Qatar Univ. Sci. J. (1995), 15 (2) : 287 - 290

THE LIKELIHOOD FUNCTION FOR A PROGRESSIVE DISEASE MODEL WITH A BIVARIATE SURVIVAL FUNCTION

By

HILMI F. KITTANI

Department of Mathematics , Faculty of Science, Qatar University, Doha - Qatar

دالــة الإمكــان لنمــوذج مــرض متقــدم ذو دالــة ثنائيــة

حــلـمـي كتــانــة

نعرض نمـوذجاً للتاريخ العادي لمرض متقـدم ، يتألف هذا النموذج من ثـلاثة مراحل مـرضية يمكن وصفهـا على شكل تـوزيع مشترك لمتغيري بقـاء عشـوائيين . يضـاف إلى هذا النمـوذج معلـومـات متغير مصاحب باستخدام نموذج المخاطرة التناسبيـة . يتناول البحث اشتقاق دالة الإمكان مع وضع الفرضيات اللازمة لهذا الإشتقاق .

ABSTRACT

A model for the natural history of a progressive disease is developed. The model has three disease states and can be expressed as the joint distribution of two survival random variables. Covariate information is incorporated into the model using proportional hazards model. The likelihood function is developed with the necessary assumptions.

KEY WORDS : Progressive disease model, Bivariate survival function, Disease states, Hazard function, Proportional hazards model, Baseline hazards.

1. INTRODUCTION

The model for the natural history of a progressive disease was introduced in a set of three papers: Albert, Gertman, Louis [1] and Albert, Gertman, Louis and Liu [2] and Louis, Albert and Heghinian [7]. Since Louis is an author in all three papers, we will refer to this model by Louis model. This model has three disease states: disease free state, preclinical state and clinical state. This model can be expressed as the joint distribution of two survival random variables X and Y, where X is the time (age) when the patient entered the preclinical state, and Y is the sojourn time in the preclinical state. For example, in cancer studies, X is the time for tumor onset and $T = X + Y$ is the time when the symptoms surface. In heart disease studies, X could be the time of getting the first heart attack, and $T = X + Y$ is the time of death of coronary heart disease, or could be the time of getting the second heart attack. In this model, X and Y are considered fixed points in a person's life, and do not change over time.

Consider a population of patients at a specific time. Associated with this population is a set of pairs (X,Y) values. This set of pairs has a probability density dunction denoted by f (x, y). Notice that allowing $X = \infty$ and $Y = \infty$ in Louis model means that f (x,y) is a mixed density with a lump of probability at infinity points and the marginal densities of X and Y are generally defective (total probability is less than unity). In this model, f (x,y) will be asumed continuous (so the lumps at infinity will have zero probability), by making the assumption that if the patient lives long enough with no other competing risks intervening in the natural history of the disease state model, then the patient will enter the preclinical state and then the clinical state eventually.

2. ASSUMPTIONS

The joint survival function for two nonnegative random variables (X, Y), given by Clayton and Cuzick [4] is

$$F(x,y) = [\, e^{\gamma\Lambda_1(x)} + e^{\gamma\Lambda_2(y)} - 1]^{-1/\gamma}$$
$$, \gamma > 0, x > 0, y > 0, \qquad (2.1)$$

where γ is an association parameter between X and Y, and Λ_1 and Λ_2 are the cumulative hazard functions for X and Y respectively. The joint density function of (X, Y) is

$$f(x,y) = (\gamma+1)\, \lambda_1(x)\, \lambda_2(y)\, e^{\gamma\Lambda_1(x)} + e^{\gamma\Lambda_2(y)}\, (D(x,y))^{-1/\gamma-2} \qquad (2.2)$$

where $D(x, y) = e^{\gamma\Lambda_1(x)} + e^{\gamma\Lambda_2(y)} - 1$,

$\gamma > 0$, $x > 0$, $y > 0$, λ_1 and λ_2 are the hazard functions associated with X and Y, respectively.

Consider a random sample of n observations (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), of the two random variables (X, Y) whose joint density function is given by equation (2.1). We partition the X axis into intervals $I_1, I_2, ..., I_M$ and the Y axis into intervals $I_1, I_2, ..., I_N$.

We will use the assumption of constant baseline hazards by [3] and [6] (i.e, $\lambda_{1i}(x) = \mu_{1i}$, $x \in I_i$ and $\lambda_{2j}(y) = \mu_{2j}$, $y \in I_j$) in the i^{th} and j^{th} intervals respectively. We model the hazard functions for the k^{th} individual whose (X,Y) values fall in rectangle $I_i \times I_j$, by assuming Cox's proportional hazards model holds for each of X and Y in each interval I_i and I_j respectively, where i = 1, . . . , M and j = 1, . . . , N. The proportional hazards model will allow us to include covariates in the model in order to study the effects of the covariates on X and Y. We assume that the vector of covariates z is p-dimensional and the same for both X and Y.

Thus the hazard functions λ_1 and λ_2 in the i^{th} and j^{th} intervals for the k^{th} individual whose observed (X,Y) value is (x_k, y_k) are defined as

$$\lambda_{1i}(x_k) = \mu_{1i} e^{\alpha' z_k},\ x_k \in I_i = (\, a_i,\ a_{i+1}\,], \qquad (2.3)$$

$$\lambda_{2j}(y_k) = \mu_{2j} e^{\beta' z_k},\ y_k \in I_j = (\, b_j,\ b_{j+1}\,], \qquad (2.4)$$

where $I_M = (a_M, a_{M+1}]$, $a_1 = 0$, $a_{N+1} = \infty$, $I_N = (b_N, b_{N+1}]$, $b_1 = 0$, $b_{N+1} = \infty$, μ_{1i} is the baseline hazard for X in I_i, i = 1,2, . . . , M and μ_{2j} is the baseline hazard for Y in I_j, j = 1, 2, . . . , N. Note that the μ_{1i} 's and μ_{1j} 's are unknown parameters to be estimated. z_k is the value of z for the k^{th} individual. $a' = (a_1, a_2, \ldots, a_p)$ are the coefficients associated with z for the failure time X, and $\beta' = (\beta_1, \beta_2, \ldots \beta_p)$ are the coefficients associated with z for the failure time Y. We will assume that the regression parameters α and β for the covariates z are constant (the same) for all intervals.

After we divide the first quadrant of XY-Plane into rectangles, then each individual observed values (x, y) will fall in one and only one rectangle.

We need to compute the cumulative hazard functions $\Lambda_1(x)$ and $\Lambda_2(y)$ associated with X and Y. First, we calculate the cumulative hazard function for the k^{th} individual whose X value falls in the i^{th} interval (assuming constant hazard over each interval) as follows.

$$\Lambda_{1i}(x_k) = \int_0^{x_k} \lambda_1(u)\, du = \int_{I_1} \lambda_{11}(u)\, du + \ldots + \int_{I_{i-1}} \lambda_{1\,i-1}(u)\, du + \int_{a_i}^{x_k} \lambda_{1\,i}(u)\, du$$
$$= \int_{I_1} \mu_{11}\, e^{\alpha' z_k} du + \ldots + \int_{I_{i-1}} \mu_{1\,i-1}\, e^{\alpha' z_k} du + \int_{a_i}^{x_k} \mu_{1\,i}\, e^{\alpha' z_k} du$$
$$= \Big[\, \sum_{r=1}^{i-1} \mu_{1r}\, (a_{r+1} - a_r) + \mu_{1i}\, (x_k - a_i) \Big]\, e^{\alpha' z_k} \qquad (2.5)$$

Similarly, the cumulative hazard function $\Lambda_2(y)$ for the k^{th} individual whose Y value, say y_k, in the j^{th} interval is

$$\Lambda_{2j}(y_k) = \Big[\, \sum_{r=1}^{j-1} \mu_{2r}\, (b_{r+1} - b_r) + \mu_{2j}\, (y_k - b_j) \Big]\, e^{\beta' z_k}\ . \qquad (2.6)$$

where $(b_{r+1} - b_r)$ is the length of the r^{th} interval..

3. LIKELIHOOD FUNCTION

In this section we will build the likelihood function when the failure times of interest (X, T), instead of (X, Y). Practically speaking, it is more advantageous to deal with the joint distribution of (X, T), where T = X + Y, rather than the distribution of (X,Y). The progressive disease model (P.D.M) has two nonnegative failure time random variables (X, T), with $X \le T$. To write the likelihood function when (X, T) are the observed variables, we cannot apply Clayton and Cuzick joint density function in this case, since $X \le T$. To apply the Clayton and Cuzick formula we model it for (X, Y) first, where X > 0 and Y > 0, as we have done earlier and then we make the transformation X = X and T = X + Y, to get the joint density function g (x, t) of (X,T) (the Jacobian is 1) as

$$g(x,t) = f(x, t-x)$$

$$= (\gamma+1)\,\lambda_1(x)\;\lambda_2(t-x)\,e^{\gamma[\Lambda_1(x)+\Lambda_2(t-x)]}\,D^{(-1/\gamma-2)}, \quad (3.1)$$

$$\Lambda_{2j}(t_k - x_k) = \Big[\sum_{r=1}^{j-1}\mu_{2r}(b_{r+1} - b_r) + \mu_{2j}(t_k - x_k - b_j)\Big]\,e^{\beta' z_k}.$$

$\gamma > 0$, $0 < x \le t$, λ_1 and λ_2 the hazard functions for X and (T − X) and Λ_1 (x) and Λ_2 (t − x) are the cumulative hazard functions for X and (T - X) respectively.

To write the likelihood function, we assume that we have n observations (X_1, T_1), (X_2, T_2), ..., (X_n, T_n) of the two random variables (X,T) whose density function g (x, t - x) is given above. The contribution to the likelihood function for the k^{th} individual whose (X,T) values falls in the I; x I_m rectangle (then T − X will fall in some interval, say j, $j \le m$) as

$$f(x_k, t_k - x_k) = \sum_{i=1}^{M}\sum_{j=1}^{N} f_{ij}(x_k, t_k - x_k)\, I[(x_k, t_k - x_k)\,\varepsilon\, I_i \times I_j]$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N} (\gamma+1)\,\mu_{1i}\,\mu_{2j}\,e^{(\alpha'+\beta')z_k}\,e^{\gamma\Lambda_{1j}(x_k)}\,e^{\gamma\Lambda_{2j}(t_k - x_k)}$$

$$\cdot\, D_{ijk}(x_k, t_k - x_k)^{(-1/\gamma-2)}\, I[(x_k, t_k - x_k)\,\varepsilon\, I_i \times I_j], \quad (3.2)$$

where

$$f_{ij}(x_k, t_k - x_k) = (\gamma+1)\,\mu_{1i}\mu_{2j}\,e^{(\alpha'+\beta')z_k}\,e^{\gamma[\Lambda_{1j}(x_k)+\Lambda_{2j}(t_k - x_k)]}\,D^{-1/\gamma-2}$$

$$D = e^{\gamma\Lambda_{1i}(x_k)} + e^{\gamma\Lambda_{2j}(t_k - x_k)} - 1.$$

Let n_{ij} = the number of individuals whose $(x_k, t_k - x_k) \in I_i \times I_j$ then

$$\sum_{i=1}^{M}\sum_{j=1}^{N} n_{ij} = n. \quad (3.3)$$

Then the overall likelihood for the n individuals is

$$L(\alpha, \beta, \gamma, \mu_1\,\mu_2, x, t)\;\prod_{k=1}^{n} f(x_k, t_k - x_k) \quad (3.4)$$

$$= \prod_{k=1}^{n}\Big[\sum_{i=1}^{M}\sum_{j=1}^{N} f_{ij}(x_k, t_k - x_k)\, I[(x_k, t_k - x_k)\,\varepsilon\, I_i \times I_j]\Big].$$

The log-likelihood becomes

$$\log L = \sum_{k=1}^{n}\log\Big[\sum_{i=1}^{M}\sum_{j=1}^{N} f_{ij}(x_k, t_k - x_k)\, I[(x_k, t_k - x_k)\,\varepsilon\, I_i \times I_j]\Big]. \quad (3.5)$$

If for example, the k^{th} individual observed values (x_k, t_k) fall in rectangle I_i x I_j then $(x_k, t_k - x_k)$ fall in I_i x I_j), then the likelihood contribution for the k^{th} individual simplifies to $f_{ij}(x_k, t_k - x_k)$, i.e

$$f(x_k, t_k - x_k) = \sum_{i=1}^{M}\sum_{j=1}^{N} f_{ij}(x_k, t_k - x_k)\, I[(x_k, t_k - x_k)\,\varepsilon\, I_i \times I_j]$$

$$= f_{ij}(x_k, t_k - x_k) = (\gamma+1)\,\mu_{1i}\,\mu_{2j}\,e^{(\alpha'+\beta')z_k}\,e^{\gamma\Lambda_{1i}(x_k)}$$

$$\cdot\, e^{\gamma\Lambda_{2j}(t_k - x_k)}\,D_{iik}^{(-1/\gamma-2)}, \quad (3.6)$$

where $D_{ijk}(x_k, t_x - x_k) = e^{\gamma\Lambda_{1i}(x_k)} + e^{\gamma\Lambda_{2j}(t_k - x_k)} - 1.$

Therefore

$$\log L = \sum_{k\varepsilon R_{11}}\log f_{11}(x_k, t_k - x_k) + \sum_{k\varepsilon R_{12}}\log f_{12}(x_k, t_k - x_k) + \ldots$$

$$+ \sum_{k\varepsilon R_{11}}\log f_{1N}(x_k, t_k - x_k) + \ldots + \sum_{k\varepsilon R_{MN}}\log f_{MN}(x_k, t_k - x_k)$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k\varepsilon R_{11}}\log f_{ij}(x_k, t_k - x_k), \quad (3.7)$$

where R_{ij} is the set of indices for those (X, T−X)s' in rectangle I_i x I_j. Substituting for $f_{ij}(x_k, t_k - x_k)$ in the above equation, we get.

$$\log L = \sum_{i=1}^{M}\sum_{j=1}^{N}\Big(n_{ij}\log[(\gamma+1)\,\mu_{1i}\mu_{2j}] + \sum_{k\varepsilon R_{ij}}\Big[(\alpha'+\beta')z_k$$

$$+ \gamma[\Lambda_{1i}(x_k) + \Lambda_{2j}(t_k - x_k)] + (-1/\gamma-2)\log D_{ijk}(x_k, t_k - x_k)\Big]\Big), \quad (3.8)$$

The above likeelihood function coincides with the likelihood function derived by Oakes in [8], where he considered the special case of Clayton and Cuzick bivariate survival function when $\Lambda_1(x) = \Lambda_2(x) = x$. This likelihood function coincides also with the likelihood function derived by [4].

This function can be maximized with respect to the parameter vector

$$\theta = (\gamma, \mu, \mu_{11}, \ldots, \mu_{1M}, \mu_{21}, \ldots, \mu_{2N}, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p),$$

where p is the number of covariates and

$\dim(\theta) = N + M + 2p + 1$.

REFERENCES

[1] **Albert, A., P.M. Gertman, and T.A. Louis,** 1978. Screening for the early detection of cancer I. The temporal natural history of a progressive disease state, *Mathematical Biosciences* 40, 1–59.

[2] **Albert, A., P.M. Gertman, and T.A. Louis, and S. Liu,** 1978. Screening for the early detection of Cancer II. The impact of screening on natural history of the disease. *Mathematical Biosciences* 40, 61–109.

[3] **Chiang, Y.K., R. J. Hardy, C.M. Hawkins, and A.S. Kapadia,** 1989. An illness-death process with timedependent covariates. *Biometrics* 45, 669–681.

[4] **Clayton, D.G. and J. Cuzick,** 1985. Multivariate generalizations of proportional hazards model. *J.R. Statistical Society Series A*, 82–117.

[5] **Cox, D.R.,** 1972. Regression models and life- tables (with discussion). *J.R. Statistical Society Series. B* 34. 187–220.

[6] **Holford, T. R.,** 1976. Life tables with concomitant information, *Biometrics* 32, 587–598.

[7] **Louis, T.A., Arthur, A., and Heghinian, S.,** 1978. Screening for the early detection of Cancer III. Estimation of disease natural history. *Mathematical Biosciences* 40, 111–144.

[8] **Oakes, D.,** 1982. A model for association inbivariate survival data. *J.R. Statistical Society Series B* 44, 412–422.