THE COX PROPORTIONAL HAZARDS MODEL WITH A PARTIALLY KNOWN BASELINE

MAI ZHOU

Department of Statistics, University of Kentucky
Lexington, KY 40506-0027, U.S.A.

The Cox proportional hazards regression model has been widely used in the analysis of survival/duration data. It is semiparametric because the model includes a baseline hazard function that is completely unspecified. We study here the statistical inference of the Cox model where some information about the baseline hazard function is available, but it still remains as an infinite dimensional nuisance parameter. We incorporate the information about the baseline hazard into the inference for regression coefficient by using the empirical likelihood method (Owen 2001) and obtained the modified test/estimator and their asymptotic distributions. The modified estimator is shown to be better than the regular Cox partial likelihood estimator in theory and in several simulations.

Some key words: Empirical likelihood; Information matrix; Log-rank test; Wilks theorem.

1. Introduction and Background

One of the most widely used regression models in survival analysis is the Cox proportional hazards model suggested by Cox (1972, 1975). Let $X_1, \ldots, X_n$ and $C_1, \ldots, C_n$ be nonnegative independent random variables. Think of C_i as the censoring time associated with the survival time X_i. Due to censoring, we can only observe (T_1, δ_1), $\ldots$, (T_n, δ_n) where

$$T_i = \min(X_i, C_i) \quad \text{and} \quad \delta_i = I_{[X_i \le C_i]}. \tag{1}$$

Also available are $z_1, \ldots, z_n$, which are covariates associated with the survival times $X_1, \ldots, X_n$ and we assume z_i do not change with time in this paper.

According to the Cox's proportional hazards model, the cumulative hazard function, $\Lambda_i(t)$, of the ith survival time is related to the covariate z_i. That relation is given by

$$\Lambda_{X_i}(t) = \Lambda_i(t) = \Lambda(t|z_i) = \Lambda_0(t) \exp(\beta_0 z_i) \ , \tag{2}$$

where β_0 is an unknown regression coefficient and $\Lambda_0(t)$ is the so called baseline cumulative hazard function. Another way to think of $\Lambda_0(t)$ is that it is the cumulative hazard function for an individual with zero covariate, $z = 0$.

The semiparametric Cox proportional hazards model assumes that the baseline cumulative hazard function $\Lambda_0(t)$ is completely unknown and arbitrary.

In this paper we study the statistical inference in the Cox model where we have some information about the baseline hazard. For example, we may know that the baseline hazard function has median equal to 40; or that the cumulative hazard is linear within the time interval $(25, 50)$. For the stratified Cox model, we may know that the two baseline cumulative hazards cross at $t = 50$, etc. In practice, when comparing a placebo against a new treatment in a two sample test, we often have additional/prior knowledge about the survival/hazard experience for the placebo group. Similarly, if a desease is well studied before then there often are some information about the baseline hazard available from prior studies. By using these information, we strike a compromise between the complete nonparametric (Cox model) and parametric models.

Empirical likelihood method is used in this paper to give inference about β_0 in the presence of this additional information. We show that the maximum empirical likelihood estimator has asymptotically a normal distribution and the (profile) empirical likelihood ratio also follows a Wilks type theorem under null hypothesis.

It is worth pointing out that in the regular Cox model, the partial likelihood estimator of β is "free" of the baseline, yet the information on the baseline does help improve the estimation of β. Our modified estimator of β is shown to be more accurate and the test have better power compared to the regular Cox partial likelihood estimator/test.

What we propose to do in this paper with the additional information is to shrink the space of the nuisance parameter, and show that this leads to improved estimation/testing for the regression parameter via empirical likelihood. Furthermore, if we keep shrinking the nuisance parameter space by adding more information about the baseline hazard, we would eventually get to the case where there is no nuisance parameter – the parametric proportional hazard model. Therefore the models and methods we study here bridges the gap between an empirical likelihood with a completely unspecified infinite dimensional nuisance parameter and a parametric likelihood with no nuisance parameter, or finite dimensional nuisance parameters. In

fact we show the Fisher information of one model approaches that of the other model with increasing knowledge about the nuisance parameter (Theorem 6).

Similar idea also work for many other semi-parametric models with infinite dimensional nuisance parameters. If there are additional information available for the nuisance parameter then by incorporating them into the empirical likelihood (by shrinking the parameter space) we can improve the estimation of the finite dimensional parameter of interest. See Chen (2005), chapters four and five for details of this approach in other models.

We end this section by presenting a few known results about the regular Cox partial likelihood estimator of the regression coefficient β_0, which can be found in Andersen and Gill (1982), and Pan (1997). For simplicity we gave detailed formula only for the case $\dim(z_i) = 1$. When $\dim(z_i) = k$, parallel results to those obtained here can be obtained similarly.

Let $\Re_i = \{j : T_j \geq T_i\}$, the risk set at time T_i. Define

$$\ell(\beta) = \sum_{i=1}^{n} \delta_i z_i - \sum_{i=1}^{n} \delta_i \frac{\sum\limits_{j\in\Re_i} z_j \exp(\beta z_j)}{\sum\limits_{j\in\Re_i} \exp(\beta z_j)}, \tag{3}$$

and

$$I(\beta_0) = \sum_{i=1}^{n} \delta_i \left(\frac{\sum\limits_{j\in\Re_i} z_j^2 \exp(\beta_0 z_j)}{\sum\limits_{j\in\Re_i} \exp(\beta_0 z_j)} - \left[\frac{\sum\limits_{j\in\Re_i} z_j \exp(\beta_0 z_j)}{\sum\limits_{j\in\Re_i} \exp(\beta_0 z_j)} \right]^2 \right) = -\ell'(\beta_0). \tag{4}$$

If $\hat{\beta}_c$ is the solution of (3), i.e. $\ell(\hat{\beta}_c) = 0$, then $\hat{\beta}_c$ is called the Cox partial likelihood estimator of the regression coefficient β_0.

