UNIVERSITÄT POTSDAM
Institut für Mathematik

Estimation in partly parametric additive Cox models

Henning Läuter

0.95-quantile
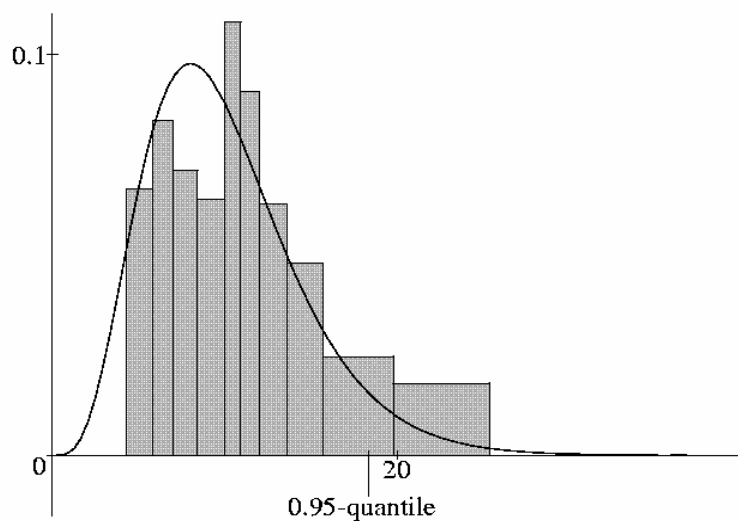
Mathematische Statistik

Estimation in partly parametric
additive Cox models

Henning Läuter

Institute of Mathematics, University of Potsdam

e-mail: laeuter@rz.uni-potsdam.de

Abstract

The dependence between survival times and covariates is described e.g. by proportional hazard models. We consider partly parametric Cox models and discuss here the estimation of interesting parameters. We represent the maximum likelihood approach and extend the results of Huang (1999) from linear to nonlinear parameters. Then we investigate the least squares estimation and formulate conditions for the a.s. boundedness and consistency of these estimators.

1 Introduction

The relationship between the survival time and covariates very often is expressed through the proportional hazards model proposed originally by Cox (1972). In several papers afterwards this model was investigated in several directions (Kalbfleisch and Prentice (1980), Andersen, Borgan, Gill, Keiding (1993), Bagdonavicius and Nikulin (2001)). For the estimation of parameters of interest exist many approaches. In this paper we consider the maximum likelihood and a least norm estimator. Both approaches have some advantages. With the maximum likelihood method we can obtain consistent and efficient estimators. With a least norm estimator it is possible to construct strongly consistent estimators. Survival models describe the probabilistic properties of a nonnegative failure time Y. We assume that Y depends on covariates Z and we model the hazard rate

$$\lambda(t \,|\, z) = \lim_{h \to 0} \frac{\mathsf{P}(t < Y \le t + h \,|\, Y > t, z)}{h}.$$

We investigate parametric and semiparametric multiplicative proportional hazard models

$$\lambda(t \,|\, z) = \lambda_0(t) \, r(\beta, z) \tag{1}$$

for a completely unspecified baseline hazard function λ_0 and the relative risk function r. The function r is assumed to be a partly parametric function. The unknown parameter β is to estimate. The deterministic or random covariate variable Z is constant in time.
We assume that $Z = (X, V) \in \mathbb{R}^k \times \mathbb{R}^m$ and

$$\log r(\beta, z) = g(x, \beta) + l(v) \tag{2}$$

for a known regression functions g, an unknown function l. The covariate X is controlled, but V is an auxilliary covariate. The aim is to estimate β and so to find the influence of $X = (X^{(1)}, ..., X^{(k)})$ on the survival time. Such models for the hazard rate were considered by several authors. Starting point was Cox(1972) with his proportional hazard rate model

$$\lambda(t) = \lambda_0(t) \exp(\beta' x), \tag{3}$$

later on e.g. Andersen&Gill (1982) proved for the regression model (3) the asymptotic normality of the maximum partial likelihood estimator (mple.) for β. The special case of (2) is the semiparametric partial linear model

$$\log r(\beta, z) = x'\beta + l(v) \tag{4}$$

which was considered by Sasieni (1992). He found an information bound for the mple. for β in (4). In Andersen et al. (1993) a broad class of multiplicative and additive models for the hazard rate and related estimates was studied. Bagdonavicius&Nikulin (2001) formulated a lot of models for uncensored and censored variables and found properties for the estimates in these models. Huang (1999) gave an approach for the estimation of β in (4) and could prove the asymptotic efficiency of the approximated mple. We will extend the model (4) to nonlinear functions g and discuss the properties of estimates in these models. In section 5 we discuss properties of least squares estimators where the basis is a nonparametric estimation of the hazard function. Here we use the Nelson-Aalen type estimator considered by several authors, for example Van Keilegom and Veraverbeke (2001) and Liero (2003). Especially we formulate conditions for the boundedness and consistency of the estimator for β.

