行政院國家科學委員會專題研究計畫　期中進度報告

含有測量誤差的 Cox 迴歸模型使用過量參數化的估計方法研究(1/2)

計畫類別：個別型計畫
計畫編號：NSC94-2118-M-032-013-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：淡江大學數學系

計畫主持人：黃逸輝

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 5 月 30 日

Midterm Report –NSC 94-2118-M-032-013-

Estimation in Cox proportional hazard model with measurement error and without extra information

Y.H. HUANG

Department of Mathematics, Tamkang University, Taipei county, Taiwan

email: yhhuang@mail.tku.edu.tw

Summary: When covariate in survival data are subject to measurement error, the estimation in Cox regression usually requires repeat measurements or extra information about the measurement error to proceed. Without such information, it seems that the naive analysis is the only choice. In this paper, we confirm the possibility of consistent estimation under the scenario of no extra information. Without extra information, we need to construct an additional constraint to determine the measurement error's variance, thus we apply the technique of over-parameterization to a weighted corrected score to obtain enough estimating equations. Since these estimating equations are zero-unbiased, the resultant estimates are consistent. A small simulation is conducted to assess the performance of our estimators.

Key words: Cox regression; extra information; over-parameterization; measurement error

1 Notation and the conventional approaches when measurement error's variance is known

Denote the failure time, censoring time and noncensoring indicator by T_i, C_i and δ_i, respectively. Where δ_i is 1 if $T_i \leq C_i$ and is 0 otherwise. Also let Z_i denote the true covariate of the ith individual and is assumed a scalar variable for simplicity. The Cox proportional hazard regression model assumes that the hazard function of the life time distribution has the form

$$\lambda(t; Z_i) = \lambda_0(t)e^{\beta Z_i}, t \geq 0, \tag{1.1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function.

Let $R_i = \{j : T_j \geq T_i, C_j \geq T_i\}$ be the risk set at time T_i, then the standard inference for the regression parameter is based on the partial likelihood

$$L(\beta) = \prod_{i=1}^{n} [\frac{e^{\beta Z_i}}{\sum_{j \in R_i} e^{\beta Z_j}}]^{\delta_i}, \tag{1.2}$$

which has the derivative—the partial score function

$$S(\beta) = \sum_{i=1}^{n} \delta_i \{Z_i - \frac{\sum_{j \in R_i} Z_j e^{\beta Z_j}}{\sum_{j \in R_i} e^{\beta Z_j}}\}. \tag{1.3}$$

For the measurement error model, we assume the measurement error e_i is additive and i.i.d. $N(0, \sigma^2)$ distributed. Let $X_i = Z_i + e_i$ denote the observable surrogate. The naive analysis ignores the measurement error and uses

$$S_0(\beta) = \sum_{i=1}^{n} \delta_i \{X_i - \frac{\sum_{j \in R_i} X_j e^{\beta X_j}}{\sum_{j \in R_i} e^{\beta X_j}}\}. \tag{1.4}$$

for solving estimate of β. As in other regression problems, the naive approach results biased estimate as expected.

A corrected score of (1.3) is

$$S_1(\beta) = \sum_{i=1}^{n} \delta_i \{X_i + \beta\sigma^2 - \frac{\sum_{j \in R_i} X_j e^{\beta X_j}}{\sum_{j \in R_i} e^{\beta X_j}}\}, \tag{1.5}$$

which was proposed by Nakamura (1992). Since no unbiased corrected score exist, Nakamura proposed (1.5) as an approximate unbiased estimating function and also provided further correction of (1.5) based on $2nd$ order correction. However, some recent work discover that (1.5) is asymptotically unbiased. Augustin (2004) showed that (1.5) is the corrected score of the score function derived from the differentiation of Breslow's likelihood (Breslow 1972, 1974).

Another unbiased estimating function can be derived through conditioning. The conditional score in generalized linear model was developed by Carroll, Ruppert and Stefanski (chp. 6, 1995). And the conditional score in the Cox proportional hazard model was developed by Tsiates and Davidian (2001), which is equivalent to the following estimating function

$$S_2(\beta) = \sum_{i=1}^{n} \delta_i \{X_i + \beta\sigma^2 - \frac{\sum_{j\in R_i, j\neq i} X_j e^{\beta X_j} + (X_i + \beta\sigma^2) e^{\beta(X_i+\beta\sigma^2)}}{\sum_{j\in R_i, j\neq i} e^{\beta X_j} + e^{\beta(X_i+\beta\sigma^2)}}\}. \qquad (1.6).$$

It is known that these two conventional approaches–Nakamura first order corrected score and conditional score are asymptotically equivalent, but the conditional approach may perform better in finite samples (Song and Huang, 2005).