Theorem 1 (Andersen and Gill 1982) *Under some mild regularity conditions we have the following results:*
(1). If $\hat{\beta}_c$ is the solution of (3), then, as $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_c - \beta_0) \xrightarrow{\mathcal{D}} N(0, \Sigma^{-1}), \tag{5}$$

where $\Sigma = Plim_{n\to\infty} \frac{1}{n} I(\beta_0)$.
(2). If $\beta_n^\star \xrightarrow{\mathrm{P}} \beta_0$, then $\frac{1}{n} I(\beta_n^\star) \xrightarrow{\mathrm{P}} \Sigma$.

Before we present the next theorem we need some definitions of the empirical likelihood. The contribution to the asymptotic (Poisson) empirical

likelihood function from (T_i, δ_i) is

$$(\Delta\Lambda_i(T_i))^{\delta_i} \exp\{-\Lambda_i(T_i)\}.$$

Under Cox's proportional hazards model,

$$\Delta\Lambda_i(T_i) = \Delta\Lambda_0(T_i)\exp(\beta z_i), \quad \text{and} \quad \Lambda_i(T_i) = \Lambda_0(T_i)\exp(\beta z_i).$$

Also, we write $\Lambda_0(T_i) = \sum_{j:T_j \le T_i} \Delta\Lambda_0(T_j)$. Thus the empirical likelihood function under the Cox's model is

$$\mathcal{A}L^c(\beta, \Lambda_0) = \prod_{i=1}^{n} (\Delta\Lambda_0(T_i)e^{\beta z_i})^{\delta_i} \exp\{-e^{\beta z_i} \sum_{j:T_j \le T_i} \Delta\Lambda_0(T_j)\}, \qquad (6)$$

where we shall require $\Lambda_0 \ll \hat{\Lambda}_{NA}$, the Nelson-Aalen estimator based on the data $(T_i, \delta_i), i = 1, 2, \cdots, n$. This requirement is similar to that of $F \ll \hat{F}_n$ for CDFs imposed by Owen, see Owen (1988) for some discussions.

Theorem 2 (Pan 1997) *Under the same conditions as in Theorem 1, we have the following empirical likelihood ratio result:*

$$-2\log \frac{\max\limits_{\{\Lambda_0:\ \Lambda_0 \ll \hat{\Lambda}_{NA}\}} \mathcal{A}L^c(\beta_0, \Lambda_0)}{\max\limits_{\{\beta,\ \Lambda_0:\ \Lambda_0 \ll \hat{\Lambda}_{NA}\}} \mathcal{A}L^c(\beta, \Lambda_0)} = I(\xi)(\beta_0 - \hat{\beta}_c)^2 \xrightarrow{\mathcal{D}} \chi^2_{(1)} \ ,$$

where ξ is between β_0 and $\hat{\beta}_c$.

2. Inference of β_0 with Information on the Baseline

The simplest form of the additional information on the baseline is given in terms of the following equation:

$$\int g(s)d\Lambda_0(s) = \sum g(T_i)\Delta\Lambda_0(T_i) = \theta \ , \qquad (7)$$

where θ is a given constant, and $g(\cdot)$ is a given function. The second expression above assumes a discrete hazard that only have possible jumps at the observed survival times, T_i (like the Nelson-Aalen estimator). This type of information includes many familiar cases. For example, if $g(s) = I_{[s \le 45]}$ and $\theta = -\log 0.5$, then the extra information corresponds to "median equal to 45".

For simplicity, we assume $T_1 < T_2 < \cdots < T_n$ for the rest of this paper. The modified Cox estimator is defined via the empirical likelihood. Let

$w_i^0 = \Delta\Lambda_0(T_i)$ for $i = 1, 2, \cdots, n$. We rewrite the log empirical likelihood (6) in terms of w_i^0:

$$\log\mathcal{A}L^c(\beta, \Lambda_0) = \sum_{i=1}^{n} \delta_i \log w_i^0 + \sum_{i=1}^{n} \delta_i \beta z_i - \sum_{i=1}^{n} w_i^0 \sum_{j=i}^{n} \exp(\beta z_j).$$

To maximize the above with respect to β and w_i^0, at the same time keep in mind of the extra information (7) imposed on the baseline hazard, we form the target function to be used by Lagrange multiplier method

$$G = \sum_{i=1}^{n} \delta_i \log w_i^0 + \sum_{i=1}^{n} \delta_i \beta z_i - \sum_{i=1}^{n} w_i^0 \sum_{j=i}^{n} \exp(\beta z_j) - n\lambda[\sum_{i=1}^{n} g(T_i)\delta_i w_i^0 - \theta] \ .$$

Taking partial derivatives of G with respect to β and w_i^0, and letting them equal to zero, we obtain

$$\frac{\partial G}{\partial w_i^0} = \frac{\delta_i}{w_i^0} - \sum_{j=i}^{n} \exp(\beta z_j) - n\lambda g(T_i)\delta_i = 0, \quad i = 1, 2, \cdots, n \tag{8}$$

and

$$\frac{\partial G}{\partial \beta} = \sum_{i=1}^{n} \delta_i z_i - \sum_{i=1}^{n} w_i^0 \sum_{j=i}^{n} z_j \exp(\beta z_j) = 0. \tag{9}$$

Solving (8), we have

$$w_i^0 = \frac{\delta_i}{\sum_{j=i}^{n} \exp(\beta z_j) + n\lambda g(T_i)\delta_i} \quad i = 1, 2, \cdots, n. \tag{10}$$

The λ in the above equation is the solution of

$$m(\beta, \lambda) = \sum_{i=1}^{n} \frac{\delta_i g(T_i)}{\sum_{j=i}^{n} \exp(\beta z_j) + n\lambda g(T_i)\delta_i} - \theta = 0. \tag{11}$$

Substituting (10) into (9), we get the equation

$$\ell^*(\beta, \lambda) = \sum_{i=1}^{n} \delta_i z_i - \sum_{i=1}^{n} \delta_i \frac{\sum\limits_{j \in \Re_i} z_j \exp(\beta z_j)}{\sum\limits_{j \in \Re_i} \exp(\beta z_j) + n\lambda g(T_i)\delta_i} = 0. \tag{12}$$

Solving (12) and (11) for β and λ simultaneously requires iterative methods. We notice that the computation for the regular Cox estimator of β needs iteration too. We shall discuss the computation in more detail in the next section. Let us use $\hat{\beta}$, $\hat{\lambda}$ to denote the solution of (12) and (11). The solution $\hat{\beta}$ is also the modified estimator of the regression coefficient.

