Proportional Hazards Regression with Unknown Link Function

Jane-Ling Wang and Wei Wang
University of California at Davis, Department of Statistics
One Shields Ave.
Davis, CA 95616
wang@wald.ucdavis.edu

Qihua Wang
Beijing University, Department of Statistics, Beijing, China

1. Introduction

Proportional hazards regression model has played a pivotal role in survival analysis since Cox proposed it in 1972. Let T represent survival time and Z its associate covariate vector. Under the proportional hazards model, the hazard function for T, given a particular value z for the covariate Z, is defined as

$$\lambda\{t \mid z\} = \lambda_0(t) \exp\big\{\psi(\beta^T z)\big\}, \tag{1}$$

where $\lambda_0(t)$ is the unknown baseline hazard function corresponding to $z = (0, \cdots, 0)$, and $\psi(0) = 0$.

This model assumes that the covariates affect the survival time through a link function ψ and an index $\beta^T z$. The link function is assumed to be known in the literature, and the most popular choice is the identity link function. In reality, the link function is often unknown and thus needs to be either estimated from the data, or to be validated before a specific form of link function is applied. The goal of this paper is to fill this gap by estimating both the unknown link function, ψ, and the unknown parameter, β, simultaneously. The baseline hazards function can then be estimated afterwards following Breslow's (1974) procedure.

Previously related works with unknown link functions are restricted to univariate covariate or low dimensional covariates without the index structure. The main difference among the various approaches is in the smoother employed to estimate the unknown link function. Tibshirani and Hastie (1987) used local likelihood smoothing technique, and Gentleman and Crowley (1991) gave another approach to local likelihood estimation. Theoretical properties of the local likelihood or local partial likelihood estimators for the link function were explored in Fan, Gijbels and King (1997). Spline smoother based on Penalized partial likelihood was employed by O'Sullivan (1988, 1993), and subsequently by Gray (1992) to breast cancer data, while Sleeper and Harrington (1990) employed the regression splines to estimate the unknown

link function. The approach of Fan, Gijbels and King can be easily extended to k-dimension covariates by estimating a multivariate unknown link function $\Psi(x_1, \cdots, x_k)$. However, as any nonparametric smoothing procedure, such an approach is subject to the curse of dimensionality. The dimension reduction model in (1) has the advantage that it can cope with high dimensional covariates commonly encountered in medical studies, and is more flexible than the conventional proportional hazards model where a known link function is assumed. Once the link function is estimated in model (1), it can further be used to guide the choice of a particular link function or to check the validity of a particular link function such as the identity link function in Cox proportional hazards model routinely employed by practitioners.

In this paper, a two-step iterative algorithm to estimate the link function and the covariate effects is proposed along with the baseline hazard estimate. The link function is estimated by a smoothing method based on a local version of partial likelihood. Asymptotic properties of the estimators are derived for both the parametric covariate effects and the non-parametric estimated link function. The approach is illustrated through a liver disease data and simulations. Both the theory and methods are applicable to censored survival data.

2. Models and Estimation Procedure

Consider the general proportional hazards model (1) with an unknown link function ψ and a k-dimensional covariate vector Z. Suppose Z is a time-independent covariate and the p-th order derivative of $\psi(\beta^T Z)$ at point z exists. Let $Z_1, \cdots, Z_n$ be i.i.d. copies of Z. Then, by Taylor's expansion, for $\beta^T Z$ in a neighborhood of $\beta_0^T z$,

$$\psi(\beta^T Z) \approx \psi(\beta_0^T z) + \psi'(\beta_0^T z)(\beta^T Z - \beta_0^T z) + \cdots + \frac{\psi^{(p)}(\beta_0^T z)}{p!}(\beta^T Z - \beta_0^T z)^P. \tag{2}$$

