

A supplement to the SAS survival guide - nonparametric regression

Karl Ernst Siegler, CRS-Mannheim GmbH, Grünstadt, Germany

ABSTRACT

Along with the well-known Cox-model in the area of survival analysis there is an interesting alternative: The Additive Hazards Regression model, which is sometimes referred to as Aalen's Linear Regression model (or simply Aalen-model). In contrast to the semi-parametric Cox-model, the Aalen-model is nonparametric. The concept permits an assessment of time dependency in covariate effects with regression functions and corresponding graphics. Those graphics are denoted as Aalen-plots. Beyond, the model allows the inclusion of time-varying covariates.

The basic principles of this model will be presented throughout this paper, including estimates, confidence intervals and test statistics. Based on the statistical theory of counting processes and martingales the calculations turn out to be reasonable simple. Programming was done with SAS® software, incorporating SAS/Graph® and SAS/IML®.

The model is applied to data from EVITA-HF (Evidence Based Treatment - Heart Failure) survey. Results are presented in short and different types of Aalen-plots are identified.

INTRODUCTION

In clinical trials the focus might be on the statistical analysis of life expectancies or survival times. Throughout this paper, survival times are considered as elapsed time until a certain event e.g. death or another "absorbing" state occurs.

A SHORT HISTORY OF SURVIVAL TIME ANALYSIS

There is a long tradition in the statistical analysis of life times or survival times. In the beginning it was applied to demographic objectives (Lexis 1875). Later an interest in analysing life cycles (e.g. life periods of light bulbs) grew in engineering sciences (Weibull 1939).

Translation of these methods to clinical trials is not straightforward. So-called "censored" observations are usually an issue in this context. This means that some patients are alive at the end of the observation period in a study. In this case a certain "survival time" is known. But this knowledge is incomplete. It is regarded as a partly known survival time. At least it covers the time under observation. In another case it might happen that a patient deceases due to circumstances which are not directly linked to the study, e.g. death by accident. Ignoring these types of information would lead to an under-estimation of survival probabilities. This type of censoring is sometimes denoted as "right censoring", because the individual time axis is cut off at the right-hand side. Other censoring schemes are possible. "Left censoring" describes cases where the start of the observation period is not known. This might occur, if the starting point of a certain disease is unclear. "Interval censoring" with intermittent observation periods is possible as well. In this paper the use of censoring is restricted to right censored data.

The statistical analysis of censored survival times started with the estimation of the survival probability functions (Kaplan, Meier 1985). Due to the importance in clinical trials, a broad range of methods was developed afterwards. Starting with the log-rank test (Peto, Peto 1972) for the statistical comparison of survival probability functions, a vast amount of methods and applications was published. The best know contribution was made with the Cox proportional hazards model (Cox 1972). It allows estimating the influence of several covariates on survival with regression methods.

Growing interest in counting processes and martingale theory in the 1980s and 1990s led to the development of another regression model for the estimation of covariate influence on survival - the nonparametric additive hazard model. Sometimes it is referred to as "Aalen-model", named after the contributions made by O.O. Aalen 1980, 1989 and 1993.

For the application of regression models in survival analysis, it is important to know that hazard functions (or rates) are modeled, rather than survival probability functions. A hazard function describes the instantaneous probability of shifting into an absorbing state (i.e. death) as a function of time.

AVAILABLE SAS PROCEDURES

For the statistical analysis of censored survival times, at least two variables are required for each patient: The first is the observed survival time (in days, weeks or years) regardless of any censoring information. This information is coded in a second variable with two realizations ("censored", "complete observation until death"). The information about treatment arms might be requested as well for each patient.

PhUSE 2012

In the analysis of survival times with censoring, Kaplan-Meier estimates are used for the estimation of the survival experience in different treatment arms or other groups. These survival curves are given as a function over time in appropriate graphs (Kaplan-Meier curves). A statistical test (log-rank test) is applied to assess the difference in survival experience.