Theorem 3 *As $n \to \infty$, the regression estimator, $\hat{\beta}$, incorporating the additional information (7) on the baseline has the following limiting distribution:*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, (\Sigma^*)^{-1}) \ ,$$

where $\Sigma^* = \Sigma + BA^{-1}B$ and

$$A = \lim A_n = \lim \sum_{i=1}^{n} \frac{\delta_i g^2(T_i)}{\left[\sum_{j \in \Re_i} \exp(\beta_0 z_j)\right]^2} \ ;$$

$$B = \lim B_n = \lim \sum_{i=1}^{n} \frac{\delta_i g(T_i) \sum_{j \in \Re_i} z_j \exp(\beta_0 z_j)}{\left[\sum_{j \in \Re_i} \exp(\beta_0 z_j)\right]^2} \ .$$

Notice the variance is smaller than that of the regular Cox estimator.

The proof of Theorem 3 is deferred to appendix.

We also have the Empirical Likelihood Theorem (Wilks) for the modified regression estimator.

Theorem 4 *Assume all the conditions of Theorem 1. In addition we assume $g(\cdot)$ is integrable. Finally assume the true baseline hazard satisfy (7). Then we have, as $n \to \infty$,*

$$-2 \log \mathcal{A}LR^c(\beta_0, \Lambda_0) = -2 \log \frac{\sup \mathcal{A}L^c(\beta_0, \Lambda_0)}{\sup \mathcal{A}L^c(\beta, \Lambda_0)} \xrightarrow{\mathcal{D}} \chi^2_{(1)} \ ,$$

where the numerator $\mathcal{A}L^c$ is maximized with β fixed at β_0 and Λ_0 satisfy (11). The denominator $\mathcal{A}L^c$ is maximized with Λ_0 satisfy (11) but β may change freely.

See appendix for proof.

Remark 2.1. If the regression coefficient β is a vector, then a similar proof will show that the limiting distribution in Theorem 4 becomes a $\chi^2_{(p)}$ where $p = \dim(\beta)$.

We may also consider the situation where the additional information is not given by (7) but by an interval type constraint,

$$C_1 \leq \sum w_i^0 g(T_i) \leq C_2$$

or equivalently replace equation (11) by $k_1 \leq m(\beta, \lambda) \leq k_2$. This is probably more practical since people are often reluctant to assume a precise value for the baseline feature, but a range is much more reasonable.

As sample size tends to infinity this type of information may not yield any improvements in estimation/testing since $\sum w_i^0 g(T_i) \to \theta$ and thus $k_1 \le m(\beta, \lambda) \le k_2$ holds with probability approach one (assuming k_1 and k_2 are fixed). But for finite samples, there is always some probability that the inequalities will be violated and the adjustment that forces the summation value into the interval $[C_1, C_2]$ will lead to improvements of the estimation for β. This means we only need to adjust the estimator when the feature of the (un-adjusted) baseline falls outside the interval and when it does, we only do minimal adjustment to pull the feature to the boundary of the interval. We call this "finite sample adjustment" in simulation section next.

The above discussion assumes the true value θ satisfy $C_1 < \theta < C_2$. If however θ equals to one of the boundaries, then the asymptotics are more complicated. If θ is outside the interval $[C_1, C_2]$, then the modified estimate/test will not be consistent.

3. Computation of the Modified Estimator and Simulations

We have modified the programs for the regular Cox model in R language (Gentleman and Ihaka 1996) to do the computations for the Cox estimator with additional information on the baseline hazard, (available at `http://www.ms.uky.edu/~mai/splus/library/coxEL`). These programs solve (by iterative method) equation (12) for a given λ value and $g(\cdot)$ function. It does not solve (11) but rather, it will give the value (13) in the output. If you happen to pick θ equal to this value then this λ also solves (11). In general, we need to solve another equation in terms of λ to obtain the λ for a given θ.

The package is called `coxEL`. The relevant function is `coxphEL()`. This function is similar to the function `coxph()` in the `survival` package. But you need to supply an extra value `lam` and a function $g(\cdot)$ when calling the function `coxphEL()`. The program will output, among other things, $\hat{\beta}$, the modified regression coefficient estimator, and the value

$$\sum_{i=1}^{n} \frac{\delta_i g(T_i)}{\sum_{j \in \Re_i} e^{\hat{\beta} z_j^*} + n\lambda g(T_i)\delta_i} = \sum \delta_i g(T_i) w_i = \int g(t) d\Lambda_0(t) \ , \qquad (13)$$

where $z_j^* = z_j - \bar{z}$.

R/Splus (also SAS) re-centers the covariates, z, automatically, therefore the baseline hazard is actually the hazard for a subject with $z = \bar{z}$ instead

of $z = 0$. If you would rather recover the constraint value for the hazard at $z = 0$, you need to multiply the value obtained in (13) by $\exp(-\hat{\beta}\bar{z})$. This is because we are in a proportional hazards model and the ratio of hazards for $z = \bar{z}$ and $z = 0$ is $\exp(\hat{\beta}\bar{z})$.

If the constraint value (13) in the output is not what you wanted, then you should adjust the value of `lam`. Keep changing the input `lam` value until you get the desired output value. This is relatively easy due to the fact that the value (13) is monotone in λ and one dimensional. In the simulation, we achieve that by calling the `uniroot( )` function in R.

For `lam`= 0, you get the regular Cox estimator, $\hat{\beta}_c$, and the constraint value (13) is the NPMLE of the integral $\int g d\Lambda_0$.

