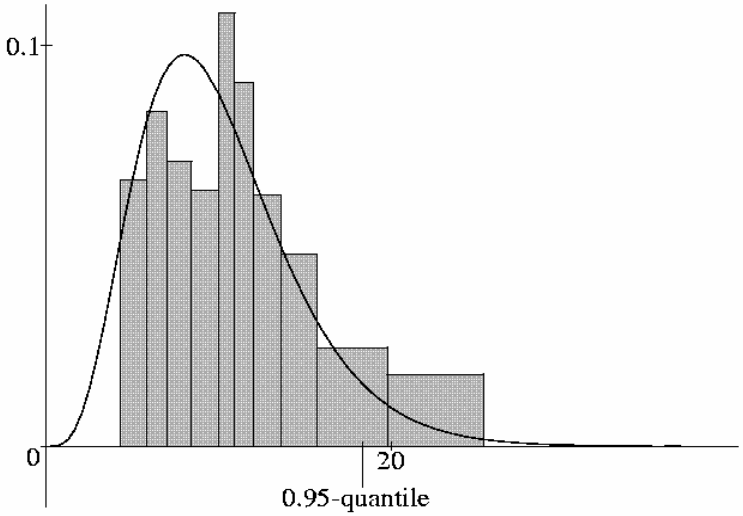
UNIVERSITÄT POTSDAM
Institut für Mathematik

On Approximate Likelihood in Survival Models

Henning Läuter



Mathematische Statistik und
Wahrscheinlichkeitstheorie

Universität Potsdam – Institut für Mathematik

Mathematische Statistik und Wahrscheinlichkeitstheorie

On Approximate Likelihood in Survival Models

Henning Läuter

Institute of Mathematics, University of Potsdam
e-mail: laeuter@uni-potsdam.de

Preprint 2006/04

November 2006

Impressum

Herausgeber: Mathematische Statistik und Wahrscheinlichkeitstheorie
am Institut für Mathematik

Adresse: Universität Potsdam
Am Neuen Palais 10
14469 Potsdam
Telefon:
Fax: +49-331-977 1500
E-mail: +49-331-977 1578
neisse@math.uni-potsdam.de

ISSN 1613-3307

On Approximate Likelihood in Survival Models

Henning Läuter

Institute of Mathematics, University of Potsdam, Germany
e-mail: laeuter@uni-potsdam.de

Abstract: We give a common frame for different estimates in survival models. For models with nuisance parameters we approximate the profile likelihood and find estimates especially for the proportional hazard model.

Key words and phrases: Approximate likelihood; profile likelihood; proportional hazard model;
AMS subject classification: Primary 62N02

1 Introduction

We consider a random life time Y which depends on some explanatory variable X. For describing the dependence between Y and X there are different possibilities. One well-known model is a proportional hazard model which was introduced by Cox (1972, 1975) and he considered the partial likelihood and conditional likelihood estimates. Anderson et al. (1993) discussed a lot of new ideas for the inference in survival models. Bagdonavičius and Nikulin (2002) investigated accelerated life models and several models for time depending covariates. Dabrowska (1997) considered models where the baseline hazard rate also depends on the covariates. Nonparametric estimates are

considered in Liero (2003). We give some results and proposals for estimating the influence of covariates. The problem is formulated as the estimation of finite dimensional parameters β if nuisance parameters η are included. Here the proportional hazard model is a good example.
If $L_n(\beta, \eta)$ is the full likelihood then the profile likelihood for β is defined by

$$pL_n(\beta) = \sup_{\eta} L_n(\beta, \eta).$$

This profile likelihood has nice properties at least if it is finite. The aim of the paper is to give a common frame for the different estimators in survival models. The starting point is the profile likelihood and with different approximations of the profile likelihood we obtain corresponding estimates. We discuss the resulting estimates in examples. One of these examples is the proportional hazard model.

