

Cause-Deleted Life Expectancy Improvement for Left and Right Censored Data

P. Adamic¹

Abstract. In this paper, an extension of the Self-Consistent Competing-Risks (SC-CR) Algorithm will be developed to estimate the cause-deleted life expectancy (CDLE) improvement. The CDLE improvement, which is an important actuarial indicator, refers to the increase in life expectancy that results when a cause of death, such as AIDS or cancer, is eliminated as a potential cause of death. The novelty of the proposed method stems from the unique characteristics of the SC-CR Algorithm, in that the method is carried out in the presence of left and right censored data, can handle the contingency of masked observations, and is fully nonparametric. The paper concludes by applying the algorithm to a real cancer data set.

Keywords: life expectancy, censoring, masking, SC-CR Algorithm, survival function

1 Introduction

Modeling the increase in life expectancy when a certain cause of death is eliminated is an important actuarial indicator. Actuaries, demographers, sociologists, and biostatisticians, to name a few, all use this concept when assessing the degree of influence certain causes of death (or failure, in a broader context) have on different populations. On a broader spectrum, the concept of cause-deleted life expectancy (or CDLE, as Newman (1987) defined it) improvement is also utilized in a wide variety of different public health initiatives. See Fang (2007) for further details.

To calculate the CDLE improvement, a multiple decrement² life table first needs to be constructed so that the probability (and cause) of death at each time point is quantified. However, in many cases the observations of death are subject to censoring, where the actual time of death is not known exactly. A common scenario is right censoring, where death is known only to occur *after* a certain point in time (or after a certain age, depending on the context). Another common scenario is left censoring, where the time of death is only known to occur *before* some known time. Following Zhou (2004), we say

¹ Corresponding Author: Dr. Peter Adamic, Department of Mathematics and Computer Science, Laurentian University, Sudbury, Ontario, Canada.

² In this paper, we will use the notions of decrement, cause, competing risk, and risk interchangeably, and assume that they are all mutually equivalent.

that observed failure times are subject to double censoring when both left and right censoring are present. Furthermore, it is also common that the cause of failure is not known precisely — a dynamic that is referred to as *masking*. The purpose of this paper is to propose a method whereby the CDLE improvement can be calculated in a statistically viable manner when failure time data are subject to left and right censoring (in addition to exact observations) as well as masking. To this end, the SC-CR Algorithm of Adamic (2008) will be used as the primary vehicle to estimate the CDLE improvement. All of the advantages of the SC-CR Algorithm will be retained, stemming from the fact that it is a fully nonparametric method that produces self-consistent estimators of the cause-specific mortality rates. This is the only distribution-free method currently available in the actuarial literature that can be employed to model multiple risks in the presence of censoring and masking.

In the cause-deleted framework, there are essentially only two decrements competing with one another: the cause that is being deleted, and all other causes. Thus, this is essentially a two-decrement problem that must be tackled using multiple decrement theory. The first step is to derive the cumulative incidence functions for each of these two risks. This will be accomplished using the SC-CR Algorithm, since the data are assumed to doubly censored. The next step is to generate the associated single risk survival functions, as defined in Bowers *et al.* (1997), for each risk, using standard actuarial assumptions. Finally, once the associated single risk survival functions are obtained, the life expectancies for both risks can be computed and the CDLE improvement can be obtained. Once the theory is presented, a real cancer data set will be modeled and analyzed.

2 Definitions, Notation, and Assumptions

A familiarity with the notation and methodology associated with the SC-CR Algorithm, as found in Adamic (2008), is necessary to appropriate the ensuing theory. Only concepts not previously defined will be formally stated here. The International Actuarial Notation (IAN) guidelines will also be adopted throughout the whole paper, as summarized in Bowers *et al.* (1997).

Definition 2.1. *The life expectancy for an individual currently aged x quantifies the number of years the indi-*

vidual is expected to survive beyond x . The life expectancy will be denoted by \hat{e}_x .