3.1. *Some Simulation Results*

We use a two sample setup and both samples are exponentially distributed, and having the same sample size. Sample one $\sim$ exp(0.2). Sample two $\sim$ exp(0.3). We use a binary covariate, z, to indicate the samples: if $z_i = 0$ then y_i is from sample one; if $z_i = 1$ then y_i is from sample two.

The risk ratio or hazard ratio is 0.3/0.2. In Cox model, this imply the true β should be $\log(0.3/0.2) = 0.4054651$.

We did not impose censoring in this simulation. The extra information we suppose we have on the baseline hazard is

$$\int \exp(-t) d\Lambda_0(t) = \theta = 0.2 \ .$$

We generated 400 such samples (each of size 400) and for each sample we computed the Cox estimator of the regression coefficient, $\hat{\beta}_c$, and the modified estimator, $\hat{\beta}$. The sample means and sample variances below are based on 400 simulation runs.

	sample mean	sample variance
Regular Cox estimator $\hat{\beta}_c$	0.4160447	0.009736113
Modified estimator $\hat{\beta}$	0.4129862	0.008310867

Table 1. Comparison of two estimators.

The simulation show $\hat{\beta}$ has smaller bias and smaller variance compared to that of $\hat{\beta}_c$. We expect the improvements for smaller samples to be more visible.

3.2. *Additional Information Given as Interval*

The extra information about the baseline may take the form $\int g(t)d\Lambda_0(t) \in [C_1, C_2]$, instead of assume we know its exact value.

In the next two simulations, we only adjust the regular Cox estimator when the value of the integration $\int g(t)d\hat{\Lambda}_0(t)$ falls outside the interval $[\theta - \epsilon, \theta + \epsilon]$ where θ is the true theoretical value of the integration (=0.2 in this case). For sample size $n = 400$, $\epsilon = 0.05$ we obtained the following results:

	sample mean	sample variance
Regular Cox estimator $\hat{\beta}_c$	0.4160447	0.009736113
Modified estimator $\hat{\beta}$	0.4085578	0.009332247

Table 2. Sample size 400. 400 simulations.

For sample size $n = 180$ (equal sample size of 90 each), $\epsilon = 0.1$, the results of 500 simulation runs gave the following table 3:

	sample mean	sample variance
Regular Cox estimator $\hat{\beta}_c$	0.4194698	0.02715997
Modified estimator $\hat{\beta}$	0.4187548	0.02708653

Table 3. Sample size 180. 500 simulations.

We see that the adjusted estimator is again having smaller bias and smaller variance. Although as ϵ increases the improvement diminished.

4. Growing information about $\Lambda_0(t)$

In this section we suppose there are many more information available about the nuisance parameter $\Lambda_0(t)$. If the additional information is given in the form of *several* equations like (7),

$$\text{for } i = 1, 2, \cdots, k; \qquad \int g_i(t)d\Lambda_0(t) = \theta_i \ , \tag{14}$$

then analysis similar to those in section two leads to

Theorem 5 *Let the maximum empirical likelihood estimator in the Cox model with additinal information (14) be denoted by $\hat{\beta}$. Under some mild regularity conditions the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$ is normal with zero mean and variance given by*

$$[\Sigma^*]^{-1} = [\Sigma + B^T A^{-1} B]^{-1} \ ,$$

where the vector B and matrix A is defined by

$$A = (A_{rs}), \quad A_{rs} = \lim \sum_{i=1}^{n} \frac{\delta_i g_r(T_i) g_s(T_i)}{\left[\sum_{j \in \Re_i} \exp(\beta_0 z_j)\right]^2} \ ;$$

$$B_m = \lim \sum_{i=1}^{n} \frac{\delta_i g_m(T_i) \sum_{j \in \Re_i} z_j \exp(\beta_0 z_j)}{\left[\sum_{j \in \Re_i} \exp(\beta_0 z_j)\right]^2} \ .$$

Next we take a closer look at the asymptotic variance of $\hat{\beta}$. Let us call $\Sigma^* = [\Sigma + B^T A^{-1} B]$ the Fisher information for β in the Cox model with additional information (14) on the baseline hazard. In view of Theorem 1, we see that the quantity $B^T A^{-1} B$ is the increment of the Fisher information due to the additional information (14) on the baseline.

When $g_i(t)$ are the indicator functions: $g_i(t) = I_{[t \le u_i]}$ for some constants $u_i,\ i = 1, 2, \cdots, k$, the increment in the Fisher information, $B^T A^{-1} B$, takes a particularly simple form (see Kim and Zhou (2002) for a proof):

$$B^T A^{-1} B = \sum_{i=1}^{k} \frac{[h(u_i) - h(u_{i-1})]^2}{V(u_i) - V(u_{i-1})} \ ,$$

where $h(u_i) = B_i$ and $V(\min(u_i, u_j)) = A_{ij}$. When $k \to \infty$ and u_i become dense, this summation will approach from below the integral

$$\int \frac{[h'(t)]^2}{V'(t)} dt \ .$$

This integration can also be written as

$$\int \frac{[h'(t)]^2}{V'(t)} dt = \lim \int \frac{[\sum z_j e^{\beta z_j} / \sum e^{\beta z_j}]^2}{n / \sum e^{\beta z_j}} d\Lambda_0(t) = \lim \sum_{i=1}^{n} \left(\frac{\sum z_j e^{\beta z_j}}{\sum e^{\beta z_j}} \right)^2 \frac{\delta_i}{n}.$$

In view of the expression of $I(\beta_0)$ in (4), we see that the Fisher information, Σ^*, in Theorem 5 can approach but never exceed the upper bound

$$\Sigma^* = [\Sigma + B^T A^{-1} B] \le \Sigma^{**}$$

with

$$\Sigma^{**} = \lim \frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{\sum_{j \in \Re_i} z_j^2 \exp(\beta_0 z_j)}{\sum_{j \in \Re_i} \exp(\beta_0 z_j)} \ .$$

The relation between Σ and Σ^{**} is like that of a variance and a second moment.