2 Weighted and reweighted corrected score functions

Recall that the original partial score function when $Z_i's$ are observed is

$$S(\beta) = \sum_{i=1}^{n} \delta_i \{Z_i - \frac{\sum_{j\in R_i} Z_j e^{\beta Z_j}}{\sum_{j\in R_i} e^{\beta Z_j}}\}.$$

As noted in Nakamura (1992), the exact corrected score of $S(\beta)$ does not exist. If we apply weights "$\sum_{j\in R_i} e^{\beta Z_j}$" to each summand, than $S(\beta)$ becomes

$$\sum_{i=1}^{n} \delta_i [Z_i \sum_{j\in R_i} e^{\beta Z_j} - \sum_{j\in R_i} Z_j e^{\beta Z_j}]. \qquad (2.1)$$

It is easy to show that (2.1) is also zero unbiased (Huang, 2004). Note that unlike $S(\beta)$, there is no fraction term in (2.1), and it is very easy to find a corrected version of (2.1). By the normal assumption of e_i, the estimating function

$$\sum_{i=1}^{n} \delta_i [X_i \sum_{j \in R_i} e^{\beta X_j - \frac{1}{2}\beta^2\sigma^2} - \sum_{j \in R_i} (X_j - \beta\sigma^2) e^{\beta X_j - \frac{1}{2}\beta^2\sigma^2} - \beta\sigma^2 e^{\beta X_i - \frac{1}{2}\beta^2\sigma^2}] \tag{2.2}$$

is a corrected version of (2.1) and has conditional expectation (2.1) when condition on $Z_i's$ and $R_i's$. Though (2.2) is zero-unbiased, it is less efficient since it estimates the weighted partial score but not the original partial score. Hence we divide each summand in (2.2) by $\sum_{j \in R_i} e^{X_j - \frac{1}{2}\beta^2\sigma^2}$ which is a predictor of $\sum_{j \in R_i} e^{\beta Z_j}$. After simplication, our proposed estimator when σ^2 is known will be the root of the equation

$$S_3(\beta) = \sum_{i=1}^{n} \delta_i [X_i - \frac{\sum_{j \in R_i} X_j e^{\beta X_j}}{\sum_{j \in R_i} e^{\beta X_j}} + \beta\sigma^2 - \beta\sigma^2 \frac{e^{\beta X_i}}{\sum_{j \in R_i} e^{\beta X_j}}]. \tag{2.3}$$

2.1 Extended estimating function by over-parameterization

When σ^2 is unknown, there are two unknowns β and σ^2 in (2.2). We need another zero-unbiased estimating function to determine their estimates. We use the idea of over-parameterization (Huang, 2005) and extend the original model (1.2) to a $2nd$ order one, that is

$$L(\beta) = \prod_{i=1}^{n} [\frac{e^{\beta Z_i + \gamma Z_i^2}}{\sum_{j \in R_i} e^{\beta Z_j + \gamma Z_i^2}}]^{\delta_i}. \tag{2.4}$$

Then the partial scores for β and γ is

$$\sum_{i=1}^{n} \delta_i \{ \begin{pmatrix} Z_i \\ Z_i^2 \end{pmatrix} - \begin{pmatrix} \sum_{j \in R_i} Z_j e^{\beta Z_j + \gamma Z_j^2} / \sum_{j \in R_i} e^{\beta Z_j + \gamma Z_j^2} \\ \sum_{j \in R_i} Z_j^2 e^{\beta Z_j + \gamma Z_j^2} / \sum_{j \in R_i} e^{\beta Z_j + \gamma Z_j^2} \end{pmatrix} \}. \tag{2.5}$$