2　Types of regression functions

Under a precise knowledge of the considered or observed process one has some special models for the hazard rate. The baseline function is unrestricted. The model specification consists in (1) the function r or the function g. We consider different choices of g.

2.1　Linear models

The linear models play a important role in all analyses. In the most of the cases one understands these models as approximations for g or λ. In the model

$$g(x,\beta) = x'\beta = x^{(1)}\beta_1 + ... + x^{(k)}\beta_k \tag{5}$$

with $x \in \mathbb{R}^k, \beta \in \mathbb{R}^k$ the dependence is linear in the controlled covariates and the parameters. If the covariates have a nonlinear influence then

$$g(x,\beta) = w(x)'\beta = \beta_1 + \beta_2\, x + ... + \beta_k\, x^{k-1} \tag{6}$$

is a polynomial model in one covariate x. Here $w(x) \in \mathbb{R}^k$, $\beta \in \mathbb{R}^k$. Polynomial models in higher dimensional covariates are also possible.

2.2 Nonlinear models

The simplest type of functions in this class is

$$g(x, \beta) = \beta_1 \exp(\beta_2 x) \tag{7}$$

for $x \in \mathbb{R}^1, \beta \in \mathbb{R}^2$. These are convex or concave curves. Another examples are the logistic curves which one can describe in different form. The solution $g(x, \beta) := \nu(x)$ of the differential equation

$$\dot{\nu}(x) = \beta_1 \nu(x)(\beta_2 - \nu(x)) \tag{8}$$

is used very often. These curves have a saturation point β_2. There is a huge set of possible other regression curves. In difference to linear regression curves the statistical properties of estimates in nonlinear models depend strongly on the space of parameters. In general the estimation problem leads to ill-posed problems and one needs additional restrictions for β. We are interested in estimating β. There are different possibilities. One approach bases on the likelihood principle. Here one expects that one gets consistent and asymptotically efficient estimates. Another approach bases on a least squares principle. Here one hopes that the corresponding estimates are strongly consistent. We will discuss both possibilities.

---

3 Random covariates and partial likelihood

We consider the nonnegative failure time Y and the censoring time C. Y depends on the covariate Z and we assume that Y and C are conditionally independent given z. $\Delta = \mathbf{1}_{Y \leq C}$ is the censoring indicator and we set $T = \min\{Y, C\}$. Our observable variable is

$$(T, \Delta, Z) \in \mathbb{R}^+ \times \{0, 1\} \times \mathbb{R}^{k+m}. \tag{9}$$

There is given a random sample

$$(T_i, \Delta_i, Z_i), \quad i = 1, ..., n$$

from the distribution of (T, Δ, Z). Here we have

$$Z_i = (X_i, V_i)$$

with $X_i = (X_i^{(1)}, ..., X_i^{(k)})^t, \quad V_i = (V_i^{(1)}, ..., V_i^{(m)})^t$.

The idea of Sasieni (1992) and Huang (1999) consists in approximating the

nonparametric term $l(\cdot)$ in (1) and (2) by a smooth function. They proposed a spline approximation. Assuming

$$l(v) = \sum_{j=1}^{m} l_j(v^{(j)}) \tag{10}$$

then

$$l(v) \approx \sum_{j=1}^{m} s_j(v^{(j)}) \tag{11}$$

for splines s_j of order $q = q(n) \geq 1$. We assume that $a \leq V^{(j)} \leq b$ and choose knots $\xi_1, ..., \xi_p$ with $a \leq \xi_1 \leq ... \leq \xi_p \leq b$ and $p = p(n)$. Furthermore

$$\max_s |\xi_s - \xi_{s-1}| = O(n^{-\alpha})$$

is assumed. In the space $\tilde{\mathcal{S}}_n$ of polynomial splines of order $q \geq 1$ there exists a local basis $B_1, ..., B_{\nu_n}$ such that the representation

$$s_j(\cdot) = \sum_{t=1}^{\nu_n} b_{jt} B_t(\cdot), \quad 1 \leq j \leq m \tag{12}$$

holds. Consequently we have with (11) the approximation

$$l(v) \approx \tilde{s}_n(v, b) := \sum_{j=1}^{m} \sum_{t=1}^{\nu_n} b_{jt} B_t(v^{(j)}) \tag{13}$$

where b is the vector of unknown coefficients $b_{11}, ..., b_{m\nu_n}$.