As the next lemma and theorem reveals, the expectation of this information upper bound, Σ^{**}, is identical to that of the parametric model Fisher information.

Lemma 4 *In the parametric proportional hazards model where the baseline is completely specified, the expected information for β (when there is no censoring) is*

$$I_{para}(\beta) = \sum_{i=1}^{n} z_i^2 \ .$$

With censoring, the information is

$$I_{para}(\beta) = E \sum z_i^2 H_i(\min(Y_i, C_i)) = \sum z_i^2 E H_i(\min(Y_i, C_i)) = \sum_{i=1}^{n} z_i^2 E\delta_i.$$

The proof of the lemma is straight forward and is omitted.

Theorem 6 *We have the following equality concerning the expected informations*

$$E\Sigma^{**} = \frac{1}{n} I_{para} \ ,$$

where the expectation is over all the possible ordering of the observations.

PROOF: This can be proved by induction. Assuming no censoring and for two observations, notice the probability

$$p = 1 - q = P(Y_1 < Y_2) = \frac{e^{\beta_0 z_1}}{e^{\beta_0 z_1} + e^{\beta_0 z_2}} \ .$$

We can now compute the expectation:

$$p(z_1^2 p + z_2^2 q + z_2^2) + q(z_2^2 q + z_1^2 p + z_1^2) = z_1^2 + z_2^2 \ .$$

Assume the Theorem is true for $(n-1)$ observations, then for n observations the expectation is

$$\sum p_{i1} p_{i2} \cdots p_{in} \left[z_{i1}^2 p_{i1} + \cdots + z_{in}^2 p_{in} + \sum_{i=1}^{n-1} \frac{\sum z_j^2 \exp(\beta_0 z_j)}{\sum \exp(\beta_0 z_j)} \right]$$

$$= \sum_i (z_i^2 p_i + p_i \sum_{j \neq i} z_j^2) = \sum_{i=1}^{n} z_i^2 \ .$$

For details of the induction proof in the censored data cases, see Luan (2004). $\diamondsuit$

This theorem gives us the following picture: additional information on the baseline hazard increases the Fisher information of β. These Fisher informations form a *continuous spectrum* from the completely unspecified baseline model (i.e. Cox model with Fisher information Σ) to completely specified baseline model (parametric proportional hazards model with Fisher information Σ^{**}).

The fact that the maximum empirical likelihood estimators achieve all these Fisher informations in the spectrum reinforces the view that empirical likelihood is the extension of parametric likelihood.

Appendix

Lemma 1 (Joint CLT): *Assume the same conditions as in Theorem 1. In addition, assume that $g(\cdot)$ is square integrable with respect to $\Lambda_0(\cdot)$, then we have*

$$\left[\frac{\ell(\beta_0)}{\sqrt{n}}, \sqrt{n}\cdot m(\beta_0, 0)\right] \xrightarrow{\mathcal{D}} N(0, V),$$

as $n \to \infty$ where m is defined as

$$m(\beta, 0) = \sum_{i=1}^{n} \frac{\delta_i g(T_i)}{\sum_{j\in\Re_i} e^{\beta z_j}} - \int g(t) d\Lambda_0(t) \ .$$

The variance-covariance matrix V is diagonal, $V = diag(\Sigma, V_{22})$, Σ is defined in Theorem 1 and

$$V_{22} = \lim_{n\to\infty} \int \frac{g^2(s) d\Lambda_0(s)}{\frac{1}{n}\sum_j \exp(\beta_0 z_j) I_{[T_j \geq s]}} = \lim \sum_{i=1}^{n} \frac{n g^2(T_i)\delta_i}{\left[\sum_{j\in\Re_i} \exp(\beta_0 z_j)\right]^2} \ .$$

Proof: It is now a standard result that we have the following martingale representation:

$$m(\beta_0, 0) = \sum_{i=1}^{n} \frac{\delta_i g(T_i)}{\sum_{T_j \geq T_i} e^{\beta_0 z_j}} - \int g(s) d\Lambda_0(s) = \sum_{i=1}^{n} \int \frac{g(s)}{\sum_{j=1}^{n} e^{\beta_0 z_j} I_{[T_j \geq s]}} dM_i(s)$$

and

$$\ell(\beta_0) = \sum_{i=1}^{n} \int \left(z_i - \frac{\sum_{j=1}^{n} z_j e^{\beta_0 z_j} I_{[T_j \geq s]}}{\sum_{j=1}^{n} e^{\beta_0 z_j} I_{[T_j \geq s]}} \right) dM_i(s)$$

where

$$M_i(t) = I_{[T_i \le t, \delta_i = 1]} - \int^t I_{[T_i \ge s]} \exp(\beta_0 z_i) d\Lambda_0(s)$$

with

$$\langle M_i(t) \rangle = \int_0^t I_{[T_i \ge s]} \exp(\beta_0 z_i) d\Lambda_0(s) \ .$$

Standard computation of the predictable quadratic variation process for the martingales yields

$$\left\langle \sqrt{n} m(\beta_0, 0), \ell(\beta_0)/\sqrt{n} \right\rangle = 0 \quad \text{and} \quad \left\langle \sqrt{n} m(\beta_0, 0) \right\rangle = \int \frac{g^2(s) d\Lambda_0(s)}{\frac{1}{n} \sum_j \exp(\beta_0 z_j) I_{[T_j \ge s]}}.$$

By the martingale central limit theorem, we see the Lemma is proved. $\diamondsuit$

Lemma 3 *The simultaneous solution of equations (12) and (11) has the following representation:*

$$\sqrt{n}[\hat{\beta} - \beta_0, \hat{\lambda}] = -\sqrt{n} \left[\frac{\ell(\beta_0)}{\sqrt{n}}, \sqrt{n} m(\beta_0, 0) \right] D^{-1} + o_p(1)$$

where D is a matrix

$$D = \begin{pmatrix} -I(\beta_0)/\sqrt{n} & -\sqrt{n}B \\ \sqrt{n}B & -\sqrt{n}A \end{pmatrix} \ .$$

The quantity A and B is defined in Theorem 3.

