The Annals of Statistics
1984, Vol. 12, No. 2, 730–736

ASYMPTOTIC EQUIVALENCE BETWEEN THE COX ESTIMATOR AND THE GENERAL ML ESTIMATORS OF REGRESSION AND SURVIVAL PARAMETERS IN THE COX MODEL[1]

BY KENT R. BAILEY

National Heart, Lung and Blood Institute

The usual approach to estimating regression parameters in the Cox regression model uses the partial likelihood. If the covariates are not time-dependent, the model can be stated in terms of the survival function, which allows one to derive a generalized likelihood containing both regression and survival curve parameters. It is shown that, in the absence of ties, an estimator results which is asymptotically equivalent to the partial likelihood estimator. A joint information matrix leads simply to standard errors for both regression and survival curve parameters which are asymptotically correct.

Introduction. The empirical cumulative distribution function and the Kaplan Meier product-limit estimator [12] are maximum likelihood estimates (in the generalized sense of likelihood) of the distribution function for complete or right-censored data, respectively, when the underlying distribution is completely unrestricted. Johansen [10] gives a formal argument.

The Cox model [7] implies a family of survival distributions

$$S(t \mid z) = (S_0(t))^{\exp(\beta' z)}, \tag{1}$$

if the covariates z are assumed not to vary with time. Cox assumed the arbitrary survival function S_0 to be continuous, but if this assumption is dropped, then the general maximum likelihood argument leads to a joint maximum likelihood estimate of β and S_0 based on a discrete distribution S^* which places mass only on observed death times. The likelihood function is formally identical to that presented in Prentice-Gloeckler [16], in which they discuss the grouped form of the Cox model.

It is the purpose of this note to show that a) maximization of the joint loglikelihood of β and S^* when there are no ties (see equation (2)) leads to an estimator $\hat{\beta}_{\mathrm{ML}}$ of β which is asymptotically equivalent to the partial likelihood estimator $\hat{\beta}_0$ of Cox; that b) the asymptotic joint distribution of $\hat{\beta}_{\mathrm{ML}}$ and $\hat{\Lambda}_{\mathrm{ML}}(t)[\equiv -\log(\hat{S}^*(t))]$ is equivalent to that of the estimators studied in Tsiatis [17] and Bailey [4]; and that c) the joint information matrix [see (12)–(14)], when inverted, yields an asymptotically correct covariance matrix for $\hat{\beta}_{\mathrm{ML}}$ and $\hat{\Lambda}_{\mathrm{ML}}(t)$ for arbitrary t. (This provides the simplest approach to determining

Received June 1982; revised December 1983.

[1] This work was done while the author was a visitor at Stanford University Department of Statistics, in 1980, and was supported by Grant # MCS-7919155.

AMS 1980 *subject classifications*. Primary 62E20, secondary 62F12.

Key words and phrases. Generalized likelihood, Cox regression, joint estimation, asymptotic distribution.

confidence intervals for survival curve estimates associated with the Cox model.) Thus in this case, simultaneous estimation of a large number of nuisance parameters does not lead to difficulty.

Preliminaries. Let $\{(Z_i, T_i^*, \delta_i), i = 1, \cdots, n\}$ be observed, where Z_i is the covariate vector for the ith person, T_i^* is the minimum of the survival time T_i and the censoring time τ_i, and δ_i is the indicator of observed failure. Let us assume no ties—i.e. all times of death are distinct. We make no distributional assumptions about the Z's or the τ's. Let (1) hold, and let S^* represent any discrete survival distribution which places mass only at observed death times $T_{(1)}, \cdots, T_{(k)}$. Let $\lambda_i \equiv \log[S^*(T_{(i)})/S^*(T_{(i-1)})]$ represent (−) the ith increment in $\Lambda^* \equiv -\log(S^*)$. Assuming no ties, we can express the "log likelihood function," in the sense of generalized likelihood, as

$$(2) \quad \log l(\beta, S^*(t)) = \textstyle\sum_{i=1}^k [\lambda_i \sum_{j\in R_i'} \exp(\beta' Z_j) + \log(1 - \exp[\lambda_i \exp(\beta' Z_{(i)})])],$$