2 Likelihood in proportional hazard models

We study the problem of estimating the conditional distribution of Y given $X = x$. Let C be a random censoring time independent from Y. Assuming there are independent copies (Y_i, C_i, X_i), $i = 1, ..., n$ of (Y, C, X) and we observe (T_i, Δ_i, X_i), $\quad i = 1, ..., n$ for $T_i = \min(Y_i, C_i)$, $\Delta_i = \mathbf{1}(Y_i \leq C_i)$. The conditional hazard rate of Y_i given $X = x$ is $\lambda(y_i \mid x)$. For continuous random life times the likelihood function is proportional to

$$\prod_{i=1}^{n} \lambda(t_i \mid x_i)^{\delta_i} \, \mathrm{e}^{-\Lambda(t_i|x_i)}$$

for

$$\Lambda(z \mid x) = \int_0^z \lambda(\xi \mid x) \, \mathrm{d}\xi.$$

For the proportional hazard model we have

$$\lambda(y_i \mid x) = \lambda_0(y_i)\psi(x, \beta)$$

where the baseline hazard rate λ_0 and the finite dimensional parameter β are unknown. The parametric form of the function ψ is known. Then this leads to the full likelihood function (up to factors)

$$L_n(\beta, \lambda_0) = \prod_{i=1}^{n} \lambda_0(t_i)^{\delta_i} \psi(x_i, \beta)^{\delta_i} \exp\Big(- \psi(x_i, \beta) \int_0^{t_i} \lambda_0(\xi) \, \mathrm{d}\xi \Big). \qquad (1)$$

Here β is the finite dimensional parameter of interest, and λ_0 is a infinite dimensional nuisance parameter η.

3 Likelihood in parametric models

In parametric models some expressions are given explicitly. Moreover we use the representations given here in section 4.1. Considering parametric models of independent and identically distributed Y_i, $i = 1, ..., n$ where the distribution is known up to an unknown parameter $\mu \in \mathbb{R}^{k+s}$. The asymptotic inference in parametric models goes back to LeCam and Hajek. A summary is given e.g. in Bickel et al. (1993). The m.l.e. $\hat{\mu}_n$ maximizes the log-likelihood $l_n(\mu)$ and under mild conditions it is an efficient estimator and we have

$$\sqrt{n}(\hat{\mu}_n - \mu) \longrightarrow \mathsf{N}_{k+s}(0, \mathcal{J}^{-1}(\mu)),$$

where $\mathcal{J}(\mu)$ is the Fisher information matrix. We estimate $\mathcal{J}(\mu)$ by the "observed" information matrix

$$\mathcal{J}_n(\mu) = -\frac{1}{n}\left(\frac{\partial^2}{\partial\mu_i\partial\mu_j} l_n(\mu) \right)_{i,j=1,...,k+s}. \tag{2}$$

If the parameter μ is partioned as $\mu = (\beta, \eta)$ and $\beta \in \mathbb{R}^k$ is a parameter of interest, while $\eta \in \mathbb{R}^s$ is a nuisance parameter then

$$[A_n(\hat{\mu}_n) - B_n(\hat{\mu}_n)C_n^{-1}(\hat{\mu}_n)B_n^t(\hat{\mu}_n)]^{-1} \tag{3}$$

with

$$\mathcal{J}_n(\mu) = \begin{pmatrix} A_n(\mu) & B_n(\mu) \\ B_n^t(\mu) & C_n(\mu) \end{pmatrix}$$

for a $k \times k$ matrix A_n and a $s \times s$ matrix C_n is an asymptotic unbiased estimator for the asymptotic variance of $\hat{\beta}_n$. With the similar block representation for $\mathcal{J}(\mu)$ we find

$$\sqrt{n}(\hat{\beta}_n - \beta) \longrightarrow \mathsf{N}_k(0, [A(\mu) - B(\mu)C^{-1}(\mu)B^t(\mu)]^{-1}).$$

For computing the observed information matrix or the mentioned variances we have to know the second derivatives of l_n and especially $\frac{\partial^2}{\partial\mu_i\partial\mu_j} l_n(\hat{\mu}_n)$ as good approximations of $\frac{\partial^2}{\partial\mu_i\partial\mu_j} l_n(\mu)$. This is one reason why we want to work

with this full log-likelihood function. In the parametric case we have under general mild conditions

$$\frac{1}{n}\, l_n(\hat{\mu}_n) \longrightarrow \mathsf{E}_\mu \ln f(Y_1, \mu).$$

4 Profile likelihood

We consider distributions $\mathsf{P}_{(\beta,\eta)}$ where $\beta \in \mathbb{R}^k$ and η is a high dimensional nuisance parameter. The profile likelihood $pL_n(\beta, \eta)$ has many properties of the original likelihood, at least if the nuisance parameter is finite dimensional (Barndorff-Nielsen and Cox (1994), Murphy and van der Vaart (2000). In any case the m.l.e. $\hat{\beta}_n$ maximizes pL_n. In some models pL_n is infinite. Then the likelihood principle fails. For instance in the proportional hazard model we have

$$pL_n(\beta) = \infty \quad \forall \beta$$

with $\beta = \beta$, $\eta = \lambda_0$. A similar result we meet in a standard situation of nonparametric regression.