If covariates are considered, they have to be available as further variables, e.g. blood pressure, tumor staging etc. In clinical trials there are often more variables (covariates) which exhibit a prognostic effect on survival. The Cox-models is applied to assess the influence of those covariates. This influence on survival is modeled with appropriate regression methods. For each covariate a regression parameter β is estimated. A statistical test for the hypothesis $H_0: \beta=0$ is applied. A p-value<0.05 (or another appropriate significance level) indicates that the corresponding covariate has a significant effect on survival.

Several SAS procedures are available for the analysis of survival times:

- proc lifereg - Parametric models for failure time data, e.g. for Weibull and exponential distribution, considering censored observations.
- proc lifetest - Estimation of the survival probability function in terms of Kaplan-Meier curves and comparison of such curves with log-rank tests (and many more applications).
- proc phreg - Regression analysis of survival data, based on the Cox proportional hazards model, estimates of covariate influence and of survival probability functions with tests and confidence intervals (and many more applications). This SAS procedure allows for time varying covariates within the counting process style of input.
- proc surveypreg - Regression analysis of survival data based on the Cox proportional hazards model for complex survey sample designs.

Besides the Cox-model, the Aalen-model might be applied for regression analysis of censored survival times. As this model is not available in SAS software it will be introduced in the following sections.

THE NONPARAMETRIC ADDITIVE HAZARD MODEL

The nonparametric additive hazard model (Aalen-model) finds its application in the same area as the Cox-model. In addition to the assessment of covariate influence on hazard functions with statistical tests and confidence intervals, the Aalen-model offers interesting graphical results. In fact, the primary estimates from this model are cumulated regression functions in contrast to regression parameters as described above. These functions might be graphically displayed as so-called Aalen-plots. They allow a visual assessment of the influence of each covariate on the survival experience over time.

A statistical test is applied in order to assess if the cumulated regression-function is different from zero, i.e. if a significant covariate effect is present.

LINEAR HAZARD MODEL

The model is used to estimate the influence of several covariates on survival. Again, survival is expressed as hazard rate. Aalen suggests a linear hazard model, postulating a linear combination of a covariate matrix $Z(t)$ and a vector of unknown regression functions $\gamma_j(t)$ which model the hazard rate $\alpha_i(t)$ with patients $i=1, \dots, n$.

$$\alpha_i(t) = \gamma_0(t) + \gamma_1(t) \cdot Z_{i1}(t) + \dots + \gamma_p(t) \cdot Z_{ip}(t)$$

$Z_{ij}(t)$ denotes the elements of a $n \times (p+1)$ covariate-matrix, i.e. the matrix containing the covariate information (in each of p columns) for each patient (in each of n rows). Most important at this place is the linear association between hazard rate, regression functions and covariates.

CUMULATIVE REGRESSION FUNCTIONS

In the model above, the regression-functions are to be estimated. From reasons lying in the theory of counting processes and martingales, it is not possible to estimate the regression functions directly. Therefore cumulative (or integrated) regression functions are estimated by integration over time.

$$\Gamma(t) = \int_0^t \gamma(s) ds$$

This denotes the integration of all instantaneous increments. It might be interpreted as a sequence of cumulative sums over time.

NONPARAMETRIC ESTIMATION OF CUMULATIVE REGRESSION FUNCTIONS

Aalen (1989) shows that the estimation of the integrated hazard function is a least square estimator for each survival time.

$$\hat{\Gamma}(t) = \sum_{T_k \leq t} \left[\left(Y(T_k)^T Y(T_k) \right)^{-1} Y(T_k)^T \right] \cdot I_k$$

- T_k denotes the distinct, observed survival times.

PhUSE 2012

- I_k denotes a $nx1$ -vector, indicating the current survival time, i.e. for the k -th survival time, the vector consists of a 1 in the k -th row and 0 in all other rows. There will be no 1 at places with censored data. Therefore no estimates will be given at time points corresponding to censored data, i.e. the increments are zero.