If we let $\beta^T \mathbf{Z} = \{\beta^T Z - \beta_0^T z, \cdots, (\beta^T Z - \beta_0^T z)^p\}^T$, and
$\beta^T \mathbf{Z}_i = \{\beta^T Z_i - \beta_0^T z, \cdots, (\beta^T Z_i - \beta_0^T z)^p\}^T$,
then, locally around $\beta_0^T z$, $\psi(\beta^T Z)$ can be modeled as

$$\psi(\beta^T Z) \approx \psi(\beta_0^T z) + (\beta^T \mathbf{Z})^T \gamma(\beta_0^T z), \tag{3}$$

where

$$\gamma(\beta_0^T z) = \{\psi'(\beta_0^T z), \cdots, \psi^{(p)}(\beta_0^T z)/p!\}^T.$$

Let K be a kernel function and h be the bandwidth, and define $K_h(z) = h^{-1}K(z/h)$. Let N be the number of uncensored observed survival times. The local partial likelihood, similar

to the one in Fan et al. (1997), is

$$l_n\{\beta,\gamma(\beta_0^T z)\} = \sum_{j=1}^{N} K_h\big\{(\beta^T Z_{(j)}) - \beta_0^T z\big\} \cdot \Bigg[(\beta^T \mathbf{Z}_{(j)})^T \gamma(\beta_0^T z) - \log\Bigg\{\sum_{i\in\mathcal{R}_j} \exp\{(\beta^T \mathbf{Z}_i)^T \gamma(\beta_0^T z)\} K_h\big\{(\beta^T Z_i) - \beta_0^T z\big\}\Bigg\}\Bigg], \tag{4}$$

where $\mathcal{R}_j$ is the risk set for the j^{th} observed time.

In order to solve for both β and γ, we propose a two-stage iterative algorithm.

Step 0. Assign an initial value to β and call it $\hat{\beta}^{[1]}$.

Step 1. Plug $\hat{\beta}^{[1]}$ into the log local partial likelihood, for a given z, we get

$$l_n\{\hat{\beta}^{[1]},\gamma\{(\hat{\beta}^{[1]})^T z\},z) = \sum_{j=1}^{N} K_h\big\{(\hat{\beta}^{[1]})^T (Z_{(j)} - z)\big\} \cdot \Bigg[[(\hat{\beta}^{[1]})^T \mathbf{Z}_{(j)}]^T \gamma\{(\hat{\beta}^{[1]})^T z\} - \log\Bigg\{\sum_{i\in\mathcal{R}_j} \exp\{[(\hat{\beta}^{[1]})^T \mathbf{Z}_i]^T \gamma\{(\hat{\beta}^{[1]})^T z\}\} K_h\big\{(\hat{\beta}^{[1]})^T (Z_i - z)\big\}\Bigg\}\Bigg].$$

Maximize this likelihood w.r.t γ to get the estimate $\hat{\gamma}^{[1]}\{(\hat{\beta}^{[1]})^T z\}$, and repeat this for all data points to get $\hat{\gamma}^{[1]}\{(\hat{\beta}^{[1]})^T Z_i\}, \quad i = 1,\cdots,n$.

Step 2. Plug $\hat{\gamma}^{[1]}\{(\hat{\beta}^{[1]})^T Z_i\}, \quad i = 1,\cdots,n$, into the log (global) partial likelihood

$$l_n(\beta) = \sum_{j=1}^{N} \Bigg[[\beta^T \mathbf{Z}_{(j)}]^T \hat{\gamma}^{[1]}\{(\hat{\beta}^{[1]})^T Z_{(j)}\} - \log\Bigg\{\sum_{i\in\mathcal{R}_j} \exp\big\{[\beta^T \mathbf{Z}_i]^T \hat{\gamma}^{[1]}\{(\hat{\beta}^{[1]})^T Z_i\}\big\}\Bigg\}\Bigg],$$

and maximize it w.r.t β to get the estimate $\hat{\beta}^{[2]}$.

