行政院國家科學委員會專題研究計畫　成果報告

Cox 迴歸模式中解釋變數有測量誤差時的估計方法研究

計畫類別：個別型計畫
計畫編號：NSC92-2118-M-032-005-
執行期間：92 年 08 月 01 日至 93 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 93 年 11 月 22 日

A Simple Estimation in Cox Proportion Hazard Model when Covariates are Subject to Measurement Errors

Y.H. HUANG

Department of Mathematics, Tamkang University, Taipei, Taiwan
(Email: huang@math.tku.edu.tw)

Abstract: We consider the estimation in Cox proportion hazard model for censored survival data when the covariates are subject to measurement errors. To construct an unbiased estimating function when errors are present, we did not invoke the conditional distribution of the partial score function conditioned on the surrogates directly, but try to estimate a ”weighted” version of the partial score functions. This ”weighting” makes the estimation easy to justify, and more efficiency can be gained by applying some weights to the estimated weighted partial score function. This procedure needs no transformation, imputation or complicate integration to compute the estimating function. The resultant estimator is shown to be consistent. A small simulation study is provided to examine the performance of such estimation.

1 Introduction

In a regression analysis, it happens that some covariate are measured with errors. While there is a vast literature on measurement error problem in regression setting (see, e.g., Carroll et al. 1995), less work has been done in the context of failure time data. This research discusses a modification that can deal with the random measurement error in the Cox model. The Cox proportional hazards model has been the most popular regression model for analysis of censored data, however the analysis that cope with measurement error usually require some approximation or distribution assumption on the mismeasured covariate. For examples, the approach of induced hazard proposed by Prentice (1982) requires the dependence of the covariate error distribution on the regression parameter and baseline hazard is negligible. The corrected score approach suggested by Nakamura (1992) using a Taylor expansion that assume the measurement error is small. Cheng and Wang (2001) assume that there exist a transformation of the life time so that the mean function is linear in covariates, they also require that the normality assumption on the mismeasured covariates. Therefore, it is worthwhile to develop an analysis that needs no approximation or the distribution assumption.

Let $(T_i, C_i, \delta_i), i = 1, \cdots, n$, be the failure time, censoring time and noncensoring indicators for the ith study subject, and Z_i be the covariate that is related to the life time of the subject. For simplicity, we consider the case that Z_i is a scalar covariate only. The vector case can be handled through a straightforward generalization. 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 the time T_i, then a standard inference for the regression parameter is base 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, called the partial likelihood score function

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

However, it may happen that the covariate Z_i is mismeasured for some subjects. In such case, denote the surrogate of Z_i by X_i, and their difference by ν, that is $X_i = Z_i + \nu_i$. Furthermore, we assume that ν_i is i.i.d. $N(0, \sigma^2)$ distributed and is independent of $Z_i's$. In such scenario, the Cox model is no longer suitable since the intensity given X_i is

$$\tilde{\lambda}(t; X_i) = E(\lambda_0(t) e^{\beta Z_i} \mid T_i \geq t, X_i) = \lambda_0(t) E(e^{\beta Z_i} \mid T_i \geq t, X_i), \tag{1.4}$$

will not be a proportional hazard in general (Prentice, 1982). As a consequence, the naive approach which replace Z_i by X_i is not valid and will result inconsistent estimates.

In section 2, we will introduce a weighting in the partial score function, and then estimate the resultant function. The section 3 contains a small simulation study that compare the proposed estimator with other existing method. Finally, a conclusion is given in section 4.

2 Methodology

As demonstrated in Anderson and Gill (1982), define the process $C(\beta, t)$ as

$$C(\beta, t) = \sum_{i=1}^{n} \int_0^t \beta Z_i dN_i(s) - \sum_{i=1}^{n} \int_0^t \ln[\sum_{j=1}^{n} Y_j(S) e^{\beta Z_j}] dN_i(s), \tag{2.1}$$

where $Y_i(s) = I(T_i \leq s, C_i \leq s)$, and $N_i(s) = \delta_i I(T_i \leq s)$. Then we have $C(\beta, 1) = \ln L(\beta)$, and the usual estimate of β is the solution of $U(\beta, 1) = 0$, where

$$U(\beta, t) = \sum_{i=1}^{n} \int_0^t Z_i dN_i(s) - \sum_{i=1}^{n} \int_0^t \frac{\sum_{j=1}^{n} Y_j(s) Z_j e^{\beta Z_j}}{\sum_{j=1}^{n} Y_j(s) e^{\beta Z_j}} dN_i(s). \tag{2.2}$$

It is obvious that the counting process $N_i(s)$ has intensity function $Y_i(s)\lambda_0(s)\exp(\beta Z_i)$, it follows that $M_i(s) = N_i(s) - \int_0^s Y_i(u)\lambda_0(u) \exp^{\beta Z_i} du$ is a martingale. A straightforward computation shows that

$$U(\beta, t) = \sum_{i=1}^{n} \int_0^t Z_i dM_i(s) - \sum_{i=1}^{n} \int_0^t \frac{\sum_{j=1}^{n} Y_j(s) Z_j e^{\beta Z_j}}{\sum_{j=1}^{n} Y_j(s) e^{\beta Z_j}} dM_i(s) \qquad (2.3).$$

Note that the estimating function has mean zero 0 for $t \in [0,1]$, and the inferences base on (2.1) were well developed in Anderson and Gill (1982).