The basic structure of the estimate is that of a regular least squares estimate which is calculated at several time points. Different is the time dependency of the Y -matrix, containing the covariate information, which is worth a further description in the next section.

THE DESIGN-MATRIX

The design matrix at the time of first observation contains all covariate values from the covariate matrix $Z(t)$.

$$Y(T_1) = \begin{pmatrix} 1 & Z_{11}(T_1) & \cdots & Z_{1p}(T_1) \\ \vdots & \vdots & & \vdots \\ 1 & Z_{n1}(T_1) & \cdots & Z_{np}(T_1) \end{pmatrix}$$

The design matrix at the time of i -th observation shows a different shape. Deceased individuals do not make any contribution to the next estimates. So the corresponding lines are set to zero.

$$Y(T_i) = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \\ 1 & Z_{i1}(T_i) & \cdots & Z_{ip}(T_i) \\ \vdots & \vdots & & \vdots \\ 1 & Z_{n1}(T_i) & \cdots & Z_{np}(T_i) \end{pmatrix}$$

And the design matrix at the time of last observation comes out with only one non-zero line present.

$$Y(T_n) = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \\ 1 & Z_{n1}(T_n) & \cdots & Z_{np}(T_n) \end{pmatrix}$$

From this scheme it is easy to see, that every kind of censoring scheme might be taken into account by setting the lines in this matrix to zero or not. This comprises left or right censoring and even interval censoring. Time varying covariates can be included into the model as well by updating values. If a covariate value in a patient changes during the course of the trial, say after d days under observation, the respective line in the design matrix is updated for all time points after day d .

CONFIDENCE INTERVALS AND TEST

The calculation of test statistics and confidence intervals is similar to the calculations explained above, incorporating a matrix with appropriate weights. The covariance matrix provides standard errors for confidence intervals for each time point (point-wise confidence intervals).

$$\hat{\Omega}(t) = \sum_{T_k \leq t} Y^-(T_k) \cdot \text{diag}(I_k) \cdot Y^-(T_k)'$$

Statistical tests are requested for testing if the effect of a covariate on the hazard rate is different from zero. To do so, a process of test statistics is estimated. The process of the test statistic is a weighted sum of the increments.

$$\hat{H}(t) = \sum_{T_k \leq t} L(T_k) \cdot Y^-(T_k) \cdot I_k$$

There are several different proposals for the weight process $L(t)$. The simplest is to take the number of individuals at risk. Aalen (1998) uses a weight process, proportional to the variance process. The overall statistic is derived from this process. The last time point is used and multiplied by a weighted version of the covariance matrix which yields a normally distributed statistic.

PhUSE 2012

SAS PROGRAMMING

The data requirements are straightforward: The survival times T_k and the information on censoring is needed for construction of I_k . The design-matrix, containing the covariate information is requested as well. This is a fixed $n \times (p+1)$ -matrix as long as no time-varying covariates are involved.

Consider data to be as follows.

Patient Number	Survival time (days)	Censored observation (yes/no)	Treatment (A/B)	Age (years)	Gender (male/female)	Tumor grading	<other covariate values>
001	15	no	A	79	female	I	
002	26	yes	A	65	male	III	
003	43	no	B	75	Male	15 days: III 43 days: IV	

The columns on the right, starting with "Treatment" define the design matrix. After an observation period of 15 days, the first patient dies (=not censored observation) and the second is lost to follow up after 26 days (=censored observation). The third patient dies after 43 days. Estimation and Aalen-plots are to be estimated for survival times at 15 days and 43 days. There are no calculations for 26 days, because this is a censored observation. If time-varying covariates are included, this matrix has to be updated. For example, consider the covariate "Tumor grading". In this example it might happen that patient 003 contributes with tumor grade III to the first survival time, i.e. to the calculations for day 15, and with grade IV to the calculations at day 43. This is an illustration for a time varying covariate, which enters the calculations by updating the design matrix.