Repeat these two iterated steps 1 and 2, until some convergence criterion is met.

3. Asymptotic Results

Under some regularity conditions, for any $\sqrt{n}$ consistent estimator $\hat{\beta}$ of true parameter β_0, let $\hat{\psi}$ be the corresponding estimator for the true link ψ, if $h \to 0$, $nh/\log n \to \infty$, then

$$\sup_{|z|\le B} \|\hat{\psi}(\hat{\beta}^T z) - \psi(\beta_0^T z)\| \to_p 0 \tag{5}$$

A possible choice of a $\sqrt{n}$ consistent estimator can be found in Chen, Li and Wang (1999). Let $\hat{\gamma}(\hat{\beta}^T z)$ be the corresponding estimator for the derivative vector $\gamma_0(\beta_0^T z)$ of the true link ψ, if $nh \to \infty$ and nh^{2p+3} is bounded, then

$$\sqrt{nh}\Bigg\{H(\hat{\gamma}(\hat{\beta}^T z) - \gamma_0(\beta_0^T z)) - \frac{\psi^{(p+1)}(\beta_0^T z)}{(p+1)!} A^{-1} b h^{p+1}\Bigg\} \to_D N\Bigg\{0, \frac{\sigma^2(\beta_0^T z)}{f(\beta_0^T z)} A^{-1} D A^{-1}\Bigg\}. \tag{6}$$

Here $\nu_1 = \int \mathbf{u}K(u)du$, $b = \int u^{p+1}(\mathbf{u} - \nu_1)K(u)du$, $A = \int \mathbf{u}\mathbf{u}^T K(u)du - \nu_1\nu_1^T$, $D = \int K^2(u)(\mathbf{u} - \nu_1)^{\otimes 2}du$ and $\sigma^2(\beta_0^T z) = E\{\delta|\beta^T Z = \beta_0^T z\}^{-1}$. If

$$\sup_{|z|\leq B} \|\hat{\psi}''(z) - \psi''(z)\| \to_p 0,$$

then we have

$$\sqrt{n}\big(\hat{\beta} - \beta_0\big) \to_D N\bigg(0, \Sigma^{-1}(\beta_0)\bigg),$$

where $\Sigma(\beta_0)$ is positive definite.

4. Simulation Studies

To see how the algorithm works for the proposed model, we did several simulation studied. We chose a linear link function and a quadratic link function. We use the measure as standard deviation of the difference between the fitted value $\hat{\psi}(\cdot)$ and the true value of $\psi(\cdot)$ function at all data point. The results of 100 simulations based on the average value of this measure and its standard deviation are not reported here.

For the quadratic link function, our method performed much better than using the traditional Cox model. Even when the true link function is linear which corresponds to the Cox model, our estimates are still quite efficient.

REFERENCES

Breslow, N. (1974), Covariance analysis of censored survival data. *Biometrics* **80** 89-99.

Cox, D. R. (1972), Regression models and lift-table (with discussion). *J. Roy. Statist. Soc. Ser.* **4** 187-220.

Fan, J., Gijbels, I. and King, M. (1997), Local likelihood and local partial likelihood in hazard regression. *Ann. Statist.* **25** 1661-1690.

Gentleman, R. and Crowley. J. (1991), Local full likelihood estimation for the proportional hazards model. *Biometrics* **47** 1283-1296.

Chen, C.H., Li, K. C., and Wang, J. L. (1999), Dimension reduction for censored regression data. *Ann. Statist.* **27** 1-23.

O'Sullivan F. (1988), Nonparametric estimation of relative risk using splines and cross-validation. *SIAM J. Sci. Statist. Comput.* **9** 531-542.

O'Sullivan F. (1993), Nonparametric estimation in the Cox model. *Ann. Statist.* **21** 124-145.

Sleeper L. A. and Harrington D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *J. Amer. Statist. Assoc.* **85** 941-949.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82** 559-567.

RESUME