Example: Consider a nonparametric regression model with nonrandom regressors

$$Y_i = m(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathsf{N}(0, \sigma^2),\ i.i.d.$$

$\beta = \sigma^2$ is the parameter of interest and $\eta = m$ is the nuisance parameter. The points $x_i, i = 1, ..., n$ should be different. Then we have

$$\sup_\beta pL_n(\beta) = \infty \quad \forall n.$$

Here a function $\hat{m}$ with $\hat{m}(x_i) = y_i$ is an estimate for m. So one cannot estimate the variance of the errors with such an unrestricted estimate $\hat{m}$. Obviously all information from the observations is used for the estimation of the nuisance parameter and the estimation of β is impossible.
We learn from these two cases that the estimation of the nuisance parameter without restrictions about η can lead to undesired problems. One has to work with restrictions and this is expressed by approximations of the likelihood or profile likelihood. A first possibility for finding an estimation of the parameter β is to restrict the space of the nuisance parameter.

4.1 Smoothness classes

Let $\mathcal{M}$ be a linear space which contains all possible η. Then we define a sequence of d_j-dimensional sets $\mathcal{M}_j$ and

$$\mathcal{M}_j \subset \mathcal{M}_{j+1}, \quad \mathcal{M}_j \longrightarrow \mathcal{M}, \quad j = 1, 2, ...$$

Here $\mathcal{M}_j \longrightarrow \mathcal{M}$ is understood in the convergence of some norm in $\mathcal{M}$. Then L_n can be maximized over the $k + d_j$-dimensional space $\mathbb{R}^k \times \mathcal{M}_j$. Let be

$$\max_{\mathbb{R}^k} \max_{\mathcal{M}_j} L_n(\beta, \eta) = L_n(\hat{\beta}_n^{(j)}, \hat{\eta}_n^{(j)}). \tag{4}$$

Of course $\hat{\beta}_n^{(j)}$ is an approximation of β and at the same time $\hat{\eta}_n^{(k)}$ is an estimation for η. The set $\mathcal{M}_j$ is to be chosen in such a way that $(\beta, \eta) \in \mathbb{R}^k \times \mathcal{M}_j$ is identifiable. The rate of convergence of $(\hat{\beta}_n^{(j)}, \hat{\eta}_n^{(j)})$ to (β, η) is determined by the dimension d_j of $\mathcal{M}_j$.
Usually the Fisher information $\mathcal{J}(\beta, \eta)$ or $\mathcal{J}_n(\beta, \eta)$ are defined as linear operators. We choose a basis in $\mathcal{M}$ in such a way that $\eta \in \mathcal{M}_j$ is represented by a infinite dimensional vector where all components are 0 except the first d_j components. So the elements of $\mathcal{M}_j$ are "smoother" in comparison with the elements of $\mathcal{M}$. Then under $(\beta, \eta) \in \mathbb{R}^k \times \mathcal{M}_j$ the observed information matrix as in (2) has the form

$$\mathcal{J}_n(\beta, \eta) = \begin{pmatrix} A_n(\beta, \eta) & \tilde{B}_n(\beta, \eta) & 0 \\ \tilde{B}_n^t(\beta, \eta) & \tilde{C}_n(\beta, \eta) & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{5}$$

for a $k \times k$ matrix A_n, $d_j \times d_j$ matrix $\tilde{C}_n$, $k \times d_j$ matrix $\tilde{B}_n$ and is considered as an approximation for $\mathcal{J}_n(\beta, \eta)$ for $(\beta, \eta) \in \mathbb{R}^k \times \mathcal{M}$. Then as in (3) we approximate the variance of $\hat{\beta}_n^{(j)}$ by

$$\mathsf{Var}\hat{\beta}_n^{(j)} \approx [A_n(\hat{\beta}_n^{(j)}, \hat{\eta}_n^{(j)}) - \tilde{B}_n(\hat{\beta}_n^{(j)}, \hat{\eta}_n^{(j)})\tilde{C}_n^{-1}(\hat{\beta}_n^{(j)}, \hat{\eta}_n^{(j)})\tilde{B}_n^t(\hat{\beta}_n^{(j)}, \hat{\eta}_n^{(j)})]^{-1}. \tag{6}$$

Really we are only interested in the convergence of the estimate of the parameter of interest. Considering again the nonparametric regression example.