Since (2.4) are zero-unbiased for true parameter β and $\gamma = 0$, hence

$$\sum_{i=1}^{n} \delta_i \{ \begin{pmatrix} Z_i \\ Z_i^2 \end{pmatrix} - \begin{pmatrix} \sum_{j \in R_i} Z_j e^{\beta Z_j} \\ \sum_{j \in R_i} Z_j^2 e^{\beta Z_j} \end{pmatrix} / \sum_{j \in R_i} e^{\beta Z_j} \}.$$

are zero-unbiased. Again, multiply the weight $\sum_{j \in R_i} e^{\beta Z_j}$ to each summand, we have two zero-unbiased estimating functions

$$S_4(\beta) = \begin{pmatrix} \sum_{i=1}^{n} \delta_i [Z_i \sum_{j \in R_i} e^{\beta Z_j} - \sum_{j \in R_i} Z_j e^{\beta Z_j}] \\ \sum_{i=1}^{n} \delta_i [Z_i^2 \sum_{j \in R_i} e^{\beta Z_j} - \sum_{j \in R_i} Z_j^2 e^{\beta Z_j}] \end{pmatrix} \tag{2.6}$$

To find a corrected version of $S_4(\beta)$ function, the following lemma is helpful.

Lemma 1. Let $X_i = Z_i + e_i$, where $e_i's$ are i.i.d. normal r.v.'s with mean 0 and common variance σ^2, then we have

$$\begin{aligned} E(e^{\beta X_i - \frac{1}{2}\beta^2\sigma^2} \mid Z_i) &= e^{\beta Z_i} \\ E((X_i - \beta\sigma^2)e^{\beta X_i - \frac{1}{2}\beta^2\sigma^2} \mid Z_i) &= Z_i e^{\beta Z_i} \\ E((X_i - 2X_i\beta\sigma^2 + \beta^2\sigma^4 - \sigma^2)e^{\beta X_i - \frac{1}{2}\beta^2\sigma^2} \mid Z_i) &= Z_i^2 e^{\beta Z_i}. \end{aligned}$$

proof. By the moment generating function of e_i, we have $E(e^{\beta X_i} \mid Z_i) = e^{\beta Z_i - \frac{1}{2}\beta^2\sigma^2}$, hence the first equation follows. The $2nd$ and $3rd$ equations can be derived by differentiae the $1st$ equation with respect to β once and twice. □

Denote these unbiased predictors of $e^{\beta Z_i}$, $Z_i e^{\beta Z_i}$ and $Z_i^2 e^{\beta Z_i}$ by $A_0(X_i)$, $A_1(X_i)$ and $A_2(X_i)$. That is $A_0(X_i) = e^{\beta X_i - \frac{1}{2}\beta^2\sigma^2}$, $A_1(X_i) = (X_i - \beta\sigma^2)e^{\beta X_i - \frac{1}{2}\beta^2\sigma^2}$ and $A_2(X_i) = (X_i - 2X_i\beta\sigma^2 + \beta^2\sigma^4 - \sigma^2)e^{\beta X_i - \frac{1}{2}\beta^2\sigma^2}$. Replace the functions of Z_i in (2.6) by their predictors, we have a set of zero-unbiased estimating functions for β and σ^2,

$$\sum_{i=1}^{n} \delta_i [A_1(X_i) + X_i \sum_{j \in R_i, j \neq i} A_0(X_j) - \sum_{j \in R_i} A_1(X_j)],$$

$$\sum_{i=1}^{n} \delta_i [A_2(X_i) + (X_i - \sigma^2) \sum_{j \in R_i, j \neq i} A_0(X_j) - \sum_{j \in R_i} A_2(X_j)]$$

Note that these functions come from estimating the weighted score functions. To gain more efficiency, we reweight these estimating functions by dividing the estimated weights "$\sum_{j \in R_i} e^{\beta X_i - \frac{1}{2}\beta^2\sigma^2}$", it turns out that the proposed estimating functions when there is no extra information are

$$S_5(\beta, \sigma^2) = \sum_{i=1}^{n} \delta_i \begin{pmatrix} A_1(X_i) + X_i \sum_{j \in R_i, j \neq i} A_0(X_j) - \sum_{j \in R_i} A_1(X_j) \\ A_2(X_i) + (X_i - \sigma^2) \sum_{j \in R_i, j \neq i} A_0(X_j) - \sum_{j \in R_i} A_2(X_j) \end{pmatrix} / \sum_{j \in R_i} A_0(X_j). \tag{2.7}$$