Proof: The $\hat{\beta}$ and $\hat{\lambda}$ are the solutions of $(0,0) = [\frac{\ell^*(\beta,\lambda)}{\sqrt{n}}, m(\beta,\lambda)\sqrt{n}]$. By Taylor expansion

$$[\frac{\ell^*(\beta,\lambda)}{\sqrt{n}}, m(\beta,\lambda)\sqrt{n}] = [\frac{\ell^*(\beta_0,0)}{\sqrt{n}}, m(\beta_0,0)\sqrt{n}] + (\beta - \beta_0, \lambda) \cdot D + o(|\beta - \beta_0| + |\lambda|)$$

where D is the matrix of the first derivatives of the vector. We let $\beta = \hat{\beta}$ and $\lambda = \hat{\lambda}$ in the above to get

$$(0,0) = [\frac{\ell^*(\beta_0,0)}{\sqrt{n}}, m(\beta_0,0)\sqrt{n}] + (\hat{\beta} - \beta_0, \hat{\lambda}) \cdot D + o(|\hat{\beta} - \beta_0| + |\lambda|) \ .$$

Notice $\ell^*(\beta_0, 0) = \ell(\beta_0)$, which gives

$$[\hat{\beta} - \beta_0, \hat{\lambda}] = -[\frac{\ell(\beta_0)}{\sqrt{n}}, \sqrt{n} m(\beta_0, 0)] * D^{-1} + o_p(1/\sqrt{n}) \ . \ \diamondsuit$$

Proof of Theorem 3: From Lemma 3 we have

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{\frac{\ell(\beta_0)}{\sqrt{n}} A + \sqrt{n} m(\beta_0, 0) B}{A I(\beta_0)/n + B^2} + o_p(1) \ .$$

The asymptotic normality is immediate from Lemma 1. We need to compute the asymptotic variance. Since ℓ and m are asymptotically independence (Lemma 1), we compute

$$Var(\sqrt{n}(\hat{\beta}-\beta_0)) = \lim \frac{A^2\Sigma + B^2 V_{22}}{(\Sigma A + B^2)^2} \ .$$

Since $\lim V_{22} = \lim A$, we have

$$= \frac{A}{\Sigma A + B^2} = \frac{1}{\Sigma + B^2/A} = \frac{1}{\Sigma^*} \ .$$

Notice $A > 0$, therefore we have

$$\lim Var(\sqrt{n}(\hat{\beta}-\beta_0)) = (\Sigma^*)^{-1} < \Sigma^{-1} = \lim Var\sqrt{n}(\hat{\beta}_c - \beta_0) \ . \ \diamondsuit$$

Lemma G (Graybill 1976) *Suppose* $Y \xrightarrow{\mathcal{D}} N(0,V)$ *and* M *is a symmetric matrix. Then* $YMY^T \xrightarrow{\mathcal{D}} \chi^2_p$ *if and only if* MV *is idempotent and* $rank(MV) = p$.

Proof of Theorem 4: In the Wilks theorem, the log of the empirical likelihood ratio becomes the difference of two terms. We shall compute each term separately:

Step I: We first compute the maximum of the log empirical likelihood (6) when β is fixed at β_0, and with the additional information (7).

Let $w_i^0 = \Delta\Lambda_0(T_i)$ for $i = 1, 2, \cdots, n$. We write the logarithm of $\mathcal{A}L^c(\beta_0, \Lambda_0)$ in terms of w_i^0's as follows

$$\begin{aligned} \log\mathcal{A}L^c(\beta_0,\Lambda_0) &= \sum_{i=1}^n \delta_i \log w_i^0 + \sum_{i=1}^n \delta_i\beta_0 z_i - \sum_{i=1}^n \sum_{j=1}^i w_j^0 \exp(\beta_0 z_i) \\ &= \sum_{i=1}^n \delta_i \log w_i^0 + \sum_{i=1}^n \delta_i\beta_0 z_i - \sum_{i=1}^n w_i^0 \sum_{j=i}^n \exp(\beta_0 z_j). \end{aligned}$$

To maximize the above empirical likelihood under the constraint (7) via Lagrange multiplier, we form the target function:

$$G = \sum_{i=1}^n \delta_i \log w_i^0 + \sum_{i=1}^n \delta_i\beta_0 z_i - \sum_{i=1}^n w_i^0 \sum_{j=i}^n \exp(\beta_0 z_j) - n\gamma[\sum g(T_i) w_i^0 - \theta]$$

Taking derivatives of G with respect to w_i^0 for $i = 1, 2, \cdots, n$, and letting them equal to zero, we obtain the equations

$$\frac{\partial G}{\partial w_i^0} = \frac{\delta_i}{w_i^0} - \sum_{j=i}^n \exp(\beta_0 z_j) - n\gamma g(T_i)\delta_i = 0.$$

It follows that

$$w_i^0 = \frac{\delta_i}{\sum_{j=i}^n \exp(\beta_0 z_j) + n\gamma g(T_i)\delta_i}$$

for $i = 1, 2, \cdots, n$. The value of the γ in the above can be obtained as the solution of the equation

$$0 = m(\beta_0, \gamma) = \sum_{i=1}^n \frac{g(T_i)\delta_i}{\sum_{j=i}^n \exp(\beta_0 z_j) + n\gamma g(T_i)\delta_i} - \theta \ . \tag{15}$$

The derivative of $m(\beta_0, \gamma)$ with respect to γ is always negative, so there is a unique γ solution, for the feasible values of θ.

By using the Taylor expansion on (15), it is easy to see that the solution, $\hat{\gamma}$, of (15) with $\theta = \int g(s)d\Lambda(s)$ has the following representation

$$\hat{\gamma} = m(\beta_0, 0) \times A^{-1} + o_p(1/\sqrt{n})$$

where $A = \lim A_n$ is defined in Theorem 3. We notice that $A = \lim A_n = V_{22}$.