If we let T denote the future lifetime random variable, then,

$$\hat{e}_x = E[T] = \int_0^\infty t \cdot f(t) dt = \int_0^\infty t \cdot {}_t p_x \mu_{x+t} dt,$$

where \hat{e}_x is also called the *complete expectation of life*. A similar measure of life expectancy that is used in discrete time domains is the *curtate expectation of life*, denoted by e_x . If R denotes the future lifetime in full years only, then,

$$e_x = E[R] = \sum_{r=0}^{\infty} r \cdot f(r) = \sum_{r=0}^{\infty} r \cdot {}_r p_x q_{x+r},$$

will capture the expected number of complete years remaining beyond age x . Clearly, $\hat{e}_x \geq e_x$. It can also be shown that $e_x = \sum_{k=1}^{\infty} k p_x$ (see Bowers *et al.* (1997) for details).

A common simplifying assumption in multiple decrement theory is the uniform distribution of decrement (UDD) assumption. Under this assumption, it is postulated that over a unit time interval,

$${}_t q_x^{(j)} = t \cdot q_x^{(j)}, \text{ and } {}_t q_x^{(\tau)} = t \cdot q_x^{(\tau)}, \quad \forall 0 < t \leq 1.$$

This assumption of uniformly distributed decrements over each time period (or age) should only be used when there is little evidence of a strong seasonal pattern within each time period. The UDD assumption is fairly safe when dealing with human mortality. Under the UDD assumption over each year, we have the intuitive result that $\hat{e}_x = e_x + \frac{1}{2}$, which is explicitly proven in Brown (1997).

Definition 2.2. *The cause-deleted life expectancy (CDLE) improvement quantifies, in years, the increase in life expectancy in the event that a particular cause of death is completely eliminated. Following the notation of Brown (1997), when cause j is eliminated, the CDLE improvement at age x equals $\hat{e}_x^{(-j)} - \hat{e}_x$.*

Notice that under a UDD assumption over each year of age, $e_x^{(-j)} - e_x = \hat{e}_x^{(-j)} - \hat{e}_x$, implying that the CDLE improvement is the same regardless of whether or not the curtate expectation of life method is used or the complete expectation of life.

To calculate the life expectancy when cause j is deleted, we use the well known formula,

$$\hat{e}_x^{(-j)} = \int_0^\infty {}_t p_x'^{(-j)} dt,$$

where ${}_t p_x'^{(-j)}$ is the associated survival function if cause j is eliminated as a competing risk. As noted in Bowers *et al.* (1997), ${}_t p_x'^{(-j)}$ is not technically a valid survival function since it is not required that the limit of ${}_t p_x'^{(-j)}$, as

$t \rightarrow \infty$, goes to 0. However, ${}_t p_x'^{(-j)}$ certainly behaves as a valid survival function up to the maximum observed age, which is sufficient for calculating the CDLE improvement $\forall x$.

The associated survival function in the absence of j can be expressed in terms of the cause-deleted hazard function as,

$${}_t p_x'^{(-j)} = \exp\left(-\int_0^t h_x^{(-j)}(s) ds\right),$$

which also implies the well known result,

$${}_t p_x^{(\tau)} = {}_t p_x'^{(-j)} {}_t p_x'^{(j)}.$$

In the next section, a discrete formula for ${}_t p_x'^{(i)}$, where $i = j, -j$, or τ , will be developed by invoking the UDD assumption. This result will then be integrated into the SC-CR scheme.

It must be emphasized here that a CDLE improvement calculation assumes that each of the multiple decrements are independent of one another — or, at a minimum, that the decrement being deleted is independent of all other decrements. If this were not the case, and the cause being deleted was highly correlated with some other decrement(s), then the gain in life expectancy would actually be higher than calculated. For example, if the probability of dying in a car accident is correlated to the probability of dying from diseases of the liver (due to, say, the confounding factor of excessive alcohol consumption), then the elimination of deaths due to liver diseases, which would come about, in part, from decreased alcohol consumption, would in turn produce less alcohol related car accidents, which would decrease the overall probability of dying in a car accident. Examples of this sort of dependence are varied and quite numerous.

In short, the assumption of independence between risks provides a conservative estimate of the CDLE improvement, which can also be thought of as a lower bound on the true CDLE improvement. The only time this would not be true is when the risks are negatively correlated - a contingency that is fairly rare. Generally speaking then, if the cause being deleted is not significantly correlated with other modes of decrement, the calculated improvement in life expectancy will be a reliable estimator of the actual improvement.