The accumulation of increments includes all estimates up the current survival time. Therefore a recursive calculation process is necessary. All this is easily done with SAS/IML, where a do-loop is requested besides multiplication and inversion of matrices. The results are displayed as cumulative regression functions over time applying SAS/Graph.

APPLICATION TO DATA FROM EVITA-HF SURVEY

The EVITA-HF (Evidence Based Treatment - Heart Failure) is a prospective multi-center survey. Data are collected in 13 hospitals throughout Germany. The observation period started in 2009. In the current state, approx. 2800 patients are included. Follow-up is completed for approx. 1420 patients, who are available for survival analysis.

Observations with missing values in covariates are deleted from any calculations. There are 1347 observations left with 187 events (=deaths) and 1160 censored observations. The survival times (censored or not) are given in days. They are in the range from 1 day to 811 days (approx. 2.25 years).

The model is applied to the data and eight covariates are analyzed:

- Gender (male / female)
- Age at entry in years
- LVEF = left ventricular ejection fraction, the volume of blood pumped out of the ventricle (heart) with each heart beat; lower values represent a worse heart-function
- NYHA classification with values from I to IV, a measure for physical performance of patients suffering from heart failure, published by the "New York Heart Association"
- ICM = ischemic cardiomyopathy, a primary reason for heart failure, can be understood as poor oxygen supply of the heart muscle
- CMP = cardiomyopathy or "heart muscle disease", a primary reason for heart failure caused by deterioration
- MI = previous myocardial infarction
- Renal failure, a concomitant disease with a large prognostic impact

For exploratory purposes, the results in terms of p-values from the Cox-model (proc phreg) and from the Aalen-model are given side by side:

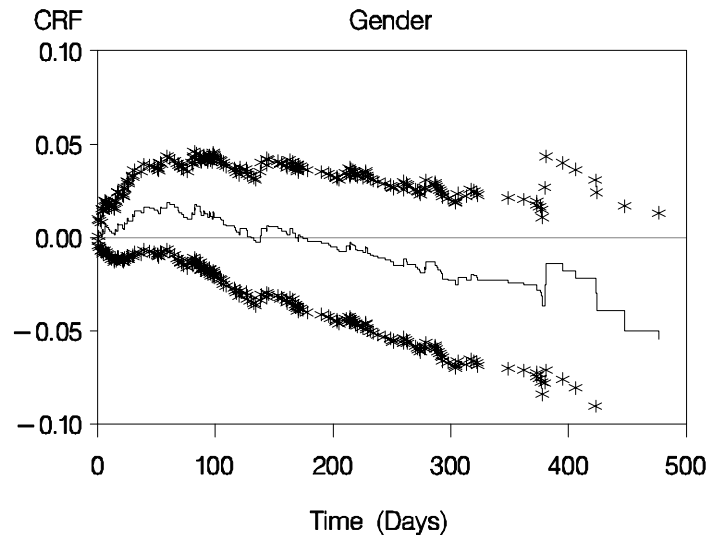
Covariate	p-values for influence of covariates	
	Cox-model	Aalen-model
Gender	0.359	0.277
Age	0.007	0.005
LVEF	<0.001	<0.001
NYHA	0.005	0.010
ICM	0.304	0.174
CMP	0.194	0.146
MI	0.137	0.102
Renal failure	<0.001	<0.001

PhUSE 2012

In addition, there are the so-called Aalen-plots which might be used for a further investigation of the influence of covariates on survival times. The whole set is given at the end of this paper. Selected Aalen-plots are briefly discussed in the following sections.

