

A Monte Carlo Study to Evaluate the Robust Standard Error Feature in the PHREG procedure

Ji Li and Xu Zhang

Department of Mathematics and Statistics, Georgia State University

ABSTRACT

Cox proportional hazards model is the most commonly used regression model to analyze survival data. The regression parameters in a Cox model can be estimated by maximizing the partial likelihood (Cox 1972, 1975). When the association is presented in the data due to matching design, recurrent events or multiple types of failure, the marginal approach can be used to adjust for association. In this approach, a robust estimator of variance-covariance for the estimated regression parameters is developed. The robust estimator has been implemented as a new feature in the PHREG procedure in SAS[®] version 9. In this study, we employ the new feature of the PHREG procedure to conduct a simulation study. The aim of the study is to investigate the performance of the robust estimator with various degrees of association.

INTRODUCTION

The Cox proportional hazards model (Cox, 1972) is commonly used to explore the effects of different covariates on survival. In a Cox model, the hazard function is specified as the product of an unconstrained baseline hazard and an exponential function of the covariates. The partial likelihood has been proposed for estimation of the regression coefficients (Cox, 1972, 1975). To evaluate the precision of the maximum likelihood estimates (MLE) of regression coefficients, the inverse of the estimated information matrix can be used as the naïve estimator of variance-covariance. The Cox model has been implemented in many statistical computing packages such as SAS[®], Splus, R, SPSS, and Strata. In SAS/STAT[®], the Cox model is implemented in the PHREG procedure. The output of this procedure includes the MLE of the regression coefficients and the standard errors using the naïve estimator.

The routine inference procedure of the Cox model is applicable to a sample with independent observations. For clustered survival data, appropriate extensions to the Cox model or the inference procedures are needed to adjust for the association among the failure time data within clusters. The frailty model is one solution. For this approach, the hazard function is multiplied by a cluster specific frailty and the frailties are treated as realizations from a probability distribution. Gamma, positive stable, and log-normal are the commonly chosen probability distributions for frailty (Hougaard, 2000). The Expectation Maximization (EM) algorithm is often employed in estimation of the parameters in a frailty model. The statistical software for the frailty model includes the Splus/R function “coxph” with “frailty” option, which implements the gamma and the log-normal frailty models, and a SAS[®] macro for positive stable frailty model (Shu and Klein, 2007).

The other solution related with the Cox model to deal with clustered survival data is the robust inference (Lin and Wei, 1989; Wei, Lin and Weissfeld, 1989; Lee, Wei, and Amato, 1992; Lin, 1994). It can be shown that the score equation of the partial likelihood can be treated as the sum of independent and identically distributed random vectors across the clusters, and a moment estimator can be used for estimating the variance-covariance of the estimated score equation. Regarding the MLE of the regression coefficients, a Sandwich type estimator for variance-covariance, which is also known as the robust estimator or the WLW method, was proposed. In SAS[®] version 9, the robust inference has been added to the PHREG procedure by using the option COVS (AGGREGATE). The enhanced PHREG procedure can be used in a wider context including the clustered data. The robust inference has been demonstrated to be effective for some clustered data (Lin, 1994). Its performance is not clear for data with strong association. In this study, we employ the new feature of the PHREG procedure to perform simulation, focusing on coverage of the confidence intervals using the robust estimator on the clustered survival data with various degrees of association.

THE ROBUST INFERENCE FOR COX MODEL

In the Cox model, the hazard function $\lambda_i(t)$ has the form:

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta^T Z_i\}, \quad i = 1, \dots, n \quad (1)$$

where $\lambda_0(t)$ is the unspecified baseline hazard function, β is the vector of regression coefficients and Z_i is the vector of covariates for the i th subject. To estimate the regression coefficient β , a partial likelihood function is constructed (Cox 1972, 1975),

$$L(\beta) = \prod_{i=1}^n \left[\frac{\exp\{\beta^T Z_i\}}{\sum_{j \in R_i} \exp\{\beta^T Z_j\}} \right]^{\delta_i} \quad (2)$$