However, when the measurement errors present, we can't observe $Z_i's$ but its surrogate $X_i's$. Hence the estimating function (2.2) is not available. A common technique to handle the measurement error problem is the corrected score suggested by Nakamura (1990), which is essentially, seeking a conditional unbiased "estimator" of the score function, and then making inference base on this estimating function. The resultant estimates is consistent since the corrected score has mean zero as the score function does. Nevertheless, the partial score function (2.2) has terms involved mismeasured covariates in the denominator. It is difficult to find any unbiased "estimators" of them base on the surrogate. To make the denominator disappear, we combine the summand of (2.2) with some weights $\sum_{i=1}^{n} Y_j(s)e^{\beta Z_j}$, and come up with the weighted score function

$$U^*(\beta,t) = \sum_{i=1}^{n}\int_0^t Z_i(\sum_{i=1}^{n} Y_j(s)e^{\beta Z_j})dM_i(s) - \sum_{i=1}^{n}\int_0^t \sum_{j=1}^{n} Y_j(s)Z_j e^{\beta Z_j}dM_i(s), \tag{2.4}$$

which is also a martingale transformation and still has mean zero. Note that $U^*(\beta,1)$ equals

$$\sum_{i=1}^{n}\delta_i[Z_i\sum_{j\in R_i}e^{\beta Z_j} - \sum_{j\in R_i}Z_j e^{\beta Z_j}]. \tag{2.5}$$

Now, finding the unbiased estimator of terms in (2.5) becomes easy. For example, let $g_0(\beta,X_i) = \exp(\beta X_i - \frac{1}{2}\beta^2\sigma^2)$, then $g_0(\beta,X_i)$ has conditional mean $E(g_0(\beta,X_i) \mid Z_i) = e^{\beta Z_i}$. Consequently, when $i \neq j$, $X_j g_0(\beta,X_i)$ has conditional mean $Z_j e^{\beta Z_i}$ and $(X_i - \beta\sigma^2)g_0(\beta,X_i)$ has mean $Z_i e^{\beta Z_i}$. Substitute the terms in (2.5) by $g_0(\beta,X_i), X_j g_0(\beta,X_i)$, and $(X_i-\beta\sigma^2)g_0(\beta,X_i)$, respectively. We have an unbiased "estimator" of (2.5)

$$\sum_{i=1}^{n}\delta_i[X_i\sum_{j\in R_i}e^{\beta X_j-\frac{1}{2}\beta^2\sigma^2} - \sum_{j\in R_i}(X_j-\beta\sigma^2)e^{\beta X_j-\frac{1}{2}\beta^2\sigma^2} - \beta\sigma^2 e^{\beta X_i-\frac{1}{2}\beta^2\sigma^2}] \tag{2.6}$$

The solution of setting (2.6) to 0, denoted by $\tilde{\beta}_1$, is one of our proposed estimator. Since (2.6) has mean zero, under some mild conditions, the estimator is consistent and asymptotically normal distributed.

Consider the situation when measurement error is small, or more specifically when σ approaches zero, then obviously (2.6) will tends to (2.5) instead of the partial score (1.3). This transpires that the estimator $\tilde{\beta}_1$ derived from solving (2.6) to 0 is less efficient than the naive approach when the measurement error is small enough. To improve the efficiency, we try to reweights the summand in (2.6) and requires that the resultant function tends to the partial score (1.3) as measurement error tends to zero. Note that (1.3) can be derived form (2.5) by applying the weights "$1/\sum_{j\in R_i}\exp(\beta Z_j)$ ", and since "$1/\sum_{j\in R_i}\exp(\beta X_j - \frac{1}{2}\beta^2\sigma^2)$" is an estimate of the weight. Hence we divide the summand in (2.6) by $\sum_{j\in R_i} e^{\beta X_j - \frac{1}{2}\beta^2\sigma^2}$ and have the estimating function

$$\sum_{i=1}^{n}[X_i - \frac{\sum_{j\in R_i} X_j e^{\beta X_j}}{\sum_{j\in R_i} e^{\beta X_j}} + \beta\sigma^2 - \beta\sigma^2 \frac{e^{\beta X_i}}{\sum_{j\in R_i} e^{\beta X_j}}]. \tag{2.7}$$

Note that (2.7) becomes (1.3) when σ converges to 0, and hence the estimator derived from solving (2.7) being 0, denoted by $\tilde{\beta}_2$, is expected to be more efficient than the estimator derived from solving (2.6) to 0. However, it is not easy to justify that (2.7) has mean 0, the consistency of the estimator still remains a question and will be pursued in the future.

3 Simulation studies

Simulation studies were carried out to investigate the finite-sample properties of the proposed estimators. In addition to the previous estimators, we also introduce the estimator proposed by Nakamura (1992) which use the Taylor expansion to correct the bias in the partial score function. This corrected score estimator performs well when $\mid \beta\sigma \mid$ is less than 0.7, and also converge to the partial maximum likelihood estimator when measurement error tends to 0.