where R_i' is the set of individuals at risk at $t_{(i)}$ who survive the ith epoch, and $Z_{(i)}$ is the covariate vector for the individual dying at $t_{(i)}$. Note that the ith parameter λ_i occurs only in the ith summand in (2), and is free to range in $(-\infty, 0)$. Therefore, (2) can be maximized overall by maximizing each term in (2) separately over λ_i, and then maximizing over β. The explicit solution is

$$(3) \qquad \hat{\lambda}_i = \exp(-\beta' Z_{(i)})\log(1 - c_i),$$

where $c_i \equiv \exp(\beta' Z_{(i)})/\sum_{j\in R_i} \exp(\beta' Z_j)$, and R_i denotes the risk set at $t_{(i)}$. The maximized loglikelihood in β is

$$(4) \qquad \log L(\beta, \hat{S}^*(t \mid \beta)) = \textstyle\sum_{i=1}^k [\log(c_i) + c_i^{-1}(1 - c_i)\log(1 - c_i)]$$

which may be compared with the Cox partial loglikelihood

$$(5) \qquad \log L_{\mathrm{Cox}}(\beta) = \textstyle\sum_{i=1}^k \log(c_i).$$

Notice that the second term in each summand in (4) is, to first order in c_i, independent of β. The maximizer of (4) will be shown to be asymptotically equivalent to the maximizer of (5).

The cumulative hazard function $\Lambda_0(t)$ has a (generalized) MLE given by

$$(6) \qquad \hat{\Lambda}_{\mathrm{ML}}(t) = -\textstyle\sum_{i=1}^{k(t)} \hat{\lambda}_i,$$

where $\hat{\lambda}_i$ is given by (3), evaluated at $\beta = \hat{\beta}_{\mathrm{ML}}$, and $k(t)$ is the number of deaths observed prior to t. It will be shown that this is asymptotically equivalent to the estimator

$$(7) \qquad \hat{\Lambda}_0(t) \equiv \textstyle\sum_{i=1}^{k(t)} 1/[\sum_{j\in R_i} \exp(\hat{\beta}_0' Z_j)],$$

where $\hat{\beta}_0$ represents the partial likelihood maximizer. The estimator (7) is one commonly used estimator of $\Lambda_0(t)$. (See [7]).

Asymptotic equivalence. Let $U_0(\beta)$, $-V_0(\beta)$, $U_{\mathrm{ML}}(\beta)$, and $-V_{\mathrm{ML}}(\beta)$ represent the first and second partial derivatives of (5) and (4) with respect to β. Let $D_i(\beta) \equiv \sum_{j\in R_i} \exp(\beta' Z_j)$, and let $\hat{\beta}_{\mathrm{ML}}$, $\hat{\beta}_0$, $\hat{\Lambda}_{\mathrm{ML}}$, and $\hat{\Lambda}_0$ represent, respectively, the maximizers of (4) and (5), and the estimators (6) and (7). Make the following assumptions:

A1. $\| Z_i \| < M < \infty$ for all i

A2. $n^{-1} \| \mathrm{EV}_0(\beta) \| > c_0 > 0$ for all sufficiently large n.

Note that A2 is an assumption about censoring, survival, and the dispersion of Z. Then the following theorem holds.

THEOREM 1. $n^{1/2}(\hat{\beta}_{\mathrm{ML}} - \hat{\beta}_0) \to_p 0$.

To prove Theorem 1, the following two lemmas are useful.

LEMMA 1. $n^{-1/2} \| U_{\mathrm{ML}} - U_0 \| \to_p 0$.

LEMMA 2. $n^{-1} \| V_{\mathrm{ML}} - V_0 \| \to_p 0$.