In the above likelihood, R_i is the risk set at time t_i . Specifically, it is the set of all subjects who are still under study at the time just prior to t_i . $\delta_i = 1$ if a failure time is observed, and $\delta_i = 0$ if a censoring time is observed. The MLE $\hat{\beta}$ is the values that maximize (2).

$$\text{Let} \quad U(\beta) = \partial \log L(\beta) / \partial \beta \quad (3)$$

$$\hat{I}(\beta) = -\partial^2 \log L(\beta) / \partial \beta^2 \quad (4)$$

$$\hat{B}(\beta) = \sum \hat{U}_i(\beta) \hat{U}_i'(\beta) \quad (5)$$

where $U_i(\theta)$ is the contribution from the i th observation to the score function (3) and $\hat{U}_i(\beta)$ is its estimate. $\hat{I}(\beta)$ is the information matrix and $\hat{I}^{-1}(\hat{\beta})$ is the naïve estimator of variance-covariance. The “sandwich” estimator or the robust estimator is defined as

$$\hat{I}^{-1}(\hat{\beta}) \hat{B}(\hat{\beta}) \hat{I}^{-1}(\hat{\beta}). \quad (6)$$

The PHREG procedure in SAS/STAT[®] routinely reports the standard errors based on the naïve estimator. With the newly added option COVS, the robust standard errors based on (6) would be included in the output as well. This method was originally proposed for the univariate survival data to protect against model misspecification (Lin and Wei, 1989).

With clustered data, the model specification should reflect the data structure,

$$\lambda_{ij}(t) = \lambda_0(t) \exp\{\beta^T Z_{ij}\}, \quad i = 1, \dots, n \quad j = 1, \dots, n_i \quad (7)$$

where Z_{ij} is the covariate for the j th subject in the i th cluster. Still, let $U(\beta)$ be the score equation. Let $\hat{I}^{-1}(\hat{\beta})$ be the naïve estimator of variance-covariance. Now

$$U(\beta) = \sum_{i=1}^n \sum_{j=1}^{m_i} U_{ij}(\beta) \quad (8)$$

Lin (1994) gave the estimator:

$$\hat{B}^*(\hat{\beta}) = \sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} \hat{U}_{ij}(\hat{\beta}) \hat{U}_{ik}(\hat{\beta})', \quad (9)$$

where $\hat{U}_{ij}(\beta)$ is estimated $U_{ij}(\beta)$. Then the robust estimator is given by

$$\hat{I}^{-1}(\hat{\beta}) \hat{B}^*(\hat{\beta}) \hat{I}^{-1}(\hat{\beta}). \quad (10)$$

This estimator is also implemented in the PHREG procedure. It can be invoked by using the COVS (AGGREGATE) option and specifying the proper variable in the ID statement that distinguishes clusters.

THE SIMULATION STUDY

We have conducted a simulation study to investigate the performance of the robust variance estimator with different degrees of association among clusters. The following regression model was used for the clustered failure time data.

$$\lambda_{ij}(t) = w_i \lambda_0(t) \exp\{\beta^T Z_{ij}\}, \quad i=1, \dots, n \quad j=1, 2. \quad (11)$$

where n is the number of clusters and there are two observations within each cluster. Here we considered the simplest type of the clustered data, the matched pairs data. The commonly sharing profile w_i causes association among the failure times within the i th cluster.

Two covariates have been considered in the study, $Z = (Z_1, Z_2)$ with the effect $\beta = (\beta_1, \beta_2)$. The continuous covariates were generated from standard normal distribution and we set $\beta_1=0.5$, $\beta_2=-0.8$; we used Bernoulli distribution with parameter value 0.5 to generate discrete covariates and we set $\beta_1=1$, $\beta_2=1.5$. We let $\lambda_0(t)=1$ for all t . The cluster effect w_i was generated from a Gamma distribution with parameter $(\alpha, 1/\alpha)$. Different α values determine the degree of association. In this simulation study, four different values for α were used, $\alpha=1, 0.5, 0.1, 0.01$. A large α value leads to a strong association and a small α value yields a weak association.