Example: (Continuation)
We consider the nonparametric normal regression model. Assuming the regression function is square-integrable and with an orthonormal basis $\{g_s, s = 1, ...\}$ we have the representation

$$m(t) = \sum_{s=1}^{\infty} \eta_s \, g_s(t).$$

With the $n \times d_j$ matrix

$$X_{(j)} = \begin{pmatrix} g_1(t_1) & \cdots & g_{d_j}(t_1) \\ & \ddots & \\ g_1(t_n) & \cdots & g_{d_j}(t_n) \end{pmatrix}, \quad y_{(j)} = \begin{pmatrix} y_1 \\ \vdots \\ y_{d_j} \end{pmatrix}$$

we have

$$\hat{\sigma}^2 = \frac{1}{n} \mid (I - P_j) y_{(j)} \mid^2$$

where P_j is the orthogonal projection on the space spanned by the columns of $X_{(j)}$ and $|\cdot|$ is the Euclidean norm. We see that with increasing n and j under $d_j/n \longrightarrow 0$ the $\hat{\sigma}^2$ is an efficient estimate for σ^2.
In other problems the rate of convergence depends on d_j in a more complicated way.
This was an approach for finding estimates of β without changing the likelihood. We restricted the space of nuisance parameters in such a way that the whole space is approximated by a sequence of finite dimensional spaces. Another possibility for a similar approximation was proposed by Huang and Stone (1998). They considered classes of splines and found convergence rates for the estimates.
Up to this point we worked with an approximation of the profile likelihood. Another approach for constructing estimates of the parameter of interest bases on the replacement of the full likelihood L_n by an appropriate function $\tilde{L}_n$. The then desired estimator is the maximizer of $\tilde{L}_n$.

4.2 Approximate likelihood function

We approximate the likelihood function in two different ways. We formulate this for proportional hazard models.

4.2.1 Λ_0 is approximated by a stepwise constant function.

Let the steps of Λ be in the observed points $t_1, \ldots, t_n$. This leads to

$$\tilde{\Lambda}_0(t_i) = \sum_{j: t_j \leq t_i} \lambda_j$$

and put $\lambda_0(t_j) = \lambda_j$ (Andersen et al. (1993), Murphy and van der Vaart (1997, 2000), Owen (2001), Lawless (2003).

Substituting this in (1) we get the approximated likelihood

$$\tilde{L}_n(\beta, \lambda_1, \ldots, \lambda_n) = \prod_{i=1}^{n} \lambda_i^{\delta_i} \psi(x_i, \beta)^{\delta_i} \exp\left(-\psi(x_i, \beta)\tilde{\Lambda}_0(t_i)\right).$$

Here $\lambda_1, \ldots, \lambda_n$, and β are unknown. Maximizing this w.r.t. $\lambda_1, \ldots, \lambda_n$ we obtain the profile likelihood. If $\mathcal{R}(t)$ denote the set of individuals which are alive and uncensored to time t and $V_s(t) = \mathbf{1}(s \in \mathcal{R}(t))$ then the profile likelihood is

$$p\tilde{L}_n(\beta) \propto \prod_{k=1}^{n} \Big(\frac{\psi(x_k, \beta)}{\sum_j V_j(t_k)\psi(x_j, \beta)} \Big)^{\delta_k}.$$

This coincide with the partial likelihood of Cox. Firstly Breslow (1974) remarked that the conditional likelihood estimate of Cox is also a partial likelihood estimate. By construction it is also a nonparametric maximum likelihood estimate. Without censoring the profile likelihood is rewritten in

$$p\tilde{L}_n(\beta) \propto \prod_{k=1}^{n} \frac{\psi(x_k, \beta)}{\sum_{j:t_j \geq t_k} \psi(x_j, \beta)}.$$

From this representation it is clear that the estimate of β is a rank statistic.

4.2.2 Λ_0 is approximated by a continuous piecewise linear function.