The Hessian matrix of $\log \mathcal{A}L^c(\beta_0, \Lambda_0)$ is negative-definite so the w_i^0's provide the maximum of the log likelihood. Thus the maximized log likelihood under the extra baseline constraint is: (maximized over baseline, with β fixed at β_0)

$$\begin{aligned}
&\log \mathcal{A}L^c(\beta_0, \hat{\Lambda}(\beta_0)) \\
&= \sum_{i=1}^n \delta_i \log \frac{\delta_i}{\sum_{j=i}^n e^{\beta_0 z_j} + n\hat{\gamma} g(T_i)\delta_i} + \sum_{i=1}^n \delta_i \beta_0 z_i - \sum_{i=1}^n \frac{\delta_i \sum_{j=i}^n \exp(\beta_0 z_j)}{\sum_{j=i}^n e^{\beta_0 z_j} + n\hat{\gamma} g(T_i)\delta_i} \\
&= \sum_{i=1}^n \delta_i \beta_0 z_i - \sum_{i=1}^n \delta_i \log \left(\sum_{j=i}^n e^{\beta_0 z_j} + n\hat{\gamma} g(T_i)\delta_i \right) - \sum_{i=1}^n \frac{\delta_i \sum_{j=i}^n \exp(\beta_0 z_j)}{\sum_{j=i}^n e^{\beta_0 z_j} + n\hat{\gamma} g(T_i)\delta_i} \ ,
\end{aligned}$$

where $\hat{\gamma}$ is the solution of the equation (15).

Step II: We now compute the maximum of the log empirical likelihood without fixing the β. The extra information on the baseline hazard, (7), shall remain in effect. Recall that the maximum is achieved at $(\beta = \hat{\beta}, \lambda = \hat{\lambda})$.

Substituting β in (10) by $\hat{\beta}$, λ by $\hat{\lambda}$, we get the expression of w_i^0:

$$w_i^0 = \frac{\delta_i}{\sum\limits_{j \in \Re_i} \exp(\hat{\beta} z_j) + n\hat{\lambda} g(T_i)\delta_i}, \quad i = 1, 2, \cdots, n. \tag{16}$$

The Hessian matrix of $\log \mathcal{A}L^c(\beta, \Lambda_0)$ is negative-definite so the stationary point of $\log \mathcal{A}L^c(\beta, \Lambda_0)$ is a maximum point. Therefore we obtain the expression for the maximum of the log likelihood

$$\log \max_{\{\beta, \Lambda_0 \ll \hat{\Lambda}_{NA}\}} \mathcal{A}L^c(\beta, \Lambda_0) =$$
$$\sum_{i=1}^{n} \delta_i \hat{\beta} z_i - \sum_{i=1}^{n} \delta_i \log \left(\sum_{j=i}^{n} e^{\hat{\beta} z_j} + n\hat{\lambda} g(T_i)\delta_i \right) - \sum_{i=1}^{n} \frac{\delta_i \sum_{j=i}^{n} \exp(\hat{\beta} z_j)}{\sum_{j=i}^{n} e^{\hat{\beta} z_j} + n\hat{\lambda} g(T_i)\delta_i}.$$

If we let

$$C(\beta, \lambda) =$$
$$\sum_{i=1}^{n} \delta_i \beta z_i - \sum_{i=1}^{n} \delta_i \log \left(\sum_{j=i}^{n} \exp(\beta z_j) + n\lambda g(T_i)\delta_i \right) - \sum_{i=1}^{n} \frac{\delta_i \sum_{j=i}^{n} \exp(\beta z_j)}{\sum_{j=i}^{n} e^{\beta z_j} + n\lambda g(T_i)\delta_i},$$

and combine step I and II, we have the Wilks statistic

$$-2\log \mathcal{A}LR^c(\beta_0, \Lambda_0) = -2 \log \frac{\max_{\{\Lambda_0 \ll \hat{\Lambda}_{NA}, \Lambda_0 \text{satisfy(7)}\}} \mathcal{A}L^c(\beta_0, \Lambda_0)}{\max_{\{\beta, \Lambda_0 \ll \hat{\Lambda}_{NA}, \Lambda_0 \text{satisfy(7)}\}} \mathcal{A}L^c(\beta, \Lambda_0)}$$

$$= 2(C(\hat{\beta}, \hat{\lambda}) - C(\beta_0, \hat{\gamma})) = T_1 - T_2. \quad (say)$$

We can verify easily that for any β value we have

$$\frac{\partial C(\beta, \lambda)}{\partial \lambda}|_{\lambda=0} \quad (17)$$
$$= -\sum \frac{n\delta_i g(T_i)}{\sum e^{\beta z_j} + n\lambda g(T_i)}|_{\lambda=0} + \sum \frac{n\delta_i g(T_i) \sum e^{\beta z_j}}{(\sum e^{\beta z_j} + n\lambda g(T_i))^2}|_{\lambda=0} = 0 \quad (18)$$

and

$$\frac{\partial^2 C(\beta, \lambda)}{\partial \lambda^2}|_{\lambda=0, \beta=\beta_0}$$
$$= \sum \frac{\delta_i n^2 g^2(T_i)}{(\sum)^2} - 2\sum \frac{\delta_i n^2 g^2(T_i)}{(\sum)^2} = -n^2 \sum \frac{\delta_i g^2(T_i)}{[\sum e^{\beta_0 z_j}]^2} = -nA < 0 \ .$$

We have the following Taylor expansion

$$T_2 = 2\{C(\beta_0, 0) + \hat{\gamma} C'(\beta_0, 0) + 1/2 C''(\beta_0, 0)\hat{\gamma}^2 + o(1)\} = 2C(\beta_0, 0) - nA\hat{\gamma}^2 + o(1) \ .$$

On the other hand

$$T_1 = 2\{C(\beta_0, 0) + (\hat{\beta} - \beta_0, \hat{\lambda})(C'_\beta(\beta_0, 0), C'_\lambda(\beta_0, 0))^T + (\hat{\beta} - \beta_0, \hat{\lambda})Q/2(\hat{\beta} - \beta_0, \hat{\lambda})^T + o(1)\}$$

$$= 2C(\beta_0, 0) + 2(\hat{\beta} - \beta_0)C'_\beta(\beta_0, 0) + (\hat{\beta} - \beta_0, \hat{\lambda})Q(\hat{\beta} - \beta_0, \hat{\lambda})^T + o(1)$$

where Q is the second derivative matrix of $C(\beta, \lambda)$ at $\beta = \beta_0, \lambda = 0$.