The proofs of these lemmas are straightforward, and are given in Appendix 1. As shown in Bailey [4], assumptions A1 and A2 imply that $V_0(\beta)$ is positive definite for all sufficiently large n with probability 1, and that $\hat{\beta}_0$ exists, is unique, consistent, and has the Taylor representation

$$\hat{\beta}_0 - \beta = [V^*]^{-1} U_0(\beta), \tag{8}$$

where the ith row of V^* is the ith row of V_0 evaluated at an ith intermediate point β_i^* between β and $\hat{\beta}_0$. By virtue of Lemma 2 and A2, the same consistency argument and Taylor expansion can be used for $\hat{\beta}_{\mathrm{ML}}$, so that

$$\begin{aligned} n^{1/2}(\hat{\beta}_{\mathrm{ML}} - \beta) &= [n^{-1} V_{\mathrm{ML}}^*]^{-1} [n^{-1/2} U_{\mathrm{ML}}(\beta)] \\ &= [n^{-1} V_0^* + o_p(1)]^{-1} [n^{-1/2} U_0(\beta) + o_p(1)], \end{aligned} \tag{9}$$

where we use the consistency, of $\hat{\beta}_0$ and $\hat{\beta}_{\mathrm{ML}}$, both Lemmas, A2 and the equicontinuity of V_0 and V_{ML} near β. Theorem 1 follows.

Now let us further assume that

A3. $S_0(T) \cdot n^{-1} \sum_{i=1}^{n} I\{\tau_i > T\} > c_1 > 0$ for all sufficiently large n.

Then the following theorem holds.

THEOREM 2. $\sup_{t<T} n^{1/2} | \hat{\Lambda}_{\mathrm{ML}}(t) - \hat{\Lambda}_0(t) | \to_p 0$.

To prove Theorem 2, note that this difference can be written as

$$\begin{aligned} n^{1/2} | \hat{\Lambda}_{\mathrm{ML}}(t) - \hat{\Lambda}_0(t) | &= n^{1/2} | \textstyle\sum_{i=1}^{k(t)} (\hat{\lambda}_i - \hat{D}_i^{-1}(\hat{\beta}_0)) | \\ &= n^{1/2} | \textstyle\sum_{i=1}^{k(t)} [\hat{\lambda}_i(\hat{\beta}_{\mathrm{ML}}) - \hat{D}_i^{-1}(\hat{\beta}_{\mathrm{ML}})] | \\ &\quad + n^{1/2} | \textstyle\sum_{i=1}^{k(t)} [(\hat{D}_i^{-1}(\hat{\beta}_{\mathrm{ML}}) - \hat{D}_i^{-1}(\hat{\beta}_0))] | . \end{aligned} \tag{10}$$

Let $m_i \equiv \mathrm{card}(R_i)$. By assumption A1 and (3), and if $\hat{\beta}_{\mathrm{ML}}$ is restricted to a compact set, the first RHS term in (10) has the bound

$$\text{first RHS term} = n^{1/2}O[\,|\textstyle\sum_{i=1}^{k(t)} m_i^{-2}|\,]. \tag{11}$$

By assumption A3, the RHS of (11) is $O_p(n^{-1/2})$ for $t < T$. By the consistency of $\hat{\beta}_{\mathrm{ML}}$ and $\hat{\beta}_0$, Theorem 1 and the bounded differentiability of $\hat{\Lambda}_0(t\,|\,\beta)$ in β, the second RHS term in (10) is $o_p(1)$, which proves Theorem 2.

Bailey [4] and Tsiatis [17] derive the joint asymptotic distribution of $\hat{\beta}_0$ and $\hat{\Lambda}_0(t)$, including weak convergence. Neither handles the case of fixed censoring times. This case is covered, however, by the counting process framework of Gill [see 2]. The present work shows that in this fixed censoring case, the asymptotic properties of the estimator $[\hat{\beta}_0,\ \hat{\Lambda}_0(t)]$ are enjoyed by the general ML estimator as well.