COVARIATE WITH NO INFLUENCE ON THE HAZARD RATE

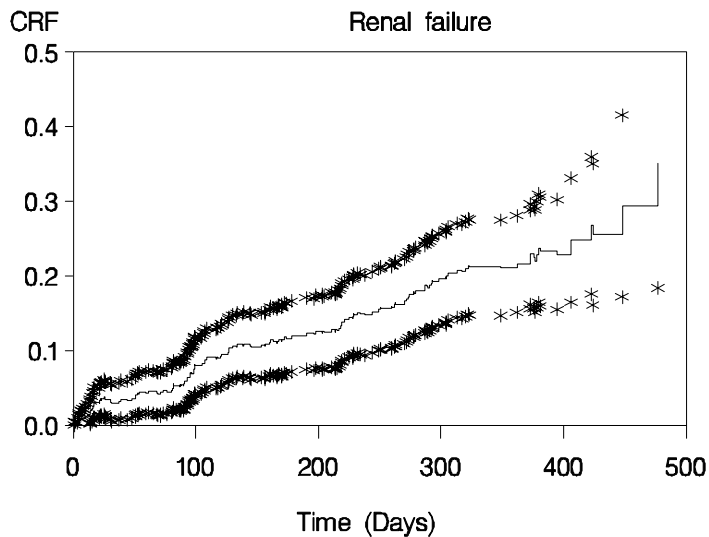
The covariate "Gender" is an example for a covariate with no influence on hazard rate. The Aalen-plot for this covariate is given below (CRF = cumulative regression function). It is obvious, that the 95% confidence intervals include the zero line at each time point. The p-value of 0.277 confirms this.



COVARIATE WITH A PERSISTENT INFLUENCE ON THE HAZARD RATE

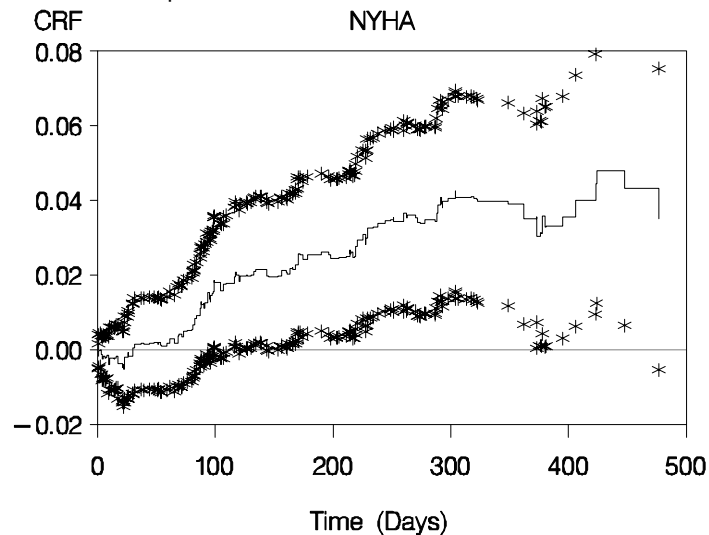
The covariate "Renal failure" is an example for a covariate with a persistent influence on hazard rate. The Aalen-plot for this covariate is given below. The 95% confidence intervals do not include the zero line at any time point.

The p-value of <0.001 confirms this. If the slope in an Aalen-plot is ascending or descending, depends on the coding of covariates. In this example, renal failure was coded as 1 and the absence of renal failure was coded as 0. Patients who experienced renal failure have a higher probability to decrease.



COVARIATE WITH A TIME-VARYING INFLUENCE ON THE HAZARD RATE

The covariate "NYHA classification" is an example for a covariate with a time-varying influence on hazard rate. The Aalen-plot for this covariate is given below. The 95% confidence intervals include zero up to ca. 100 days. Afterwards the 95% confidence intervals are beyond zero. In this case, the risk for patients with different NYHA classifications is equal up to 100 days. After this time point the risk is higher for patients within higher NYHA classes. Although the last values for the confidence intervals include zero, the p-value is 0.010, indicating that a significant influence was detected during the course of the Aalen-plot.