3 The SC-CR Algorithm for finding the CDLE Improvement

3.1 Developing the Model

The first step in the model building process is to invoke the SC-CR Algorithm, as found in Adamic (2008) and reproduced below. The steps, statements and logic of the algorithm are directly taken and/or generalized from the univariate algorithm as found in Klein & Moeschberger (1997).

The SC-CR Algorithm

Step 0: Produce legitimate³ initial estimates of the overall survival probabilities at each t_r , ${}_t p_0^{(\tau)}$. Also, find initial estimates for the probability of failing during each time interval by cause j , ${}_{t_r-t_{r-1}} q_{t_{r-1}}^{(j)}$, for $r = 1, \dots, m$. Ignore the left-censored observations when calculating these estimates.

Step 1: Using the current estimates of ${}_t p_0^{(\tau)}$ and ${}_{t_r-t_{r-1}} q_{t_{r-1}}^{(j)}$, calculate estimates of the conditional probabilities $e_{ir}^{(j)} = P[t_{r-1} < X^{(j)} \leq t_r | X^{(\tau)} \leq t_i]$, where $X^{(j)}$ represents the event of failure due to cause j , using,

$$\hat{e}_{ir}^{(j),K} = \frac{{}_{t_r-t_{r-1}} \hat{p}_0^{(\tau),K} \cdot {}_{t_r-t_{r-1}} \hat{q}_{t_{r-1}}^{(j),K}}{1 - {}_{t_i} \hat{p}_0^{(\tau),K}}, \quad \forall r \leq i.$$

Step 2: Using the results of the previous step, estimate the number of cause-specific failures at time t_i by using,

$$\hat{d}_i^{(j),K} = d_i^{(j)} + \sum_{i=r}^m c_i \hat{e}_{ir}^{(j),K}.$$

Step 3: Compute ${}_{t_r-t_{r-1}} \hat{q}_{t_{r-1}}^{(j)} \forall r = 1, \dots, m$, using a generalized cause-specific Product-Limit estimator based on the current estimates of $\hat{d}_i^{(j)}$. If,

$$\sup_{t_r, j} \left| {}_{t_r-t_{r-1}} \hat{q}_{t_{r-1}}^{(j),K} - {}_{t_r-t_{r-1}} \hat{q}_{t_{r-1}}^{(j),K-1} \right| < \epsilon,$$

(jointly for all of the time points t_i and causes j , given some small predetermined $\epsilon > 0$), stop the algorithm; otherwise return to Step 1.

The goal of this paper is to estimate the improvement of life expectancy when a cause of death is eliminated. To accomplish this, we need associated single risk survival functions for each of the two risks in question: the cause being deleted, and all causes taken together, in aggregate. In essence, a “bridge” needs to be constructed to generate the associated single risks survival functions. The uniform distribution of decrements (UDD) assumption over each year, or the constant hazard assumption over each year, will give the same result at the end of each discrete time point (a year in this case) thus providing the necessary bridge to obtain the associated single risk survival functions that are needed. Note that these assumptions apply to each decrement individually, as well as the decrements taken collectively.

Under the UDD assumption⁴ we have,

³ As noted by Day (2005), a generalization of the Karush-Kuhn-Tucker conditions from optimization theory can be used to derive legitimate starting points for Turnbull’s (1974) algorithm that will guarantee convergence. These conditions could easily be extended to the SC-CR Algorithms, if necessary.

⁴ Subsequent derivation of Equation (1) taken mostly from Bowers et al. (1997).

$${}_t q_x^{(j)} = t \cdot q_x^{(j)}, \text{ and } {}_t q_x^{(\tau)} = t \cdot q_x^{(\tau)}, \quad \forall 0 < t \leq 1.$$

Since,

$$h_x^{(j)}(t) = \frac{f(t, j)}{{}_t p_x^{(\tau)}}, \quad {}_t q_x^{(j)} = \int_0^t f(s, j) ds,$$

the following expression can be developed for $h_x^{(j)}(t)$:

$$\begin{aligned} h_x^{(j)}(t) &= \frac{1}{{}_t p_x^{(\tau)}} \left[\frac{d}{{dt}} {}_t q_x^{(j)} \right], \\ &= \frac{1}{{}_t p_x^{(\tau)}} \left[\frac{d}{dt} (t \cdot q_x^{(j)}) \right], \\ &= \frac{q_x^{(j)}}{{}_t p_x^{(\tau)}}. \end{aligned}$$