Assuming the model for the hazard rate function in the form

$$\lambda(t|x, v) = \lambda_0(t) \exp(g(x, \beta) + l(v))$$

then with

$$f_n(x, v) = g(x, \beta) + \tilde{s}_n(v, b)$$

the partial log-likelihood function is given by

$$l_n(\beta, b) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i \{ f_n(X_i, V_i) - \log \sum_{j:T_j \geq T_i} \exp[f_n(X_j, V_j)] \}. \tag{14}$$

5

Any $(\hat{\beta}_n, \hat{b}_n)$ which maximizes $l_n(\beta, b)$ is called a maximum partial likelihood estimator (mple) for (β, b). The additive constant in (14) is free, so we determine such a $\hat{b}_n$ that

$$\sum_{i=1}^{n} \Delta_i \sum_{t=1}^{\nu_n} \hat{b}_n^{(j,t)} B_t(V_i^{(j)}) = 0, \quad j = 1, ..., m$$

with

$$\hat{b}_n^T = \left(\hat{b}_n^{(1,1)}, ..., \hat{b}_n^{(m,\nu_n)} \right)^T.$$

The limit distribution of $\hat{\beta}_n$ is given by the next theorem. For this we formulate at first some assumptions.

Assumption 1 *g is continuously differentiable w.r.t. β.*

Assumption 2 *β is an inner point of a bounded set in $\mathbb{R}^k$ or for any $\tilde{\beta}$ and any sequence $x_1, x_2, ...$ of realizations of $X_1, X_2, ...$ we have*

$$\sum_{i=1}^{n} [g(x_i, \tilde{\beta}) - g(x_i, \tilde{\tilde{\beta}})]^2 \to \infty \quad for \quad \|\tilde{\tilde{\beta}}\| \to \infty.$$

Assumption 3

$$\mathsf{E}\, g(X, \beta)\Delta = 0, \quad \mathsf{E}\, l_j(V^{(j)})\Delta = 0, \quad j = 1, ..., m.$$

Assumption 4

$$\|X\| \leq K \quad a.s. \ for \ a \ finite \ constant \ K,$$

$$V \in [a, b]^m.$$

Theorem 3.1 *Under Assumptions 1 to 4 we have*

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{\mathrm{d}} \mathsf{N}(0, \Sigma), \tag{15}$$

where

$$\Sigma^{-1} = \mathsf{E}\left[\Delta \left(G(X, \beta) - a^*(T) - h^*(V) \right)^{\otimes 2} \right]$$

with $x^{\otimes 2} := xx'$ and

$$G(x, \beta) = \frac{\partial}{\partial \beta}\, g(x, \beta) \in \mathbb{R}^k.$$

Here a^*, h^* are k-dimensional functions with

$$h^*(V) = \sum_{j=1}^{m} h_j^*(V^{(j)})$$

and

$$\mathsf{E}\Delta\|G(X,\beta) - a(T) - h(V)\|^2$$

is minimized by a^*, h^*.

PROOF: The proof combines results from nonlinear regression estimation and results of Huang (1999) for estimation in partly linear models.

$\square$

This theorem formulates an expected result. Important are the assumptions. We mention especially the finite supports for the covariates. The assumption **A**2 ensures the estimability of the parameters. This assumption is important for infinite parameter spaces.

3.1 Linear case

The linear case

$$g(x,\beta) = x'\beta$$

is of special interest. For such models Sasieni (1992) found a lower information bound for the asymptotic variance of an estimator for β. Huang (1999) proved that the maximum partial likelihood estimator $\hat{\beta}_n$ achieves this lower bound.

The result of theorem 3.1 can be used for choosing an appropriate experimental design. We have to find such a random variable X that Σ is in some sense minimal. This is then a general problem of nonlinear estimation.