Comparing the estimating functions (2.3) with (1.5) or (1.6), we found that they differ only by some terms that can be neglected asymptotically. Thus (1.5), (1.6) and (2.3) are

asymptotically equivalent and can yield consistent estimates when σ^2 is known. For the case when σ^2 is unknown, we first note that (2.5) can be expressed as $\sum \delta_i\{\binom{Z_i}{Z_i^2} - E[\binom{Z_i}{Z_i^2} \mid R_i, H_i]\}$ where H_i denotes the event that a failure occurs at time T_i. These two estimating function are both zero-unbiased and can be used to yield consistent estimates of β (though they are different in finite sample). By the same reason for the consistency of (2.3), we known that (2.7) are asymptotically zero-unbiased and can determine a consistent root of (β, σ^2) whenever $Cov(Z_i, Z_i^2)$ is of full rank.

3 Simulation studies

Simulation studies were carried out to investigate the finite sample properties of the previous estimators and verify the possibility of estimation without extra information. In addition to the Nakamura $1st$ order corrected score estimate and conditional score estimate, we also introduce the reweighted corrected score estimates (2.3) and (2.7) for the situation when σ^2 is known and unknown, respectively. Note that for the case when σ^2 is unknown, only estimator from (2.7) are available and can provide an estimate of σ^2. The notations we used are

$\hat{\beta}_{naive}$: the naive estimator which is the root of (1.4).

$\hat{\beta}_1$: The root of "(1.5)=0", the Nakaruma $1st$ order corrected score estimate.

$\hat{\beta}_2$: The root of "(1.6)=0", the conditional score estimate.

$\hat{\beta}_3$: The root of "(2.3)=0", the reweighted corrected score estimate when σ^2 is known.

$(\hat{\beta}_4, \hat{\sigma}^2)$: the root of "(2.7)=(0,0)"', The reweighted corrected score estimate when σ^2 is unknown.

We chose identity function as baseline hazard, and denote n the sample size. The n is chosen to be 300 and 600, censoring time C_i has distribution function $1 - e^{-\frac{c}{2.5}}$, and the true covariate Z_i is sampled from standardized $U(0, 1)$ so it has mean 0 and variance 1. The

results were exhibited in table 1.

As we expected, the naive estimator is not satisfactory due to its bias. The other estimators $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ work fine when σ^2 is known. $\hat{\beta}_2$ and $\hat{\beta}_3$ are preferable than $\hat{\beta}_1$ according to bias and variance criterion. The reweighted corrected score estimator $\hat{\beta}_3$ seems be an intermediate of $\hat{\beta}_1$ and $\hat{\beta}_2$ both in bias and variance. Howeverm, there are not much differences among them. This is consistent with the fact that the three estimates $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ are asymptotically equivalent. The estimator $\hat{\beta}_4$ is much variable than any other estimators, this may due to estimating σ^2 or multiple roots of (2.7). However, the accuracy of $\hat{\beta}_4$ and $\hat{\sigma}^2$ had improved much if the sample size n is 600 when compare with the case $n = 300$, this indicates that the estimates are consistent and can converge to the parameter if n goes to infinity.

There are some important problems about solving (2.7) should be aware. One is the multiple roots of (2.7). Typically, there are two solutions of $\hat{\sigma}^2$. When these two roots are both positive, we chose the small one as estimate of σ^2, so that estimate of β will close to the naive estimate which corresponds to the case $\sigma^2 = 0$. Besides, in some simulation case especially for the case $\beta = 0.7$, we occasionally fail to find a root of (2.7) that is reasonable and the numerical solution is too extreme to be an estimate. We will try to solve or mitigate these problems in the near future.

4 Discussion and future work

From this report, we know that the analysis in survival data with measurement error and without extra information is possible. The basic idea is to derived an additional unbiased equation through over-parameterization. However, there are many problems remain to solve before estimation without extra information to be practical. One of the problem is to find an optimal additional unbiased equation, as we see in section 2, the additional estimation

equation of (2.7) come from estimating $Z_i^2 - E(Z_i^2 \mid R_i, H_i)$. The choice of Z_i^2 is only a convenient one. The parameter σ^2 came into the estimating function (2.7) naturally since we used the $2nd$ moment of X_i to estimate Z_i^2. There are other moments of Z_i or functions of Z_i can be used and will also introduce the parameter σ^2 into the $2nd$ equation of (2.7). Hence, an obvious problem is to find a criterion to decide when the additional equation is reasonable or optimal. Another problem may encounter is the multiple roots of (2.7), the $2nd$ equation row in (2.7) is in fact a quadratic function of σ^2, thus we will expect there are two solutions for estimates of σ^2. How to chose one between them is also important. In this report, we chose the one which is positive and closer to 0, and chose 0 if none of them are positive. This procedure is reasonable, however, when one believe that the measurement error is not severe.