We assume $\lambda_0(t)=1 \forall t$, then the cumulative hazard function is just

$$\Lambda_0(t) = \int_0^t \lambda_0(u) du = t. \quad (12)$$

The survival function for the j th subject in the i th cluster has the form:

$$S_{ij}(t) = \exp\left\{-\int_0^t w_i \lambda_0(u) \exp(\beta^T Z_{ij}) du\right\} \quad (13)$$

$S_{ij}(t)$ can be viewed as a random variable u from the uniform distribution in $[0, 1]$. For simplicity, we just treat w_i as fixed. Then

$$u = \exp\{-w_i \exp(\beta^T Z_{ij}) t\} \quad (14)$$

The failure time for the j th observation in the i th cluster can be obtained by solving t in Equation (14). We considered the low and high censoring rates (25% and 50%). The censoring times were generated from the uniform distribution in the interval $[0, a]$. The values of a were searched so that we obtained the desired censoring rates for these settings.

We let the sample size to be 100, and each sample hence consists of 50 clusters. We generated 1000 replicates for each setting. The values of regression parameter estimate $(\hat{\beta}_1, \hat{\beta}_2)$ of these 1000 repetitions were saved, and the averages of these 1000 $(\hat{\beta}_1, \hat{\beta}_2)$ were obtained. In addition, the 95% confidence intervals have been computed by using the robust standard errors and we report the coverage. Tables 1-4 are the summary of parameter estimate values and the 95% linear confidence interval coverage for 25% and 50% censoring rate.

	Continuous covariates ($\beta_1=0.5$, $\beta_2 =-0.8$) for 25% censoring rate			
	$\hat{\beta}_1$ ($\beta_1 =0.5$)	95% confidence interval coverage for β_1	$\hat{\beta}_2$ ($\beta_2 =-0.8$)	95% confidence interval coverage for β_2
$\alpha =0.01$	0.511	93.7	-0.816	93.6
$\alpha =0.1$	0.488	94.0	-0.764	90.9
$\alpha =0.5$	0.403	83.8	-0.645	70.4
$\alpha =1$	0.393	70.9	-0.671	50.7

Table 1. The 95% confidence interval coverage using the robust variance estimator with continuous covariates and 25% censoring rate.

	Continuous covariates ($\beta_1=0.5$, $\beta_2 =-0.8$) for 50% censoring rate			
	$\hat{\beta}_1$ ($\beta_1 =0.5$)	95% confidence interval coverage for β_1	$\hat{\beta}_2$ ($\beta_2 =-0.8$)	95% confidence interval coverage for β_2
$\alpha =0.01$	0.513	93.7	-0.810	92.3
$\alpha =0.1$	0.496	92.6	-0.788	90.9
$\alpha =0.5$	0.436	87.8	-0.700	83.5
$\alpha =1$	0.388	83.5	-0.631	73.4

Table 2. The 95% confidence interval coverage using the robust variance estimator with continuous covariates and 50% censoring rate.

	Discrete covariates ($\beta_1=1$, $\beta_2 =1.5$) for 25% censoring rate			
	$\hat{\beta}_1$ ($\beta_1 =1$)	95% confidence interval coverage for β_1	$\hat{\beta}_2$ ($\beta_2 =1.5$)	95% confidence interval coverage for β_2

$\alpha = 0.01$	1.008	93.0	1.516	93.6
$\alpha = 0.1$	0.934	92.5	1.491	91.1
$\alpha = 0.5$	0.747	75.8	1.151	59.8
$\alpha = 1$	0.657	54.9	0.965	28.4

Table 3. The 95% confidence interval coverage using the robust variance estimator with discrete covariates and 25% censoring rate.