4 Fix covariates and partial likelihood

For fix covariates $z_1, ..., z_n$ the random variables (T_i, Δ_i, z_i) are independent distributed but not identically. The partial log-likelihood is given by

$$l_n(\beta, b) = \sum_{i=1}^{n} \Delta_i \{f_n(x_i, v_i) - \log \sum_{j:T_j \geq T_i} \exp[f_n(x_j, v_j)]\}. \tag{16}$$

7

Then a maximum partial likelihood estimator $\hat{\beta}_n$ is given as before by such a vector which maximizes l_n. The asymptotic properties of $\hat{\beta}_n$ depends on the sequence of z_i.

---

5 Least squares estimation

We consider a proportional hazard model (1) with $\log r(\beta, z) = g(x, \beta)$. The estimation of the parameter β can be constructed as before by a maximum likelihood principle but also by a least squares criterion. For this we point out that we are interested in the hazard rate

$$\lambda(t \,|\, z) = \lambda_0(t) \, r(\beta, z) \tag{17}$$

or in the cumulative hazard rate

$$\Lambda(t \,|\, z) = \int_0^t \lambda_0(s) \, \mathrm{d}s \ r(\beta, z). \tag{18}$$

The term which includes β is the same in (17) and in (18). Therefore for the estimation of β there is no statistical difference in both equations. The estimation of the cumulative hazard rate is of higher power than those for the hazard rate λ.

In Van Keilegom and Veraverbeke (2001) one finds a nonparametric estimator for the cumulative hazard function Λ. This is a Nelson-Aalen type estimator and is constructed in the following way. We choose a kernel function W and a sequence of bandwidths $\{a_n\}$ and define for any value of the covariates $z_1, ..., z_n$ the weights

$$w_{nj}(z) = \frac{W(\frac{z - z_j}{a_n})}{\sum_{k=1}^n W(\frac{z - z_k}{a_n})} \tag{19}$$

Then we have $\sum_{j=1}^n w_{nj}(z) = 1$ for any z. An estimator for Λ is then

$$\hat{\Lambda}_n(t|z) = \sum_{i=1}^n \frac{\mathbf{1}(T_i \leq t)\Delta_i w_{ni}(z)}{\sum_{j:T_j \geq T_i} w_{nj}(z)}. \tag{20}$$

The parameter β is estimated by comparing $\hat{\Lambda}_n$ with Λ under an appropriate distance measure. We formulate this for the quadratic norm. We call $\hat{\beta}_n$ a least squares estimator if

$$arg \min_{\beta, \lambda} \sum_{j=1}^n \int \left(\hat{\Lambda}_n(t|z_j) - \int_0^t \lambda(s) \, \mathrm{d}s \cdot r(\beta, z_j) \right)^2 b_n(t) \, \mathrm{d}t = (\hat{\beta}_n \, \hat{\lambda}_n). \tag{21}$$

Here $\{b_n\}$ is a sequence of nonnegative weight functions. With (β_0, λ_0) we denote the true parameter in the hazard rate function and set

$$\Lambda_0(t) = \int_0^t \lambda_0(s)\,\mathrm{d}s.$$

We use the notations

$$W_n^2(\beta_0, \Lambda_0, \beta, \Lambda) = \frac{1}{n} \sum_{j=1}^n \int \Big(\Lambda_0(t)\,r(\beta_0, z_j) - \Lambda(t)\,r(\beta, z_j)\Big)^2 b_n(t)\,\mathrm{d}t,$$

$$Z_n(\beta_0, \Lambda_0, \beta, \Lambda) = \frac{1}{n\,W_n(\beta_0, \Lambda_0, \beta, \Lambda)} \times$$

$$\times \sum_{j=1}^n \int \Big(\hat{\Lambda}_n(t|z_j) - \Lambda_0(t)\,r(\beta_0, z_j)\Big) \Big(\Lambda_0(t)\,r(\beta_0, z_j) - \Lambda(t)\,r(\beta, z_j)\Big) b_n(t)\,\mathrm{d}t$$

if $W_n(\beta_0, \Lambda_0, \beta, \Lambda) \neq 0$ and put

$$\varepsilon_n^2 = \frac{1}{n} \sum_{j=1}^n \int \Big(\hat{\Lambda}_n(t|z_j) - \Lambda_0(t)\,r(\beta_0, z_j)\Big)^2 b_n(t)\,\mathrm{d}t.$$