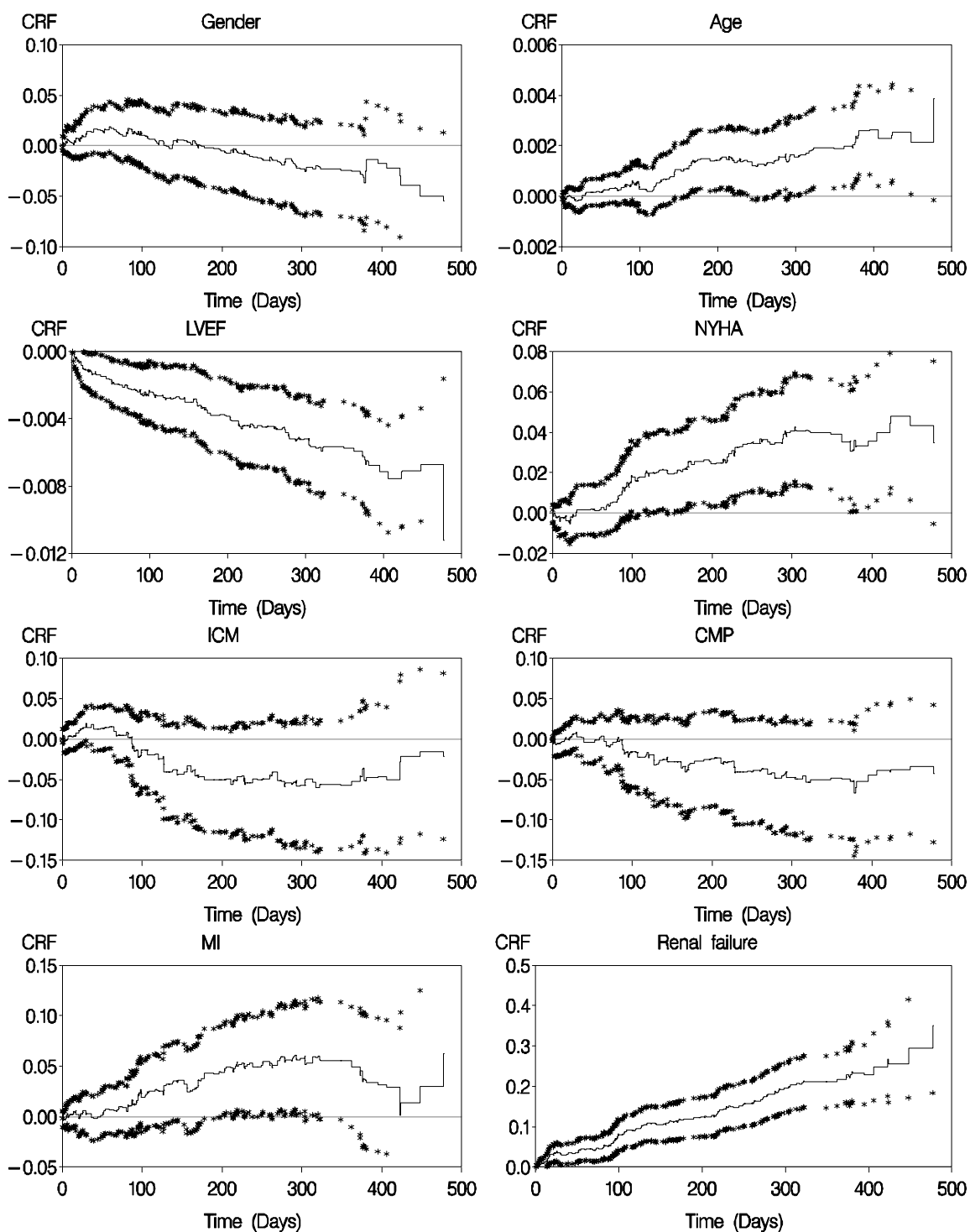
CONCLUSION

Although all calculations are easy to do and the model is described extensively in the statistical literature, the experience in clinical trials is limited.

As no SAS procedures are available and the cumulated regression function is not easy to understand in the context of clinical trials, the practical experience with this model is low. Nevertheless "Aalen-plots" can give valuable and deeper insight in results of any survival analysis. Other advantages of the nonparametric additive hazard regression model are the possibilities to involve all kinds of censoring schemes and to account for time varying covariates. Beside this, the calculations are straightforward and do not require any numerical solutions.

