
<i>Alter</i>	0.99	0.34	0.4%
<i>Histologie</i>	2.3 / 3.2 / 1.5	<0.0001	9.3%
<i>Performance Status</i>	0.97	<0.0001	19.0%

Totale EV: 36.1%

Beispiel aus Dunkler, Michiels & Schemper (2007):

Ausgangspunkt: van't Veer et al.: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002.

Ziel: Bestimmung des tatsächlichen prognostischen Nettonutzens des gene expression profiling laut obiger Arbeit.

Methode: Berechnung der Zunahme an erklärter Variabilität für Lebensdauer Daten
(nach Schemper-Henderson, *Biometrics*, 2000)

Ergebnis

	Explained Variation in %	
		Standard Error
<i>Model without predictors</i>	0	
<i>Model with clinical characteristics</i>	16	± 5
<i>Model with gene classifier</i>	12	± 4
<i>Model with “clinical characteristics” and gene classifier</i>	19	± 5
<i>Gain by adding gene classifier to “clinical characteristics”</i>	3	