According to (21) our functional which is to minimize is

$$Q_n(\beta, \Lambda) = \frac{1}{n} \sum_{j=1}^n \int \Big(\hat{\Lambda}_n(t|z_j) - \Lambda(t)\,r(\beta, z_j)\Big)^2 b_n(t)\,\mathrm{d}t \qquad (22)$$

and with the introduced notations we have the representation

$$Q_n(\beta, \Lambda) = W_n^2(\beta_0, \Lambda_0, \beta, \Lambda) +$$

$$+ 2\,W_n(\beta_0, \Lambda_0, \beta, \Lambda)\,Z_n(\beta_0, \Lambda_0, \beta, \Lambda) + \varepsilon_n^2. \qquad (23)$$

Assumption 5 *The support of X is a subset $\mathcal{X}$ in $\mathbb{R}^k$. The parameter β lies in a set $\mathcal{B}$. Both set are not necessarily bounded.*

Assumption 6 *ε_n^2 tends a.s. to a finite value σ^2.*

With (23) we obtain immediately a result on the boundedness of the least squares estimator.

Theorem 5.1 *If* **A***5 and* **A***6 and*

$$\inf_{|\beta|\geq\eta} W_n(\beta_0,\Lambda_0,\beta,\Lambda) \to \infty \text{ for } \eta \to \infty$$

hold then $\hat{\beta}_n$ *is bounded a.s.*

Sometimes one can check the behavior of Z_n. Then the following result is useful.

Theorem 5.2 *If for any* β_0 *there exist constants* $c(\beta_0)$, $d(\beta_0)$ *such that for* $\|\beta\| \geq d(\beta_0)$ *for all* $n \geq n_0$ *the condition*

$$W_n(\beta_0,\Lambda_0,\beta,\Lambda) \geq c(\beta_0)$$

and if for some positive η *and a constant* κ *the inequality*

$$\inf_{\|\beta\|\geq\eta} \frac{Z_n(\beta_0,\Lambda_0,\beta,\Lambda)}{W_n(\beta_0,\Lambda_0,\beta,\Lambda)} \leq \kappa < 0.5 \quad a.s.$$

holds then $\hat{\beta}_n$ *is bounded a.s.*

The boundedness is necessary to prove consistency of $\hat{\beta}_n$. In general one has two possibilities to prove the consistency. Either one assume that the parameter lies in a compact set or one finds such conditions that one can restrict itself on compact sets. Our formulated conditions in the previous theorems are such assumptions that we can choose compact subsets which contains the true value β_0 and the estimates $\hat{\beta}_n$ lie a.s. in these subsets.
The representation (23) leads to

$$Q_n(\beta,\Lambda) - Q_n(\beta_0,\Lambda_0) = n\,W_n^2(\beta_0,\Lambda_0,\beta,\Lambda)\Big[1 + 2\,\frac{Z_n(\beta_0,\Lambda_0,\beta,\Lambda)}{W_n(\beta_0,\Lambda_0,\beta,\Lambda)}\Big] \tag{24}$$

This representation shows that the difference of the values of the functional Q_n for different parameter values is expressed by W_n and $\frac{Z_n}{W_n}$. These terms played already a role in the theorems about the boundedness of the l.s.e.

Theorem 5.3 *If there is a compact subset* $\tilde{\mathcal{B}}$ *of the parameter space* $\mathcal{B}$ *such that* $\hat{\beta}_n \in \tilde{\mathcal{B}}$ *a.s. and if* $\inf_{|\beta|\geq\eta} W_n(\beta_0,\Lambda_0,\beta,\Lambda) \to \infty$ *for* $\eta \to \infty$ *then under* **A***6* $\hat{\beta}_n$ *is strongly consistent.*

References

1. Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer, New York.

2. Bagdonavicius, V. and Nikulin, M. (2001). *Accelerated Life Models.* Chapman&Hall.

3. Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* **27** 1536–1563.

4. Läuter, H. (1989). Note on the strong consistency of the least squares estimator in nonlinear regression. *Statistics* **20**, 199–210.

5. Liero, H. (2003). Goodness of fit tests of L_2-type. *in this Volume*

6. Sasieni, P. (1992). Information bounds for the conditional hazard ratio in a nested family of regression models. *J. Roy. Statist. Soc. Ser. B* **54**, 617–635.

7. Van Keilegom, I. and Veraverbeke, N. (2001). Hazard rate estimation in nonparametric regression with censored data, *Ann. Inst. Statist. Math.*, **53**, 730–745