Throughout this paper, not all potentials were explored. The author thinks that further research might be useful, especially in the examination of the time dependency of covariate influence in this particular regression model. There are several published approaches for describing a change point in hazard rates. These methods might be translated to the cumulative regression functions. Some references are given in the recommended reading section.

ALL AALEN-PLOTS CREATED FOR THE EVITA-HF SURVEY



PhUSE 2012

REFERENCES

- Aalen, O.O. (1980) A model for nonparametric regression analysis of counting processes. Springer Lect. Notes Statist. 2, 1-25.
- Aalen, O.O. (1989) A linear regression model for the analysis of life times. Statist. Med. 8, 907-925.
- Aalen, O.O. (1993) Further results on non-parametric regression models in survival analysis. Statist. Med. 12, 1569-1588.
- Cox, D.R. (1972) Regression models and life-tables (with discussion). J. Roy. Statist. Soc. B 4, 187-220.
- Kaplan, E.L., Meier, P. (1985) Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. 53, 457-481.
- Lexis, W. (1875) Einleitung in die Theorie der Bevölkerungsstatistik. Trübner, Strassburg.
- Peto, R., Peto, J. (1972) Asymptotically efficient rank invariant procedures. J. Roy. Statist. Soc. A 135, 185-206.
- Weibull, W. (1939) A statistical theory of the strength of materials. Ing. Vetenskaps Akad. Handl. 151, 1-45.

Comprehensive textbooks:

- Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding, N. (1993) Statistical models based on counting processes. Springer, New York.
- Klein, J.P., Moeschberger, M.L. (1997) Survival analysis: Techniques for censored and truncated data. Springer, New York.

ACKNOWLEDGMENTS

Thanks to Dr. Steffen Schneider and Dr. Matthias Hochadel from the "Institut für Herzinfarktforschung" in Ludwigshafen for providing the data used as an illustration.

And thanks to Prof. Dr. Jochen Mau (Düsseldorf) who encouraged me to work with this methodology.

RECOMMENDED READING

- Achcar, J.A., Loibel, S. (1998) Constant hazard function models with a change point: A Bayesian analysis using Markov chain Monte Carlo methods. Biometrical J. 40, 543-555.
- Anderson, J.K., Senthilselvan, S. (1982) A two-stage regression model for hazard functions. Appl. Statist. 31, 44-51.
- Gosh, J.K., Joshi, S.N., Mukhopadhyay, C. (1996) Asymptotics of a Bayesian approach to estimating change-point in a hazard rate. Commun. Statist. - Theory Meth. 25(12), 3147-3166.
- Matthews, D.E., Farewell, V.T. (1982) On testing for a constant hazard against a change-point alternative. Biometrics 23, 463-468.
- Matthews, D.E., Farewell, V.T., Pyke, R. (1985) Asymptotic score-statistic processes and tests for constant hazard against a change-point alternative. Ann. Statist. 13, 583-591.
- Mau, J. (1985) Statistical modeling via partitioned counting processes. J. Statist. Planning Inf. 12, 171-176.
- Nguyen, H.T., Rogers, G.S., Walker, E.A. (1984) Estimation in change-point hazard rate models. Biometrika 71, 299-304.
- Yao, Y.-C. (1986) Maximum likelihood estimation in hazard rate models with a change-point. Commun. Statist.-Theor. Meth. 15, 2455-2466.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name	Karl Ernst Siegler
Company	CRS Clinical Research Services Mannheim GmbH
Address	Richard-Wagner-Strasse 20
City / Postcode	67269 Grünstadt
Work Phone:	+49 6359 899 379
Fax:	+49 6359 899 352
Email:	karl-ernst.siegler@crs-group.de
Web:	www.crs-group.de

A sample program for a set of published data is available from the author (ca. 350 lines, uses SAS/IML).

Brand and product names are trademarks of their respective companies.