Now we compute Q. Notice also that

$$\frac{\partial C(\beta, \lambda)}{\partial \beta}|_{\lambda=0} = \sum \delta_i z_i - \sum \delta_i \frac{\sum z_j e^{\beta z_j}}{\sum e^{\beta z_j}} = \ell(\beta) \ .$$

and

$$\frac{\partial^2 C(\beta, \lambda)}{\partial \beta^2}|_{\lambda=0} = -I(\beta).$$

Also

$$\frac{\partial^2 C(\beta, \lambda)}{\partial \lambda \partial \beta}|_{\lambda=0} = \frac{\partial^2 C(\beta, \lambda)}{\partial \beta \partial \lambda}|_{\lambda=0} = 0.$$

Thus we have a diagonal matrix Q:

$$Q = \text{diag}[-I(\beta_0), -nA].$$

Putting these all together we have

$$\begin{aligned}
&-2\log \mathcal{A}LR^c(\beta_0, \Lambda_0) \\
&= \{nA\hat{\gamma}^2 + 2(\hat{\beta} - \beta_0)C'_\beta(\beta_0, 0) + (\hat{\beta} - \beta_0, \hat{\lambda})Q(\hat{\beta} - \beta_0, \hat{\lambda})^T + o(1)\} \\
&= (\sqrt{n}\frac{m(\beta_0, 0)}{\sqrt{A}} + o_p(1))^2 + 2((\hat{\beta} - \beta_0)C'_\beta(\beta_0, 0)) \\
&\quad + \ [\frac{\ell(\beta_0)}{\sqrt{n}}, \sqrt{n}m(\beta_0, 0)]D^{-1}\sqrt{n}\frac{Q}{n}(\sqrt{n}D^{-1})^T[\frac{\ell(\beta_0)}{\sqrt{n}}, \sqrt{n}m(\beta_0, 0)]^T \\
&= [\frac{\ell(\beta_0)}{\sqrt{n}}, \sqrt{n}m(\beta_0, 0)] \cdot (I + II + III) \cdot [\frac{\ell(\beta_0)}{\sqrt{n}}, \sqrt{n}m(\beta_0, 0)]^T + o_p(1) \\
&= [\frac{\ell(\beta_0)}{\sqrt{n}}, \sqrt{n}m(\beta_0, 0)] \cdot M \cdot [\frac{\ell(\beta_0)}{\sqrt{n}}, \sqrt{n}m(\beta_0, 0)]^T + o_p(1)
\end{aligned}$$

where

$$I = \begin{pmatrix} 0 & 0 \\ 0 & 1/A \end{pmatrix} \ , \quad II = \frac{1}{A\Sigma + B^2} \begin{pmatrix} 2A & B \\ B & 0 \end{pmatrix}$$

and

$$III = \frac{1}{A\Sigma + B^2} \begin{pmatrix} -A & 0 \\ 0 & -\Sigma \end{pmatrix} \ , \qquad M = \frac{1}{A\Sigma + B^2} \begin{pmatrix} A & B \\ B & B^2/A \end{pmatrix} \ .$$

In view of the Lemma G and Lemma 1, we only need to verify two matrix properties. To this end we compute

$$M \cdot V = \frac{1}{A\Sigma + B^2} \begin{pmatrix} A\Sigma & AB \\ B\Sigma & B^2 \end{pmatrix} \ .$$

It is easy to verify that the above matrix is idempotent and has rank 1. By Lemma 1 and Lemma G we have the desired result. $\diamondsuit$

References

1. Andersen, P.K., Borgan, O., Gill, R. and Keiding, N. (1993), *Statistical Models Based on Counting Processes.* Springer, New York.
2. Chen, M. (2005). Some contributions to the empirical likelihood method. Ph.D. Thesis. Department of Statistics, University of Kentucky.
3. Cox, D. R. (1972). Regression Models and Life Tables (with discussion) *J. Roy. Statist. Soc B.*, **34**, 187-220.
4. Cox, D. R. (1975). Partial Likelihood. *Biometrika*, **62**, 269-276.
5. Gentleman, R. and Ihaka, R. (1996). R: A Language for data analysis and graphics. *J. of Computational and Graphical Statistics*, **5**, 299-314.
6. Gill, R. (1980), *Censoring and Stochastic Integrals.* Mathematical Centre Tracts 124. Mathematisch Centrum, Amsterdam.
7. Graybill, F. (1976). *Theory and Application of the Linear Model.* Wadsworth Publishing Company Inc., Belmont, California.
8. Kim, K. and Zhou, M. (2004). Symmetric location estimation/testing by empirical likelihood *Communications in Statistics: Theory and Methods* **33**, 2233-2243.
9. Luan, J. (2004). Empirical Likelihood and Right-censoring and Left-truncation Data. Ph.D. Thesis. Department of Statistics, University of Kentucky.
10. Owen, A. (1988). Empirical Likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika*, **75**, 237-249.
11. Owen, A. (2001). *Empirical Likelihood.* Chapman and Hall, London.
12. Pan, X.R. (1997). *Empirical Likelihood Ratio Method for Censored Data. Ph.D. Thesis, Univ. of Kentucky, Dept. of Statist.*
13. Thomas, D. R. and Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* **70**, 865-871.