We computed the following estimators:

$\hat{\beta}_{pmle}$: the partial maximum likelihood estimator, which is the root of (1.3).

$\hat{\beta}_{naive}$: the naive estimator which is the also root of (1.3) but with X_i replacing Z_i.

$\tilde{\beta}_1$: the root of "(2.5)=0".

$\tilde{\beta}_2$: the root that "(2.7)=0".

$\hat{\beta}^*$: the corrected score estimator in Nakamura (1992).

Let the baseline hazard be the identity, n and m denote the number of samples and the number of failures, and (T_i, C_i, Z_i) denotes the failure times, censoring time and the covariates, respectively.

We consider two scenarios of simulation setting:

1). We fixed $n = 400$ and $m = 300$, and let $Z_i \sim (0, \sqrt{12})$. The "$m = 300$" means that only the 300 deaths before the 300th death were recorded, the remaining 100 life time were censored at the time of the 300th death. In a short, we observed $T_{(1)}, \cdots, T_{(300)}$, and $T_{(301)}, \cdots, T_{(400)}$ are only known to be larger than $T_{(300)}$. The results are shown in table 1.

2). n is chosen to be 300, and C_i has distribution function $1 - e^{-\frac{c}{2.5}}$, and Z_i is drawn from $N(0, 1)$. The ratio of m/n is about 0.7. The result is exhibited in table 2.

All the estimators work fine if measurement error is small. The naive estimator has severe bias problem when $\mid \beta\sigma \mid$ is moderate or large, and its value towards to 0 when measurement error getting large. Except the naive estimator, all estimators has slight bias. and the mean square error is almost determined by the variance. In some cases, $\hat{\beta}^*$ is better than $\tilde{\beta}_1$, but it has larger variance than $\tilde{\beta}_2$ in most cases. Obviously, $\tilde{\beta}_2$ seems to be preferable in both tables.

Table 1: Comparison of estimator's performances.

$n = 400, m = 300, Z \sim U(0, 3.464)$

β	σ	β_{pmle}	$\beta naive$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\hat{\beta}^*$
0.5	0.3	0.501 (0.062)	0.455 (0.062)	0.502 (0.077)	0.502 (0.070)	0.504 (0.073)
0.75	0.3	0.741 (0.065)	0.666 (0.064)	0.741 (0.089)	0.743 (0.076)	0.741 (0.082)
1	0.3	1.01 (0.080)	0.894 (0.072)	1.02 (0.107)	1.02 (0.090)	1.02 (0.093)
0.5	0.6	0.498 (0.060)	0.357 (0.050)	0.509 (0.081)	0.503 (0.078)	0.501 (0.088)
0.75	0.6	0.757 (0.065)	0.526 (0.056)	0.767 (0.112)	0.769 (0.106)	0.767 (0.113)
1	0.6	1.00 (0.074)	0.664 (0.066)	0.925 (0.516)	1.03 (0.164)	1.04 (0.171)

Table 2: Comparison of estimator's performances.

$n = 300, F_C(c) = 1 - e^{-\frac{c}{2.5}}, Z \sim N(0, 1)$

β	σ	β_{pmle}	$\beta naive$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\hat{\beta}^*$
0.5	0.3	0.500 (0.077)	0.455 (0.072)	0.501 (0.095)	0.503 (0.082)	0.499 (0.086)
0.75	0.3	0.758 (0.086)	0.682 (0.077)	0.761 (0.104)	0.765 (0.091)	0.764 (0.098)
1	0.3	1.01 (0.095)	0.888 (0.096)	1.03 (0.126)	1.02 (0.124)	1.02 (0.126)
0.5	0.6	0.519 (0.082)	0.365 (0.070)	0.524 (0.120)	0.519 (0.082)	0.516 (0.119)
0.75	0.6	0.757 (0.083)	0.519 (0.067)	0.778 (0.139)	0.789 (0.141)	0.788 (0.149)
1	0.6	1.01 (0.086)	0.655 (0.076)	1.05 (0.202)	1.06 (0.220)	1.06 (0.264)

4 Conclusion

A weighted score function is derived by apply some weights in combining the summand of the partial score function, the weighted function is still of mean zero and is very easy to estimate. The estimating function, which is the estimated function of the weighted partial score function is shown to be mean zero, and hence it yields consistent estimator of the parameter. Furthermore, when another weight is apply on the previous estimating function, the estimating function will converge to the original partial score function when measurement error tends to 0. The simulation results show that these estimating functions work well, and can do better than the corrected score estimator of Nakamura (1992) in some context.

From the derivation of the estimating function, the moment generating function of the measurement error is the key knowledge and is not restricted to be normal distribution. Therefore, this method is easy to modify for other nonormal errors as long as the moment generating function is known.

References

Anderson, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10**, 1100-1120

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

Cheng, S. and Wang, N (2001) Linear Transformation Models for Failure Time Data With Covariate Measurement Error. *Journal of the American Statistical Association* **96**, 706-716.

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

Prentice, R. L. (1982) Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, **69**, 331-342