Utilizing this result on the second line below provides for a useful expression for ${}_t p_x^{(j)}$:

$$\begin{aligned} {}_t p_x^{(j)} &= \exp \left(- \int_0^t h_x^{(j)}(s) ds \right) \\ &= \exp \left(- \int_0^t \frac{q_x^{(j)}}{{}_s p_x^{(\tau)}} ds \right) \\ &= \exp \left(- \int_0^t \frac{q_x^{(j)}}{1 - s q_x^{(\tau)}} ds \right) \\ &= \exp \left(- q_x^{(j)} \int_0^t \frac{1}{1 - s q_x^{(\tau)}} ds \right) \\ &= \exp \left(\frac{q_x^{(j)}}{q_x^{(\tau)}} \ln[1 - t q_x^{(\tau)}] \right) \\ &= \exp \left(\frac{q_x^{(j)}}{q_x^{(\tau)}} \ln[{}_t p_x^{(\tau)}] \right) \\ &= \left({}_t p_x^{(\tau)} \right)^{\frac{q_x^{(j)}}{q_x^{(\tau)}}}. \end{aligned} \quad (1)$$

Equation (1) holds in the case where $t = 1$, which is the only interval length that will be considered here. However, a generalization of Equation (1) can easily be derived for $t > 1$, the result being,

$${}_t p_x^{(j)} = \left({}_t p_x^{(\tau)} \right)^{\frac{{}_t q_x^{(j)}}{{}_t q_x^{(\tau)}}}, \quad \forall t > 1. \quad (2)$$

Interestingly, Equations (1) and (2) can also be derived under the piecewise-constant hazard assumption, where the hazard is constant over each distinct time period.

The question of the nature of the estimators used in the exponent, $\left(\frac{{}_t q_x^{(j)}}{{}_t q_x^{(\tau)}} \right)$, still needs to be addressed. The following theorem is useful in this regard.

Theorem 1: Assume that the SC-CR Algorithm for doubly-censored data converges for all values of $t_i = x$ and $t_{i+1} - t_i = t$. Let the iterations of the generalized algorithm be indexed by $K = 1, \dots, m$. Then,

$$\frac{{}_t q_x^{(j),K}}{{}_t q_x^{(\tau),K}} = \frac{{}_t q_x^{(j),K+1}}{{}_t q_x^{(\tau),K+1}} \quad \forall j, x, K.$$

Proof: First, note that $\frac{{}_tq_x^{(j),K}}{{}_tq_x^{(\tau),K}} = \frac{{}_td_x^{(j),K}}{{}_td_x^{(\tau),K}} \quad \forall j, x, K$, since the number at risk (i.e. the denominator) will be the same when calculating either ${}_tq_x^{(j),K}$ or ${}_tq_x^{(\tau),K}$. Now, the algorithm has the property that,

$$\frac{{}_td_x^{(j),K+1}}{{}_td_x^{(\tau),K+1}} = \frac{{}_td_x^{(j),K} + \alpha^{(j)}}{{}_td_x^{(\tau),K} + \beta},$$

where $\left(\frac{\alpha^{(j)}}{\beta}\right)$ captures the relative proportion of left-censored observations allocated to the cumulative number of failures for risk j . Since the left-censored observations are always assigned based on the failure probabilities of the latest iteration, it follows that,

$$\frac{\alpha^{(j)}}{\beta} = \frac{{}_td_x^{(j),K}}{{}_td_x^{(\tau),K}}.$$

For ease of notation, let ${}_td_x^{(j),K} = \omega^{(j)}$ and ${}_td_x^{(\tau),K} = \phi$. Then,

$$\frac{\omega^{(j)} + \alpha^{(j)}}{\phi + \beta} = \frac{\omega^{(j)} + \frac{\beta\omega^{(j)}}{\phi}}{\phi + \frac{\alpha^{(j)}\phi}{\omega^{(j)}}} = \frac{\omega^{(j)}(1 + \frac{\beta}{\phi})}{\phi(1 + \frac{\alpha^{(j)}}{\omega^{(j)}})} = \frac{\omega^{(j)}}{\phi},$$

since $\frac{\beta}{\phi} = \frac{\alpha^{(j)}}{\omega^{(j)}}$. This completes the proof.

Returning to the original question, it is now noted that the fraction $\frac{{}_tq_x^{(j)}}{{}_tq_x^{(\tau)}}$ is constant for every iteration K . So, if the constant (piecewise) hazard or uniform distribution assumption is considered given, then at convergence,

$${}_t\hat{p}_x^{(j)} = \left({}_t\hat{p}_x^{(\tau)}\right)^{\frac{{}_tq_x^{(j)}}{{}_tq_x^{(\tau)}}} = \left({}_t\hat{p}_x^{(\tau)}\right)^Q,$$

where Q is a constant for every iteration. This establishes the fact that we can safely use the associated single risk survival probabilities that are derived from the converged double decrement probabilities.

3.2 Illustrative Numerical Example

To illustrate the method, let us consider the following sample data set. We can verify empirically the result from Theorem 1 in Table 1 which shows the results from an implementation the SC-CR Algorithm. At the beginning of the second iteration, the estimated number of failures due to Cause 1 at time t_2 was 10.6258 and the estimated number of failures due to all other causes was 95.6322. The proportion of failures that were due to Cause 1 was $10.6258/(10.6258+95.6322) = 0.1$. Now, consider the estimated failure rates for iteration 4. The fraction $10.6168/(10.6168+95.5515)$ at time t_2 again equals 0.1, as it must. The same will hold true for all values of t_i . The CDLE improvement from time 0, once convergence was achieved at iteration 4 (assuming an epsilon of 0.0001), was equal to 1.0570.

Table 1. Successive Iterations (Double Censoring, Complete Masking), Iterations 1, 2, & 4

t_i	Left	Cause 1	Other	Right	${}_t\hat{p}_0^{(c1)}$	${}_t\hat{p}_0^{(-c1)}$
1	5	20	60	15	0.9306	0.8059
2	5	10	90	20	0.8775	0.4748
3	10	15	85	5	0.5558	0.0357
1	5	22.4234	67.2701	15	0.9263	0.7948
2	5	10.6258	95.6322	20	0.8723	0.4629
3	10	15.6073	88.4413	5	0.5494	0.0337
1	5	22.4776	67.4329	15	0.9261	0.7943
2	5	10.6168	95.5515	20	0.8721	0.4625
3	10	15.5882	88.3330	5	0.5493	0.0337

4 Application to a Cancer Data Set

We illustrate the method using real data. The data set gives the number of deaths (at each age) for the citizens of Denver, Colorado, for the year 2006. The data was obtained from the Colorado Department of Public Health and Environment. The failures are subject to both left and right censoring. Although data is available for each age, in the interest of space, it will be summarized as follows.

Table 2. Summary of the Denver Cancer Data

t_i	Left	Non-Cancer	Cancer	Right
(0-25]	0	148	0	2
(25-35]	7	91	3	2
(35-45]	3	170	14	4
(45-55]	4	333	78	0
(55-65]	0	356	156	7
(65-75]	2	388	183	13
(75-85]	8	880	249	4
(85-95]	6	802	106	33

The modified SC-CR Algorithm was run, and the associated single risk survival probabilities for the cancer and non-cancer risks were obtained. Figure 1 depicts the resulting survival functions with and without the cancer decrement.

The improvement in life expectancy at age zero was calculated for each iteration. The results were: {2.9983465, 2.9956080, 2.9954872, 2.9954824, 2.9954822, 2.9954821}. After six iterations, the life expectancy stabilized for eight significant digits. The improvement at each iteration is small, due to the presence of very light censoring in the original data set.

In short, the life expectancy of the citizens of Denver, estimated to be roughly seventy-four years age (from birth), would be expected to rise to about seventy-seven years of age, if all deaths due to cancer were eliminated. In Denver in 2006, from Table 2, about one in every five deaths was due to cancer. Usually, the theoretical improvement in life expectancy from the elimination of all

Figure 1. Converged Survival Functions for All Causes and All Causes minus Cancer