Joint information matrix. One of the most appealing features of the general ML approach to the Cox model is the natural way in which the asymptotic joint covariance matrix for $\hat{\beta}$ and $\hat{\Lambda}(t)$ can be estimated, by inverting the full second derivative matrix based on (2). The cumulative hazard function $\hat{\Lambda}_{\mathrm{ML}}(t)$ is simply the partial sum of the parameters $\hat{\lambda}_i$ defined in (3), up to the index of the last death prior to t, denoted $k(t)$. Let l_i be the ith term in (2). Then with $q_i \equiv \sum_{j\in R_i} Z_j Z_j' \exp(\beta' Z_j)$, $N_i \equiv \sum_{j\in R_i} Z_j \exp(\beta' Z_j)$, and $B_i \equiv 1 - \exp[\lambda_i \exp(\beta' Z_{(i)})]$, the second partials are:

$$\begin{aligned}
\partial^2 l_i/\partial\beta\partial\beta' &= \lambda_i(q_i - Z_i Z_i' \exp(\beta' Z_{(i)}) B_i^{-1}) \\
&\quad - \lambda_i^2 Z_i Z_i' \exp(2\beta' Z_{(i)}) B_i^{-2}(1 - B_i), \\
\partial^2 l_i/\partial\beta\partial\lambda_i &= N_i - Z_{(i)} \exp(\beta' Z_{(i)}) B_i^{-1} \\
&\quad - \lambda_i Z_{(i)} \exp(2\beta' Z_{(i)}) B_i^{-2}(1 - B_i),
\end{aligned} \tag{12}$$

and

$$\partial^2 l_i/\partial\lambda_i^2 = -\exp(2\beta' Z_{(i)}) B_i^{-2}(1 - B_i).$$

Noting from (3) that $\hat{\lambda}_i(\beta) = D_i^{-1}(\beta) + o(m_i^{-1})$, and that $B_i(\hat{\lambda}_i) = c_i$,

$$\begin{aligned}
(\partial^2 l_i/\partial\beta\partial\beta')_{|\hat{\lambda}_i} &= -D_i^{-1} q_i + O(m_i^{-1}), \\
(\partial^2 l_i/\partial\beta\partial\lambda_i)_{|\hat{\lambda}_i} &= N_i + O(1), \quad \text{and} \\
(\partial^2 l_i/\partial\lambda_i^2)_{|\hat{\lambda}_i} &= -D_i^2[1 + O(m_i^{-1})].
\end{aligned} \tag{13}$$

All cross partials involving λ_i and λ_j are zero, as noted earlier. Therefore, the inverse of the matrix of second partials of (2) evaluated at $\hat{\Lambda}_{\mathrm{ML}}$ can be approximated from (13) as follows:

$$\begin{aligned}
-I^{\beta\beta'}\,|_{\hat{\Lambda}_{\mathrm{ML}}} &= \textstyle\sum_{i=1}^{k} [\partial^2 l_i/\partial\beta\partial\beta' - (\partial^2 l_i/\partial\beta\partial\lambda_i)^2/\partial^2 l_i/\partial\lambda_i^2] \\
&= V_0^{-1}(1 + o(1)), \\
-I^{\beta\lambda_i}\,|_{\hat{\Lambda}_{\mathrm{ML}}} &= -V_0^{-1} N_i D_i^{-2} + o(m_i^{-2}), \quad \text{and} \\
-I^{\lambda_i\lambda_j}\,|_{\hat{\Lambda}_{\mathrm{ML}}} &= [\delta_{ij} D_i^{-2} + D_i^{-2} N_i' V_0^{-1} N_j D_j^{-2}][1 + O(m_i^{-1})].
\end{aligned} \tag{14}$$

It is important to note that the error terms in (14) are not probabilistic. The first line of (14) shows the consistency of $I^{\beta\beta'}(\beta, \hat{\Lambda}_{\mathrm{ML}})$ for estimating the covariance matrix of $\hat{\beta}$. Let 1_t be the vector of length k with one in the first $k(t)$ components and zero thereafter. Then $\hat{\Lambda}_{\mathrm{ML}}(t) = 1_t' \hat{\Lambda}_{\mathrm{ML}}$. The estimates of $\mathrm{cov}(\hat{\beta}_{\mathrm{ML}}, \hat{\Lambda}_{\mathrm{ML}}(t))$ based on (14) become:

$$\mathrm{cov}(\hat{\beta}_{\mathrm{ML}}, \hat{\Lambda}_{\mathrm{ML}}(t)) = -I^{\beta\Lambda'}1_t = V_0^{-1} \textstyle\sum_{i=1}^{k(t)} N_i D_i^{-2} + o(n^{-1}), \quad \text{and}$$

$$(15) \qquad \mathrm{var}(\hat{\Lambda}_{\mathrm{ML}}(t)) = -1_t' I^{\Lambda\Lambda'} 1_t = \{\textstyle\sum_{i=1}^{k(t)} D_i^{-2} + [\sum_{i=1}^{k(t)} N_i D_i^{-2}]' V_0^{-1} [\sum_{i=1}^{k(t)} N_i D_i^{-2}]\}\{1 + o(1)\}.$$

In Bailey [4], the asymptotic covariance matrix for $\hat{\beta}_0$ and $\hat{\Lambda}_0(t)$ was shown to be

$$(16) \qquad W(t) = \begin{pmatrix} E[V_0]^{-1} & E[V_0]^{-1}\Gamma(t) \\ \Gamma(t)'E[V_0]^{-1} & \psi(t) + \Gamma(t)'E[V_0]^{-1}\Gamma(t) \end{pmatrix}$$

where

$$(17) \qquad \begin{aligned} \Gamma(t) &\equiv \int_0^t g(s)\lambda_0(s)\, ds, \\ \psi(t) &\equiv \int_0^t h(s)\, d_\mu(s), \\ g(s) &\equiv \textstyle\sum_{i=1}^n Z_j \exp(\beta' Z_j) S(s \mid Z_j) / \sum_{i=1}^n \exp(\beta' Z_j) S(s \mid Z_j), \\ h(s) &\equiv 1/[\textstyle\sum_{i=1}^n \exp(\beta' Z_j) S(s \mid Z_j)]^2, \quad \text{and} \\ d_\mu(s) &\equiv -\textstyle\sum_{i=1}^n dS(s \mid Z_j). \end{aligned}$$

The formulae (16, 17) were derived in the case of no censoring, but apply to censored data with the replacement of $S(s \mid z)$ by the censored survival function everywhere in (17). The sums in (15) are the natural "estimators" of the integrals in (17) (except that β is unknown). The "consistency" of these estimates can be proved by arguments along the lines of Tsiatis [17], or Gill [9]. Substitutions of $\hat{\beta}_{\mathrm{ML}}$ into (15) preserves the asymptotics, by virtue of the consistency of $\hat{\beta}_{\mathrm{ML}}$.

Conclusion. The general maximum likelihood approach leads to a joint likelihood function (2) for β and $\Lambda_0(t)$ in the Cox model. Although there is no general theory which would lead to asymptotic results, the joint estimates obtained have the same asymptotic properties as the partial likelihood estimator and any of the commonly used estimates of the cumulative hazard function, in the case when no ties occur. While numerical studies suggest no small-sample advantage to the GML estimates, there is some economy in obtaining standard errors for the cumulative hazard function estimator. It is also of theoretical interest that the simultaneous estimation of β and a very large-dimensional nuisance parameter is possible without any asymptotic cost.

APPENDIX

Proof of Lemmas 2.1 and 2.2 The proofs will be given in the case of no censoring for convenience, but extend almost without change to the censoring case.

To prove these Lemmas, first compute U_{ML} and V_{ML} from (2) as

$$U_{\mathrm{ML}}(\beta) = -\textstyle\sum_{i=1}^{n} (Z_i - \bar{Z}_i)[c_i^{-1}\log(1 - c_i)], \tag{A.1}$$

and

$$V_{\mathrm{ML}}(\beta) = \textstyle\sum_{i=1}^{n} \{-c_i^{-1}\log(1 - c_i)V_i - [(1 - c_i)^{-1} + c_i^{-1}\log(1 - c_i)](Z_i - \bar{Z}_i)(Z_i - \bar{Z}_i)'\},$$

where

$$\bar{Z}_i \equiv \textstyle\sum_{j\in R_i} Z_j \exp(\beta' Z_j)/\sum_{j\in R_i} \exp(\beta' Z_j).$$

and

$$V_i \equiv \textstyle\sum_{j\in R_i} (Z_j - \bar{Z}_i)(Z_j - \bar{Z}_i)'\exp(\beta' Z_j)/\sum_{j\in R_i} \exp(\beta' Z_j).$$