Let Λ_0 be a continuous piecewise linear function. Because of

$$\tilde{\Lambda}_0(t) = \int_0^t \tilde{\lambda}_0(s)\, \mathrm{d}s$$

the corresponding hazard rate $\tilde{\lambda}_0(s)$ is piecewise constant. Let $t_{(1)}, \ldots, t_{(n)}$ be the ordered observations then we have

$$\tilde{\Lambda}_0(t) = \sum_{i=1}^{k-1} \tilde{\lambda}_0(t_{(i)})(t_{(i)} - t_{(i-1)}) + \tilde{\lambda}_0(t_{(k)})(t - t_{(k-1)})$$

for $t_{(k-1)} \leq t \leq t_{(k)}, \quad k = 1, \ldots, n$. Here is $t_{(0)} = 0$.

Consequently

$$\tilde{\Lambda}_0(t_{(i)}) = \sum_{j=1}^{i} \tilde{\lambda}_0(t_{(j)})(t_{(j)} - t_{(j-1)})$$

holds. The likelihood function is approximated by

$$\tilde{L}_n(\beta, \lambda_1, \ldots, \lambda_n) = \prod_{i=1}^{n} \tilde{\lambda}_0(t_{(i)})\psi(x_{[i]}, \beta)\, \mathrm{e}^{-\psi(x_{[i]},\beta)\sum_{j=1}^{[i]} \tilde{\lambda}_0(t_{(j)})(t_{(j)}-t_{(j-1)})}$$

where $\lambda_i := \tilde{\lambda}_0(t_{(i)})$ and $[i] = j$ if $t_{(i)} = t_j$. We consider the λ_i as nuisance parameters and determine the profile likelihood. For fixed β the function $\tilde{L}_n$ is maximized by

$$\frac{1}{\lambda_k} = (t_{(k)} - t_{(k-1)}) \sum_{i \geq k} \psi(x_{[i]}).$$

So the profile likelihood is

$$p\tilde{L}(\beta) := \max_{\lambda_1,\ldots,\lambda_n} \tilde{L}(\beta, \lambda_1, \ldots, \lambda_n) \propto \prod_{k=1}^{n} \Big(\frac{1}{t_{(k)} - t_{(k-1)}} \; \frac{\psi(x_{[k]}, \beta)}{\sum_i \; V_{[i]}(t_{(k)})\psi(x_{[i]}, \beta)} \; \mathrm{e}^{-n} \Big)^{\delta_{[k]}}$$

and therefore

$$p\tilde{L}(\beta) \propto \prod_{k=1}^{n} \Big(\frac{\psi(x_k, \beta)}{\sum_j V_j(t_k)\psi(x_j, \beta)} \Big)^{\delta_k}.$$

Consequently the maximal $\tilde{\beta}_n$ of the profile (approximated) likelihood is the same as in the previous approximation, i.e. in both cases the Cox estimator is the corresponding solution.

We remark: $\tilde{\beta}_n$ depends on the $t_1, \ldots, t_n$ only in a restricted way: It is not important how large the differences $t_{(i+1)} - t_{(i)}$ are. Without censoring it is obviously that the solution $\tilde{\beta}_n$ is a rank statistic in the observations. But these differences can have a large information about the regression part where we are interested in.

5 Statistical arguments

In section 4.2 both approximations for Λ used n parameters and the second approximation is a continuous function. But both had the same solution. From the asymptotic point of view the resulting Cox estimator is an efficient estimator (Efron 1977). For relatively small sample sizes from the statistical point of view it is very important to include the distances between different

failure times because there is an information about the influence of the covariates. So it is better - at least in these cases - to work with estimates from section 4.1. But here arises the problem of choosing an appropriate class $\mathcal{M}_j$ of smooth nuisance parameters. Often this is a step of experience or prior knowledge. In general one cannot judge which approach is the better one. A broad simulation study can help for giving recommendations. For different types of baseline rate functions and moderate sample sizes n - here I mean sample sizes four or six times the unknown parameter number - samples are generated and the variances of the resulting estimates are computed.
We give now a numerical example from Feigl and Zelen, given in Cox and Oakes (1984).
Example: Two groups of leukaemia patients are considered and the failure time (time to death) in weeks is observed. For any patient the white blood counts (WBC) are given. The two groups are characterized by a positive (17 patients) or negative (16 patients) gene AG. In Cox and Oakes (1984) a proportional hazard model with 3 exploratory variables is taken,

$$\begin{aligned} x_{(1)} &= \mathbf{1}_{AG=pos.}, \\ x_{(2)} &= \ln(WBC) - 9.5, \\ x_{(3)} &= (x_{(1)} - 0.5152)x_{(2)}. \end{aligned}$$