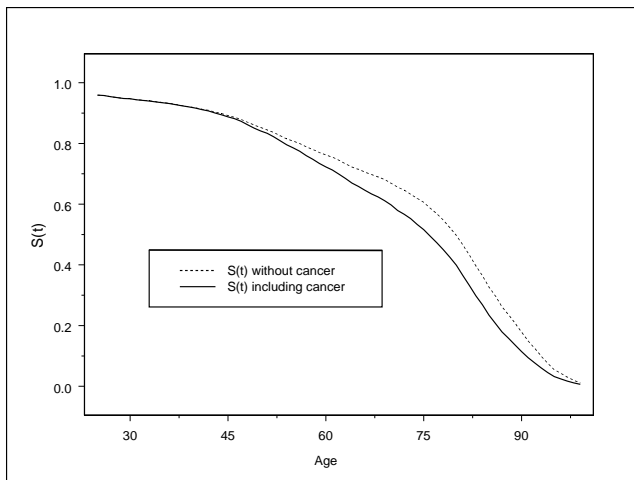
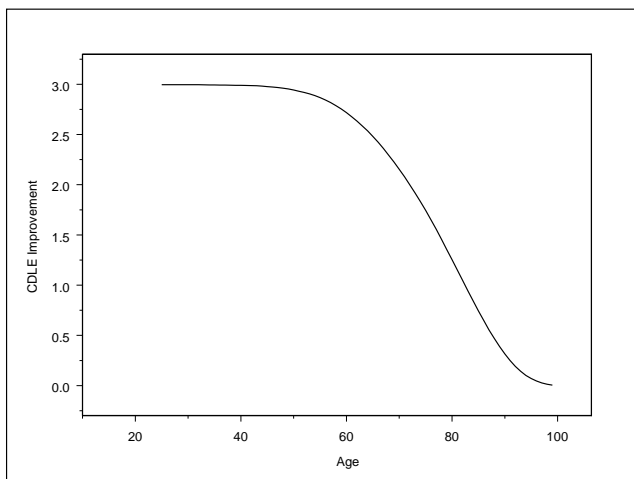


Figure 2. CDLE Improvement at Each Age



cancers is typically in the neighborhood of about three to five years (see Fang (2007) for details). Once the cause-

deleted survival probabilities are known for each age, the CDLE improvement can also be calculated at each age. Figure 2 shows the results for the Denver cancer data. It is interesting to note that, even at age 80, a life expectancy improvement of at least one year can be realized if cancer was eliminated as a potential cause of death.

ACKNOWLEDGEMENTS

I wish to thank the Colorado Department of Public Health and Environment for providing the cancer data set. I also wish to thank Dr. Hafida Boudjellaba of Laurentian University for her helpful suggestions.

REFERENCES

- [1] Adamic, P. (2008). Modeling Multiple Risks in the Presence of Double Censoring, *The Scandinavian Actuarial Journal* (DOI: 10.1080/03461230802420603).
- [2] Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A., Nesbitt, C.J. (1997). *Actuarial Mathematics, 2nd ed.* (Schaumburg, Illinois: The Society of Actuaries).
- [3] Brown, R.L. (1997). *Introduction to the Mathematics of Demography, 3rd ed.* (Winsted, CT, USA: Actex).
- [4] Day, B. (2005). Distribution Free Estimation with Interval Censored Contingent Valuation Data: Troubles with Turnbull?, http://www.uea.ac.uk/env/cserge/pub/wp/edm/edm_2005_07.pdf (Accessed December 23, 2008).
- [5] Fang, R. (2007). Life expectancy as a measure of population health: Comparing British Columbia with other Olympic and Paralympic Winter Games host jurisdictions (Summary Report), <http://www.phsa.ca/NR/rdonlyres/76D687CF-6596-46FE-AA9A-A536D61FB038/23536/PHSAreportlifeexpectancy.pdf> (Accessed March 14, 2009).
- [6] Klein, J.P., Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data* (New York: Springer).
- [7] Lawless, J.F. (1982), 2nd ed. (2003). *Statistical Models and Methods for Lifetime Data* (New York: John Wiley and Sons).
- [8] Newman, S. (1987). Formulae for Cause-deleted Life Tables, *Statistics in Medicine*, 6, 527-528.
- [9] Turnbull, B.W. (1974). Nonparametric Estimation of a Survivorship Function with Doubly Censored Data, *Journal of the American Statistical Association*, 69, 169-173.
- [10] Zhou, M. (2004). Nonparametric Bayes Estimator of Survival Functions for Doubly/Interval Censored Data, *Statistica Sinica*, 14, 533-546.