The corresponding formulae for U_0 and V_0 are:

$$U_0(\beta) = \textstyle\sum_{i=1}^{n} (Z_i - \bar{Z}_i) \quad \text{and} \quad V_0(\beta) = \sum_{i=1}^{n} V_i. \tag{A.2}$$

Note that for $m_i > 2\exp(2\,\|\beta\|\,M)$, by assumption A1,

$$\begin{aligned} &|-c_i^{-1}\log(1 - c_i) - 1| < c_i, \quad \text{and}\\ &|(1 - c_i)^{-1} + c_i^{-1}\log(1 - c_i)| < 2c_i. \end{aligned} \tag{A.3}$$

In any case, $c_i < [1 + \exp(-2\,\|\beta\|\,M)]^{-1}$, so that

$$|-c_i^{-1}\log(1 - c_i) - 1| = O(1), \tag{A.4}$$

and

$$|(1 - c_i)^{-1} + c_i^{-1}\log(1 - c_i)| = O(1).$$

Therefore, summing over all i, with $m_0 \equiv [2\exp(2\,\|\beta\|\,M)]$,

$$\begin{aligned} &\|U_{\mathrm{ML}} - U_0\| < 2M[\textstyle\sum_{i=1}^{n-m_0} c_i + m_0 O(1)] = O(\log n), \quad \text{and}\\ &\|V_{\mathrm{ML}} - V_0\| < 4M^2[3\textstyle\sum_{i=1}^{n-m_0} c_i + m_0 O(1)] = O(\log n), \end{aligned} \tag{A.5}$$

which establishes both lemmas. When there is censoring, the same bounds (A.3) and (A.4) apply, and the sums (A.5) involve fewer terms, so the same order of n result holds.

REFERENCES

[1] AALEN, O. O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6** 701–726.

[2] ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120.

[3] BAILEY, K. (1979). The general ML approach to the Cox regression model. Unpublished Ph.D. thesis.

[4] BAILEY, K. (1983). The asymptotic joint distribution of regression and survival estimates in the Cox regression model. *Ann. Statist.* **11** 39–48.

[5] BRESLOW, N. and CROWLEY, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.* **2** 437–453.

[6] BRESLOW, N. E. (1975). Analysis of survival data under the proportional hazards model. *Int. Statist. Rev.* **43** 45–58.

[7] COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. B.* **34** 187–220.

[8] COX, D. R (1975). Partial likelihoods. *Biometrika* **62** 269–276.

[9] GILL, R. D. (1980). *Censoring and Stochastic Integrals.* Mathematical Centre Tracts 124. Amsterdam. Netherlands.

[10] JOHANSEN, S. (1978). The product limit estimator as maximum likelihood. *Scand. J. Statist.* **5** 195–199.

[11] KALBFLEISCH, J. D. and PRENTICE, R. L. (1973). Marginal likelihoods based under Cox's regression and life model. *Biometrika* **39** 267–278.

[12] KAPLAN, E. L. and MEIER, P. (1956). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–581.

[13] LEHMANN, E. L. (1953). The power of rank tests. *Ann. Math. Statist.* **24** 23–43.

[14] MEIER, P. (1980). On generalized maximum likelihood. Unpublished manuscript.

[15] NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimate based on partially consistent observations. *Econometrica* **16** 1–32.

[16] PRENTICE, R. L., and GLOECKLER, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34** 57–67.

[17] TSIATIS, A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9** 93–108.

MATHEMATICAL AND APPLIED STATISTICS BRANCH
DIVISION OF HEART AND VASCULAR DISEASES
NATIONAL HEART, LUNG AND BLOOD INSTITUTE
7550 WISCONSIN AVENUE
BETHESDA, MARYLAND 20205