The proposed model is

$$\begin{aligned} \lambda(t, x, \beta) &= \lambda_0(t)\psi(x, \beta) \\ &= \lambda_0(t) \exp\Big(\beta_1 x_{(1)} + \beta_2 x_{(2)} + \beta_3 x_{(3)}\Big). \end{aligned}$$

Using the approach in section 4.1 we choose $\mathcal{M}_1$ as the exponential of quadratic polynomials and so we use

$$\tilde{\lambda}_0(t) \approx \exp(\eta_1 + \eta_2 t + \eta_3 t^2).$$

The resulting estimations are with $n = 33$

$$\hat{\beta}_{33}^{(1)} = [-1.398, 0.413, 0.44] \tag{7}$$

$$\hat{\eta}_{33}^{(1)} = [-0.832, 0.125, 0]. \tag{8}$$

Using the approach in section 4.2 then the estimate is the Cox estimator

$$\widehat{Cox} = [-1.14, 0.4, 0.5]. \tag{9}$$

The plots of the predictor $\beta_1 x_{(1)} + \beta_2 x_{(2)} + \beta_3 x_{(3)}$ with (7) and (9) in Fig. 5 show some differences between both estimates but no other tendency. We point out that in the estimate $\hat{\beta}_{33}^{(1)}$ the failure times are really used not only their ranks. With the first method for a longer region the WBC are less for the group of positive genes than for those with negative genes.

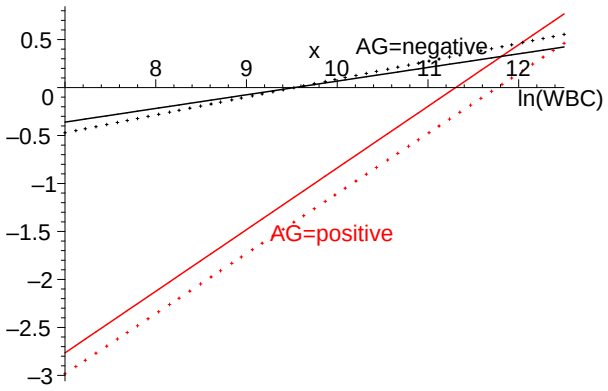


Fig. 1. Data of Feigl and Zelen

with Cox estimator ——, with approx. MLE ·····

Acknowledgement The author is very grateful to Prof. H. Liero for her helpful comments and suggestions.

References

Anderson P., Borgan O., Gill R., Keiding N.(1993). Statistical Models Based on Counting Processes. *Springer Verlag, New York.*

Bagdonavicius V. and Nikulin M. (2002). Accelerated Life Models: Modelling and Statistical Analysis. *Chapman& Hall, London.*

Barndorff-Nielsen O. and Cox D.R. (1994). Inference and Asymptotics. *Chapman& Hall London.*

Bickel P., Klaasen C., Ritov, Y. and Wellner J. (1993). Efficient and Adaptive Estimation for Semiparametric Models. *Johns Hopkins Baltimore.*

Breslow N. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89-99.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *J.Roy.Statist.Soc.Ser.B* **34** 187-220.

Cox D.R. (1975). Partial likelihood. *Biometrika* **62** 269-276.

Cox D. and Oakes D. (1984). Analysis of Survival Data. *Chapman& Hall London*.

Dabrowska D. (1997). Smoothed Cox regression. *Ann.Statist* **25** 1510-1540.

Efron B. (1977). The efficiency of Cox's likelihood function for censored data. *J.Amer. Statist. Assoc*. **72** 557-565.

Huang J.and Stone C. (1998). The L^2 rate of convergence for event history regression with time-dependent covariates. *Scand.J.Statist.*, 603-620.

Lawless J. (2003). Statistical Models and Methods for Lifetime Data. *John Wiley* .

Liero H. (2003). Goodness of fit tests of L_2 type. In: Statistical Inference for Semiparametric Models and Applications, ed. by M. Nikulin, N. Balakrishnan, N. Limnios, M. Mesbah, *Springer-Verlag.*

Murphy S. and van der Vaart A. (1997). Semiparametric likelihood ratio inference. *Ann.Statist*. **25**, 1471-1509.

Murphy S. and van der Vaart A. (2000). On profile likelihood. *J.Amer.Statist.Ass.* **95** 449-465.

Owen A. (2001). Empirical Likelihood. *Chapman& Hall, London*.