Besides the measurement error model introduced here, there are other possible applications of estimation without extra information like laten variable model or random effect model. For example, in a random effect regression model, let the predictor be $\alpha_{ij} + \beta Z_i$, where α_{ij} is the random effect and Z_i is the observable covariate, then the "covariate" βZ_i differs from the true predictor $\alpha_{ij} + \beta Z_i$ by a random term α_{ij}. If we treat α_{ij} as a random measurement error, βZ_i as the observed covariate and $\alpha_{ij} + \beta Z_i$ as the true covariate, then they look like a measurement error model. We think that the measurement error model approach is applicable and our technique is useful at least in some way.

In summary, there are many unsolved problems about the technique of estimations without extra information. And there are also interesting things worth to investigate like the application or extension of measurement error model to other useful statistical models. We will be pursued these problems in the $2nd$ year of the project.

References

Augustin, T. (2004) An Exact Corrected Log-Likelihood Function for Cox's Proportional Hazards Model under Measurement Error and Some Extensions. *Scand. J. Statist.* **31**, 43-50

Breslow, N. E. (1972). Contribution to the discussion of Cox (1972). *J. R. Stat. B Stat. Methodol.* **34**, 216-217.

Breslow, N. E. (1974). Covariate analysis of censored survival data. *Biometrics* **30**, 89-99.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Errors in Nonlinear Models,* Chapman & Hall, London.

Nakamura, T. (1992) Proportional Hazards Model with Covariates Subject to Measurement Error. *Biometrics* **48**, 829-838

Song, X. and Huang, Y. (2005) On Corrected Score Approach for Proportional Hazards Model with Covariate Measurement Error. *Biometrics* **61**, 702-714.

Tsiatis, A. and Davidian, M. (2001) A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, **88**, 447-458

黃逸輝 (2004). Cox 迴歸模式中解釋變數有測量誤差時的估計方法研究. 國科會報告 NSC 92-2118-M-032-005.

黃逸輝 (2005). 測量誤差模式在無額外訊息下的估計方法研究. 國科會報告 NSC 93-2118-M-032-006.

Table 1: Comparison of estimator's performances.

$n = 300, \quad Z \sim U(-1.732, 1.732), \quad e_i \sim i.i.d.N(0, \sigma^2)$

β	σ^2	β_{naive}	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}^2$
0.7	0.09	0.622 (0.0727)	0.695 (0.0851)	0.693 (0.0845)	0.693 (0.0847)	0.629 (0.450)	0.169 (0.131)
1.2	0.09	1.04 (0.0851)	1.21 (0.115)	1.20 (0.113)	1.20 (0.114)	1.18 (0.485)	0.105 (0.0850)
0.7	0.16	0.583 (0.0697)	0.705 (0.0931)	0.701 (0.0918)	0.702 (0.0923)	0.732 (0.259)	0.190 (0.161)
1.2	0.16	0.949 (0.0828)	1.23 (0.153)	1.21 (0.144)	1.22 (0.148)	1.16 (0.471)	0.136 (0.111)

$n = 600, \quad Z \sim U(-1.732, 1.732), \quad e_i \sim i.i.d.N(0, \sigma^2)$

β	σ^2	β_{naive}	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}^2$
0.7	0.09	0.627 (0.0498)	0.699 (0.0583)	0.698 (0.0581)	0.698 (0.0582)	0.692 (0.279)	0.151 (0.193)
1.2	0.09	1.04 (0.0659)	1.20 (0.0884)	1.20 (0.0873)	1.20 (0.0878)	1.22 (0.177)	0.0915 (0.0613)
0.7	0.16	0.590 (0.0485)	0.712 (0.0638)	0.710 (0.0633)	0.711 (0.0635)	0.702 (0.233)	0.174 (0.158)
1.2	0.16	0.948 (0.0532)	1.22 (0.0892)	1.21 (0.0870)	1.21 (0.0882)	1.21 (0.192)	0.153 (0.100)

*The numbers in parentheses are the sample standard deviation of the estimates.