	Discrete covariates ($\beta_1 = 1$, $\beta_2 = 1.5$) for 50% censoring rate			
	$\hat{\beta}_1$ ($\beta_1 = 1$)	95% confidence interval coverage for β_1	$\hat{\beta}_2$ ($\beta_2 = 1.5$)	95% confidence interval coverage for β_2
$\alpha = 0.01$	1.064	93.2	1.568	93.6
$\alpha = 0.1$	0.965	92.4	1.552	91.8
$\alpha = 0.5$	0.849	84.9	1.266	72.4
$\alpha = 1$	0.707	69.6	1.010	49.2

Table 4. The 95% confidence interval coverage using the robust variance estimator with discrete covariates and 50% censoring rate.

In each table, the first column gives different α values of the Gamma distributions, which were used to generate w_i in (14). A larger α value indicates a stronger association within clusters. All the tables show that when the α is small, the 95% confidence interval coverage is close to the nominal level. As the α value increases, the parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ depart from the true values, and the 95% confidence interval coverage is quite low. For example, in Table1, when α is =0.01, the 95% confidence interval coverage for β_1 and β_2 is 93.7 and 93.6, respectively. When α is =1, the 95% confidence interval coverage for β_1 and β_2 is only 70.9 and 50.7, respectively. We can draw the same conclusions for other tables. Tables1 and 3 are the results under 25% censoring rate. Tables 2 and 4 are the result under 50% censoring rate. For both censoring rates, the coverage is generally good when the association is weak, but becomes worse when association becomes stronger.

CONCLUSION

The robust inference for clustered survival data has been well recognized in real applications. Compared to a frailty model, the robust inference is appealing in terms of constant hazards ratio and flexible dependence structure. The PHREG procedure, with the new feature of the robust estimator, will be applicable to a wider range of survival data.

The robust inference requires the condition that the marginal distribution of the failure time satisfies the Cox model specification. In our simulation study, we have showed that the robust inference fail to give sufficient coverage for data with strong within-cluster association. It is clear that strong association leads to the departure of the marginal distribution of the failure time from the proportional hazards pattern. Based on our study result, we recommend adoption of the robust inference to clustered data with weak association. Some preliminary investigation or diagnostics procedure is needed to measure the level of association.

REFERENCES

- Cox, D. R. (1972). Regression Models and Life Tables (with discussion). *Journal of the Royal Statistical Society*, Ser. B,34, 187-220.
- Cox, D. R. (1975). Partial Likelihood. *Biometrika* **62**, 269-276.
- Hougaard, P. (2002). Analysis of Multivariate Survival Data. *Springer-Verlag*. New York, Inc.
- Lee, E., Wei, L.J., and Amato, D. (1992). Cox-Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations, 237-247, Netherlands: Kluwer Academic Publishers.
- Lin, D.Y. (1994). Cox Regression Analysis of Multivariate Failure Time Data: The Marginal Approach. *Statistics in Medicine* **13**, 233-2247.
- Lin, D. Y. and Wei, L. J. (1989). The Robust Inference for the Cox Proportional Hazards Model. *J. American Stat. Assoc.* **84**, 1074-1077.
- SAS Institute Inc. 2008. SAS/STAT® 9.2 User's Guide, Gary, NC;SAS Institute Inc.
- Shu, Y. and Klein, J. P. (1999). A SAS Macro for the Positive Stable Frailty Model. *Amerian Statistical Association Proceedings of the Statistical Computing Section*, 47-52.
- Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *J. American Stat. Assoc.* **84**, 1065-1073.

APPENDIX

SAS code for the robust standard error estimation in PHREG procedure:

```
proc phreg data=indata covs(aggregate) covm outest=out;
model time*event(0)=z1 z2;
id id;
run;
```

CONTACT INFROMATION

Name: Ji Li

Enterprise: Georgia State University

Address: 34 Peachtree St

City, State, ZIP: Atlanta, GA, 30303

E-mail: jjao_198@hotmail.com

Name: Xu Zhang

Enterprise: Georgia State University

Address: 34 Peachtree St

City, State, ZIP: Atlanta, GA, 30303

E-mail: matxxz@langate.gsu.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies.