AUXILIARY COVARIATE MEASUREMENT PROBLEM IN FAILURE TIME REGRESSION

by

Haibo Zhou and Margaret Pepe*

Department of Biostatistics, University of North Carolina at Chapel Hill, NC.

Institute of Statistics Mimeo Series No. 2112

April 1993

Auxiliary Covariate Measurement Problem in Failure Time Regression

Haibo Zhou and Margaret Pepe*

SUMMARY

A semi-parametric method, Estimated Partial Likelihood(EPL) method, is proposed for the estimation of the relative risk parameters when auxiliary covariates are present. The asymptotic distribution theory is derived for the proposed estimator for the case in which the surrogate or mismeasured covariates are categorical. The asymptotic relative efficiency of the EPL estimator with respect to the fully parametric maximum likelihood analysis and a partial likelihood analysis based on those with true covariates only are examined under the exponential model. The EPL analysis is found to compare favorably with more model-dependent analysis and more efficient than the partial likelihood analysis in most cases of practical importance. Small sample properties are investigated by simulation studies and a real data example is used to illustrate the EPL method.

Some key words: Auxiliary Covariate; Counting process; Estimated Partial Likelihood; Measurement Error; Martingale; Relative Risk Parameter Estimation; Semiparametric estimation; Stochastic Intergration; Surogate Covariate Data.

* Haibo Zhou is a Postdoctoral Fellow, Department of Biostatistics, The University of North Carolina at Chapel Hill. Chapel Hill, N.C. 27599-7400. Margaret S. Pepe is an associate member, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, Washington 98104. This research was supported in part by National Science Foundation grant No. DMS 89-02211. Part of this work was included in 1992 Department of Statistics, University of Washington Doctoral dissertation by Haibo Zhou. The authors thank Ed Davis for the helpful discussion about SOLVD data and comments.

1. INTRODUCTION

Auxiliary Covariate data is a common problem in statistical data analysis. For example, in a substudy of the Studies of Left Ventricular Dysfunction(SOLVD)(SOLVD Investigators, 1991) , the patient's left ventricular ejection fraction(EF) is an importent risk factor in predicting the patient's heart failure. However due to the complication and expensiveness of the measuring procedure, the EF, which is a standardized radionuclide measurement, is only avaiable to a subset of 179 patients in the study. For a total of 1111 patients in the substudy, a nonstandardized across clinical centers measurement of EF is avaiable to everyone.

Prentice(1982) first considered the relative risk parameter estimation in a failure time regression model by introducing the induced relative risk function for the failure rate given the mismeasured covariate. This paper provides a semi-parametric method, the Estimated Partial Likelihood(EPL) method, to the estimation problem when a subsample, the validation sample, is avaiable.

Suppose there are total n subjects in the study. The validation sample, denoted by V, is the subsample in which both true covariate X and auxiliary covariate Z are observed. The non-validation sample, denoted by $\bar{V}$, is the subsample in which only the auxiliary covariate is observed. Suppose an individual i in the validation sample has a hazard process model of the form (Cox, 1972)

$$\lambda_i(t) = Y_i(t)\lambda_0(t)e^{\beta' X_i(t)} \tag{1}$$

where $Y_i(\cdot)$ is an 'at risk' indicator process, $X_i'(\cdot) = \{X_{i1}(\cdot), \cdots, X_{ip}(\cdot)\}$ is a modeled regression vector, $\beta' = (\beta_1, \cdots, \beta_p)$ is a relative risk parameter to be estimated, and $\lambda_0(\cdot) \geq 0$ is an unspecified baseline hazard function. Prentice(1982) had shown that given only the auxiliary covariate Z avaiable, the hazard process model for an

individual j in the non-validation set has the form

$$\lambda_j(t) = Y_j(t)\lambda_0(t)E[e^{\beta' X_i(t)}|Y_j(t) = 1, Z(t)] \tag{2}$$

Let $r_i(t)$ and $\bar{r}_i(t)$ denote the relative risk function for the ith individual in the validation and non-validation sample respectively, i.e.

$$\begin{aligned} r_i(t) &= e^{\beta' X_i(t)} && if \quad i \in V \\ \bar{r}_i(t) &= E(e^{\beta' X_i(t)}|Y_i(t) = 1, Z_i(t)) && if \quad i \in \bar{V} \end{aligned}$$

For the observed data we will denote the relative risk function for an individual i as

$$r_i^*(t) = r_i(t)I_{[i\in V]} + \bar{r}_i(t)I_{[i\in \bar{V}]} \tag{3}$$

Possible methods that can be employed to draw inference about β in our situation are: (i) to specify the underlying distribution and use a fully parametric maximum likehood analysis; (ii) to discard the non-validation sample observations and use a partial likelihood analysis based on the validation sample only; (iii) to use the induced partial likelihood which is based on the whole sample. The induced partial likelihood function is of the form

$$IPL(\beta) = \prod_{i=1}^{n} [\frac{r_i^*(T_i)}{\sum_{j\in\mathcal{R}(T_i)} r_j^*(T_i)}]^{\delta_i} \tag{4}$$

The first approach is the most comprehensive parametric approach and yet the most efficient one under correctly specified model. Heowever realistic model is difficult to construct, furthmore, this method may not be robust to model misspecification. The second approach simplly discard the information in the non-validation sample and can lead to substantial reduction in efficiency. This leaves the approach (iii) a great interest. However this approach can not be carried out in general because the condition $\{Y_i(t) = 1, Z_i(t)\}$ in (3) usually imply that (3) will depend on baseline hazard function $\lambda_0(t)$ and the underline joint distribution of covariates. This paper aimed at

providing a method that assumes no stringent parametric assumption yet still utilizes the information about β in the non-validation sample.

In Section 2, we will present the proposed EPL method which is non-parametric with respect to $\lambda_0(t)$ and underline distribution and draw inference about β from a (4) type likelihood function. The asymptotic distribution theory of the EPL estimator will be given in Section 3, followed by asymptotic relative efficiency evaluation, simulation study and data example in Section 4 and 5 respectively.

2. METHOD DESCRIPTION

Assuming auxiliary covariate Z is discrete and has no prediction value given the true covariate X, we propose to estimate $\bar{r}(t)$ empirically on the basis of the validation sample observations. So the estimated relative risk function for an individual i given his observed data is:

$$\hat{r}_i^*(t) = r_i(t) I_{[i \in V]} + \hat{\bar{r}}_i(t) I_{[i \in \bar{V}]} \tag{5}$$

where $\hat{\bar{r}}_i(t)$ is the estimated relative risk function for an individual in the non-validation sample with only covariate process $Z(t)$ available. Specifically, the empirical estimate of the induced relative risk function denoted by $\hat{\bar{r}}(t)$ is

$$\begin{aligned} \hat{\bar{r}}(t) &= \hat{E}[r(\beta, X(t)) | Y(t) = 1, Z(t)] \\ &= \frac{\sum_{i \in V} I_{[Y_i(t)=1, Z_i(t)=Z(t)]} r(\beta, X_i(t))}{\sum_{i \in V} I_{[Y_i(t)=1, Z_i(t)=Z(t)]}}, \end{aligned} \tag{6}$$

The resulting estimated partial likelihood function is,

$$EPL(\beta) = \prod_{i=1}^{n} [\frac{\hat{r}_i^*(T_i)}{\sum_{j \in \mathcal{R}(T_i)} \hat{r}_j^*(T_i)}]^{\delta_i} \tag{7}$$

We call $\hat{\beta}_{EPL}$ the Maximum Estimated Partial Likelihood Estimate if $\hat{\beta}_{EPL}$ is the maximizer of (7).

The esitmated partial likelihood approach is robust to model misspecification in that no underline parametric structure is assumed. The basedline hazard function is

still left to be nuisiance. The information cantained in the non-validation sample are fully included in drawing inference about regression coefficients. The computation of the estimated partial likelihood approach is straight forward.

3. ASYMPTOTIC DISTRIBUTION THEORY

The background theory for this section and Appendix are the theories for multivariate counting processes, stochastic integrals and local martingales. We shall use the basic results from those theories without further comment. A good survey of these theories can be found in Fleming & Harrington(1991). In our model, properties of stochastic processes, such as being a local martingale or a predictable process, are relative to a right-continuous nondecreasing family $\{\mathcal{F}_t^{(n)} : t \in [0,1]\}$ of sub σ-algebras on the nth probability space $(\Omega^{(n)}, \mathcal{F}^{(n)}, \mathcal{P}^{(n)})$; $\mathcal{F}_t^{(n)}$ is the filtration and can be thought of as the history of everything that happens up to time t(in the nth model). Apart from the background theories, our basic tools are the Inverse Function Theorem(Rudin, 1964), the extented Inequality of Lenglart and the Martingale Central Limit Theorem of Rebolledo(Anderson & Gill 1982). The covariate processes $X(t)$ and $Z(t)$ are assumed to be predictable and locally bounded. The standard independent failure time and independent censorship are assumed.

Recall that T_i, $Y_i(t)$ and $\delta_i(t)$ denote the failure time, at risk indicator and failure indicator respectively, for $i = 1, ..., n$, the counting process $N_i(t)$ and martingale $M_i(t)$ are defined as

$$\begin{aligned} N_i(t) &= I(T_i \leq t, \ \delta_i = 1) \\ M_i(t) &= N_i(t) - \int_0^t Y_i(u)\lambda_i(u)\, du \end{aligned}$$

For a matrix A or vector a, we define the norm as $\|A\| = sup_{i,j}|a_{ij}|$ and $\|a\| = sup_i|a_i|$. For a vector a, define $|a| = (\sum a_i^2)^{\frac{1}{2}} = (a'a)^{\frac{1}{2}}$. We also write the matrix of

aa' as $a^{\otimes 2}$ and the matrix $(aa')(aa')'$ as $a^{\otimes 4}$. For the relative risk function r^* (as well as for $\hat{r}^*$, r, $\hat{\bar{r}}$ and $\bar{r}$), we let $r^{*(j)}$ denote the jth derivative of r^* with respect to β, $j = 0, 1, 2$, where $j = 0$ represents the function itself.

Some further definitions are:

$$\begin{aligned}
\hat{S}^{(0)}(\beta, t) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\hat{r}_i^*(\beta, t) \\
\hat{S}^{(1)}(\beta, t) &= \frac{\partial}{\partial\beta}\hat{S}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\hat{r}_i^{*(1)}(\beta, t) \\
\hat{S}^{(2)}(\beta, t) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\left(\frac{\hat{r}_i^{*(1)}(\beta, t)}{\hat{r}_i^*(\beta, t)}\right)^{\otimes 2} r_i^*(\beta_0, t)
\end{aligned}$$

We define $S^{(j)}(\beta, t)$ as the corresponding functions with $r^*(\beta, t)$ substituted for $\hat{r}^*(\beta, t)$ in the above $\hat{S}^{(j)}(\beta, t)$, $j = 0, 1, 2$. Also, we define

$$\begin{aligned}
s^{(0)}(\beta, t) &= E(Y(t)r^*(\beta, t)) \\
s^{(1)}(\beta, t) &= E(Y(t)r^{*(1)}(\beta, t)) \\
s^{(2)}(\beta, t) &= E\left(Y(t)\left(\frac{r^{*(1)}(\beta, t)}{r^*(\beta, t)}\right)^{\otimes 2} r^*(\beta_0, t)\right)
\end{aligned}$$

and

$$\begin{aligned}
\Sigma &= \int_0^1 \left[\frac{s^{(2)}(\beta_0, \omega)}{s^{(0)}(\beta_0, \omega)} - \left(\frac{s^{(1)}(\beta_0, \omega)}{s^{(0)}(\beta_0, \omega)}\right)^{\otimes 2}\right] s^{(0)}(\beta_0, \omega)\lambda_0(\omega)d\omega \\
\Sigma_1 &= \int_0^1 \left[E\left(Y_i(\omega)\frac{\bar{r}_i^{(1)}(\beta_0, \omega)^{\otimes 2}}{\bar{r}_i(\beta_0, \omega)}\right) - \frac{s^{(1)}(\beta_0, \omega)^{\otimes 2}}{s^{(0)}(\beta_0, \omega)}\right] \lambda_0(\omega)d\omega \\
\Sigma_2 &= E\left\{\int_0^1 \left(\frac{r_j^{(1)}(\beta_0, \omega)}{r_j(\beta_0, \omega)} - \frac{s^{(1)}(\beta_0, \omega)}{s^{(0)}(\beta_0, \omega)}\right) dM_j(\omega) - \right. \\
&\quad \left. \frac{1-\rho}{\rho}\int_0^1 \left(\frac{\bar{r}_j^{(1)}(\beta_0, \omega)}{\bar{r}_j(\beta_0, \omega)} - \frac{s^{(1)}(\beta_0, \omega)}{s^{(0)}(\beta_0, \omega)}\right) Y_j(\omega)\lambda_0(\omega)d\omega\right\}^{\otimes 2} \qquad (8)
\end{aligned}$$

where ρ is the limit of the validation fraction in the sample.

With above defination, the score function $\hat{U}(\beta, t)$ from estimated partial likelihood(7) can be expressed as

$$\hat{U}(\beta, t) = \sum_{i=1}^{n}\int_0^t \Delta(\hat{r}_i^*)(s)dM_i(s) + \sum_{i=1}^{n}\int_0^t \Delta(\hat{r}_i^*)(s)r_i^*(s)Y_i(s)\lambda_0(s)ds \qquad (9)$$

where $\Delta(\hat{r}_i)(s)$ is defined as

$$\Delta(\hat{r}_i)(s) = \frac{\hat{r}_i^{(1)}(s)}{\hat{r}_i(s)} - \frac{\sum_{i=1}^n Y_i(s)\hat{r}_i^{(1)}(s)}{\sum_{i=1}^n Y_i(s)\hat{r}_i(s)}$$

Since we define $\hat{\beta}_{EPL}$ as the solution of $\hat{U}(\beta, 1) = 0$, a Taylor expansion of $\hat{U}(\beta, 1)$ about β_0 evaluated at $\hat{\beta}_{EPL}$ gives

$$n^{-\frac{1}{2}}\hat{U}(\beta_0, 1) = \{-n^{-1}\frac{\partial}{\partial\beta_*}\hat{U}(\beta_*, 1)\}n^{\frac{1}{2}}(\hat{\beta}_{EPL} - \beta_0)$$

where β_* is between $\hat{\beta}$ and β_0. Therefore to prove asymptotic normality of $n^{\frac{1}{2}}(\hat{\beta}_{EPL} - \beta_0)$ it is suffficient to show $n^{-\frac{1}{2}}\hat{U}(\beta_0, t)$ converges weakly to a Gaussian process and $\frac{\partial}{\partial\beta_*}\hat{U}(\beta_*, 1)$ converges to a finite quantity in probability.

The asymptotic distribution theroy of the $\hat{\beta}_{EPL}$ are summerized in the next two theorems. The outline of the proof of the theorems are given in Appendix. Rigirous proof can be found in Zhou(1992).

The following assumptions will be used to develope the asymptotic theory.

(A) $\int_0^1 \lambda_0(t)dt < \infty$

(B) $P(Y(1) = 1|Z = z_k) > 0 \quad k = 1, 2, \cdots, q$

(C) There exists an open subset $\mathcal{B}$, containing β_0, of the Euclidean p space $\mathcal{E}^p$. $r^{(2)}$ with elements $\frac{\partial^2}{\partial\beta_i\partial\beta_j}r(\beta, t)$ exists and is continuous on $\mathcal{B}$ for each $t \in [0, 1]$, uniformly in t, and $\bar{r}(\beta, t)$ is bounded away from 0 on $\mathcal{B} \times [0, 1]$. $\Sigma(\beta_0)$ is positive definite.

(D)

$$E\{sup_{\mathcal{B}\times[0,1]}|Y(t)r^{*(j)}(\beta, t)|\} < \infty \quad j = 0, 1, 2$$

$$E\{sup_{\mathcal{B}\times[0,1]}|Y(t)(\frac{r^{*(1)}(\beta, t)}{r^*(\beta, t)})^{\otimes 2j}r^*(\beta_0, t)|\} < \infty \quad j = 1, 2$$

$$E\{sup_{\mathcal{B}\times[0,1]}|Y(t)(\frac{r^{*(2)}(\beta, t)}{r^*(\beta, t)})^{\otimes j}r^*(\beta_0, t)|\} < \infty \quad j = 1, 2$$

(E)

$$sup_{0\leq t\leq 1}|Z_v^{(K)}(t)| = O_p(1) \ \ as \ \ K = 0, 1$$

where

$$Z_v^{(K)}(t) \equiv \sqrt{v}\{\frac{1}{v}\sum_{i=1}^{v} I_{[Y_i=1, Z_i=z]} r_i^{(K)}(\beta, t) - E(I_{[Y(t)=1, Z=z]} r^{(K)}(\beta, t))\} \ \ K = 0, 1$$

Theorem 1 *Under Conditions A - E, we have the following results:*

1. $\hat{\beta}_{EPL}$ *is a consistent estimator for* β_0.

2. $n^{\frac{1}{2}}(\hat{\beta}_{EPL} - \beta_0)$ *is asymptotically normally distributed with mean zero and covariance matrix* Σ_{EPL}, *where* $\Sigma_{EPL}(\beta_0) = \Sigma^{-1}(\beta_0)((1-\rho)\Sigma_1(\beta_0) + \rho\Sigma_2(\beta_0))\Sigma^{-1}(\beta_0)^T$ *and* Σ, Σ_1 *and* Σ_2 *are as defined in (8).*

Notice that the term Σ_2 in the asymptotic covariance matrix $\Sigma_{EPL}(\beta_0)$ is the term that captures the variablity induced by the non-validation sample in the EPL. When $\rho = 1$, i.e. covariate X is available to everyone in the sample, $\Sigma_{EPL}(\beta_0)$ is the same covariance matrix that obtained from the standard partial likelihood analysis; when $\rho = 0$, $\Sigma_{EPL}(\beta_0)$ will blow up to infinity. This makes sense because one can not determine the effect of X through Z without knowing the association between X and auxiliary Z.

The asymptotic covariance matrix can be easily estimated by estimating the components in the $\Sigma_{EPL}(\beta_0)$. The following theorem gives a consistent estimator of $\Sigma_{EPL}(\beta_0)$.

Theorem 2

$$\hat{\Sigma}_{EPL}(\hat{\beta}_{EPL}) \xrightarrow{p} \Sigma_{EPL}(\beta_0) \ \ as \ n \to \infty$$

where

$$\hat{\Sigma}_{EPL}(\hat{\beta}_{EPL}) = \hat{\Sigma}(\hat{\beta}_{EPL})^{-1}((1-\rho)\hat{\Sigma}_1(\hat{\beta}_{EPL}) + \rho\hat{\Sigma}_2(\hat{\beta}_{EPL}))(\hat{\Sigma}(\hat{\beta}_{EPL})^{-1})^T$$

$$
\begin{aligned}
\hat{\Sigma}_1(\beta) &= \int_0^1 \left(\sum_{k=1}^{q} \frac{\hat{\bar{r}}_k^{(1)}(\beta,t)^{\otimes 2}}{\hat{\bar{r}}_k(\beta,t)} \frac{1}{n} \sum_{j=1}^{n} Y_j(t) I_{[z_j=z_k]} - \frac{\hat{S}^{(1)}(\beta,t)^{\otimes 2}}{\hat{S}^{(0)}(\beta,t)} \right) d\hat{\hat{\Lambda}}_0(t) \\
\hat{\Sigma}_2(\beta) &= \frac{1}{v} \sum_{j=1}^{v} \left\{ \int_0^1 \left(\frac{r_i^{(1)}(\beta,t)}{r_i(\beta,t)} - \frac{\hat{S}^{(1)}(\beta,t)}{\hat{S}^{(0)}(\beta,t)} \right) dM_i(t) \right. \\
&\quad - \frac{1-\rho}{\rho} \int_0^1 \int_0^1 \left(\frac{\hat{\bar{r}}_i^{(1)}(\beta,t)}{\hat{\bar{r}}_i(\beta,t)} - \frac{\hat{S}^{(1)}(\beta,t)}{\hat{S}^{(0)}(\beta,t)} \right) \\
&\quad \left. \times Y_j(t)(r_i(\beta,t) - \hat{\bar{r}}_i(\beta,t) d\hat{\hat{\Lambda}}_0(t) \right\}^{\otimes 2} \\
\hat{\Sigma}(\beta) &= \int_0^1 \left(\hat{S}^{(3)}(\beta,t) - \frac{\hat{S}^{(1)}(\beta,t)^{\otimes 2}}{\hat{S}^{(0)}(\beta,t)} \right) d\hat{\hat{\Lambda}}_0(t) \\
d\hat{\hat{\Lambda}}_0(t) &= \frac{\sum_{i=1}^{n} dN_i(t)}{\sum_{i=1}^{n} Y_i(t) \hat{r}_i^*(\hat{\beta}_{EPL},t)} = \frac{1}{\hat{S}^{(0)}(\hat{\beta}_{EPL},t)} \frac{1}{n} \sum_{i=1}^{n} dN_i(t)
\end{aligned} \tag{10}
$$

4. ASYMPTOTIC RELATIVE EFFICIENCY

Asymptotic relative efficiency(ARE) of $\hat{\beta}_{EPL}$, relative to the maximum likelihood estimator $\hat{\beta}_{MLE}$ and the partial likelihood estimaor $\hat{\beta}_{VAL}$ are studied under the exponential model. The $\hat{\beta}_{VAL}$ is obtained from the analysis based on the validation sample only. ARE's are evaluated with covariate X a scalor variable that has uniform distribution on $[0,2]$. Conditional on the covariate X, the density function of failure time variable T and the corresponding hazard function takes the form,

$$
\begin{aligned}
f(T;X) &= \lambda_0 e^{\beta X} e^{-T\lambda_0 e^{\beta X}} \\
\lambda(t,X) &= \lambda_0(t) e^{\beta X(t)}
\end{aligned} \tag{11}
$$

The auxiliary covariate Z is a binary variable that takes value one if X plus a $N(0,\sigma^2)$ random error e is greater than or equals to one. Note that the variance σ^2 of the random error e is the parameter that determines the strength of association between X and Z. As the value of σ^2 increases, the corretation coefficient of X and Z decreases. Assuming random censoring mechanism, the censoring variable is taken as uniform distribution on $[0,c]$ where c is the parameter that determines the censoring percentage.

Table 1 givens the AREs for β to be 0 and $ln(2)$; validation fraction ρ to be 20%, 50% and 80%; censoring percentage to be 50%; and variance of the random error to be a sequence value between 0 and ∞. The correlation coefficient of X and Z for different σ^2's are presented in colum 3.

Some points can be noted from Table 1: the EPL analysis has nearly better than 80% efficiency in estimation when compared to the maximum likelihood analysis for most cases of practical importance. When the validation fraction is relatively high($\rho \geq 50\%$) very little is to be gained by making the more stringent parametric assumptions required for the maximum likelihood estimation. The asymptotic efficiency of EPL relative to MLE estimator decreases as the strength of association between X and Z decreases. Since the random sensoring is asummed, it is not surprising that at $\beta = 0$, EPL method is fully efficient with respect to the parametric approach. Similar properties holds for the partial likelihood method(Kalbflesh, 1974). On the other hand, the EPL estimator is asymptotically more efficient than that obtained form the partial likelihood analysis based on validation sample only. The efficiency gained by employing the EPL method versus validation analysis only could be substantial($ARE(\hat{\beta}_{EPL}|\hat{\beta}_{VAL}) = 4.07$) when the validation fraction is low, or the association between X and Z is strong. As the strength between X and Z geting weaker, the asymptotic efficiency of $\hat{\beta}_{EPL}$ relative to $\hat{\beta}_{VAL}$ reduces to one.

5. SIMULATION STUDIES AND REAL DATA ILLUSTRATION

The small sample performence of $\hat{\beta}_{EPL}$ are examined under the configuration in Section 4. This is done by means of the simulation studies. Special attention are paid to the factors that may affect the small sample properties: sample size n, the validation fraction ρ and the strength of association between X and Z. Only 20% and

50% of the validation fraction are considered. For each validation fraction, the value of sample sized n considered are 100, 200 and 400; In each each combination of n and ρ, the error variance considered are 0.01, 0.1 and 1.

For all the investigations, 1,000 realizations of data were generated in accordance with the above models. For each of these 1000 data realizations, the estimate $\hat{\beta}_{EPL}$ and its variance estimate was found.

Using the GAUSS (GAUSS, 1991) programming language the score equations were solved using Newton-Raphson iterative procedure to obtain estimates. The RNDSEED command was used to set the parameters of the GAUSS random number generator, thereby ensuring the reproducibility of results, and making sure that the different methods analyze the same sequence of data in order to compare the results.

Results from the simulation studies are presented in Table 2. The last column provides the small sample efficiency of $\hat{\beta}_{EPL}$ relative to the $\hat{\beta}_{VAL}$. As in the large sample situation, $\hat{\beta}_{EPL}$ is more precise than $\hat{\beta}_{VAL}$ in all situations considered. The small sample efficiency gain of $\hat{\beta}_{EPL}$ relative to $\hat{\beta}_{VAL}$ is even higher in the small sample. This efficiency gain decreases as the validation fraction increases or if the $corr(X, Z)$ decreases. The small sample bias of $\hat{\beta}_{EPL}$ appears to be reasonably small in that the contribution to the mean square error due to the bias is less than 1% in most cases considered. The most biased situation happened in the case where the total smple size is 100, the validation fraction is 20% and the association between X and Z is not strong($corr(X, Z) = 0.41$) in which cas the bias account for 4.2% of the mean square error of $\hat{\beta}_{EPL}$. The proposed variance estimator appears to have no clear bias. The nominal approximation to the $\hat{\beta}_{EPL}$ seems quit well in most situations in that actual coverage of probability of confidency interval are close to the nominal 90%.

5.1 SOLVD DATA ANALYSIS

In the SOLVD data example, the relative risk of death of hospitalization due to con-

gestive heart failure as a function of ejection fraction is estimated. The correlation coefficient of X and Z is about 0.41. The validation analysis based on 179 patients estimates the regression coefficient EF as -0.031 with 0.021 as the estimated standard error. The analysis that uses the non-standardized EF as covariate yields -0.070 with 0.008 as the estimated standard error. This is the approach currently used by the SOLVD analysis. However it is know that the use of mismeasured covariate instead of the true covariate will causs the β bias towards zero. Applying the EPL method, we estimate β as -0.084 and the estimated standard error as 0.020. The EPL method provides a significant estimate of the relative risk parameter while the validation analysis only failed to do so.

6. Discussion

In this work, we have developed an approach that draw inference about the relative risk parameter from the subjects with only the auxiliary covariate data. The work was completed under the assumption that the auxilary covariate is discrete and the validation sample is a simple random sample from the entire sample. However due to the nature of the EPL estimator, the random sample assumption can be relaxed in some situations, e.g. the validation sample can be a stratified random sample on Z. Furthermore the validation sample can be a time dependent subset. This latter case may be useful in the large cohort studies.

In practice, one should be cautious about the bias when the validation sample is small(≤ 40) and the association between X and Z is not stronge(≤ 0.4). In the above situation, two possible methods can be used to reduce the bias from EPL to level comparable to those obtained from validation analysis. First, one can increase the validation fraction by partially including the subjects with only auxiliary covariate in the estimation process. Secondly, by censoring the late stage subjects in the non-validation set to avoid the unstability in small sample.

The conditions in in Section 3 are quite general for the partial likelihood regression with generalized relative risk function(Prentice and Self, 1983). Condition (B) is the natual restriction that enable the EPL be carried out in large sample. Condition (C) and (D) are the continuity and the regularity conditions that enable us to extend the domain and range space of the score function to Banach spaces. Condition (E) is actually fullfilled in the quite general setting, e.g. it can be shown with morden empirical theory that (E) holds if X and Z is time independent and the relative risk function r is a monotone function of X. A survey of these theory can be found in Wellner(1992).

Some work need to be done to extend the method to the continuous auxuliary covariate situation. A possiable non-paramtetric regression may be used to accomplish this task. A kernal estimation method may be used in the continuours case. (e.g. Carroll and Wand, 1991)

References

[1] Andersen, 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.

[2] Breslow, N. E. (1972), "Contribution to Discussion of Paper by D. R. Cox", *Journal of the Royal Statistical Society*, B, 34, 216-217.

[3] Breslow, N. E. (1974), "Covariance Analysis of Censored Survival Data", *Biometrics*, 30, 89-99

[4] Caroll, R. J. and Wand, M. P. (1991), "Semiparametric Estimation in Logistic Measurement Error Models", *J. R. Statist. Soc. B*, 53, 573-585

[5] Chung (1977), *propability*, Holden-Day, Inc., San Francisco.

[6] Fleming, T.R. and Harrington, D.P. (1991), *Counting Processes and Survival Analysis*, Wiley, New York.

[7] Foutz, R. V. (1977), "On the Unique Consistent Solution to the Likelihood Equations", *Journal of the American Statistical Association*, 72, 147-148.

[8] GAUSS System, Version 2.1 (1991), Kent, WA: Aptech Systems Inc.

[9] Pepe, M. S., Self, S. G. and Prentice, R. L. (1989), "Further Results on Covariate Measurement Errors in Cohort Studies with Time to Response Data", *Statistics in Medicine*, 8, 1167-1178.

[10] Prentice, R. L. (1982), "Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model" *Biometrika* 69, 2, 331-342

[11] Prentice, R. L. and Self, S.S (1983), "Asymptotic Distribution Theory for Cox-Type Regression Models with General Relative Risk Form", *The Annals of Statistics*, Vol. 11, 3, 804-813.

[12] Rudin, W (1964), *Principles of Mathematical Analysis*, New York: McGraw-Hill Book Co..

[13] Shorack, G "Probability", manuscript

[14] The SOLVD Investors (1991) "Effect of Enalapril on Surbival In patients with reduced Left Ventricular Ejection Fraction and Congestive Heart Failure", *New England Iournal of Madicine* , 325:293-302(August 1), 1991

[15] Jon Wellner Manuscript. "Empirical Processes In Action: A Review"

[16] Haibo Zhou (1992), "Auxiliary Covariate Data in Failure Time Regression Analysis" *Ph.D Dissertation, Unversity of Washington*,

APPENDIX

Outline proofs of Proof of Theorem 1

Existence of a unique consistent solution to the EPL equations is obtained as a consequence of the Inverse Function Theorem(Rudin & Walter, 1964, p. 193). The following theory and its outline proof is an extension of standard likelihood(Foutz, 1977). OUTLINE PROOF OF CONSISTENCY: By Condition (C), $-\Sigma(\beta_0)$ is negative definite, we may define $\lambda = \frac{1}{4\|-\Sigma^{-1}(\beta_0)\|}$. Since it can be shown that $\frac{1}{n}\frac{\partial}{\partial\beta}\hat{U}(\beta)$ converges uniformaly to $-\Sigma(\beta)$, we can choose δ sufficiently small that the above uniform convergence holds and $\|\Sigma(\beta_0) - \Sigma(\beta)\| < \frac{\lambda}{2}$ whenever $|\beta - \beta_0| < \delta$. Denote the neighborhood of β_0 with radius δ as $\mathbf{U}_\delta$.

Also since $\frac{\partial}{\partial\beta}\hat{U}(\beta_0)$ converges to $-\Sigma(\beta_0)$ in probability, it ensures that $\frac{1}{n}\frac{\partial}{\partial\beta}\hat{U}(\beta_0)$ is negative definite with probability going to one. Let $\lambda_n = \frac{1}{4\|\frac{\partial}{\partial\beta}\hat{U}^{-1}(\beta)\|}$ whenever $\frac{1}{n}\frac{\partial}{\partial\beta}\hat{U}(\beta)$ is negative definite. Then $\lambda_n \xrightarrow{p} \lambda$. Hence we have

$$\begin{aligned}\|\frac{1}{n}\frac{\partial}{\partial\beta}\hat{U}(\beta) - \frac{1}{n}\frac{\partial}{\partial\beta}\hat{U}(\beta_0)\| \;&\leq\; \|\frac{1}{n}\frac{\partial}{\partial\beta}\hat{U}(\beta) - (-\Sigma(\beta))\| \\ &\quad + \|\Sigma(\beta) - \Sigma(\beta_0)\| + \| - \Sigma(\beta_0) - \frac{1}{n}\frac{\partial}{\partial\beta}\hat{U}(\beta_0)\| \\ &< \;\lambda < 2\lambda_n\end{aligned}$$

with probability going to one as $n \to \infty$ if $|\beta - \beta_0| < \delta$. By the Inverse Function Theorem(Rudin &Walter, 1964) $\frac{1}{n}\hat{U}$ is a one-to-one function from $\mathbf{U}_\delta$ on to $\frac{1}{n}\hat{U}(\mathbf{U}_\delta)$ and the image set $\frac{1}{n}\hat{U}(\mathbf{U}_\delta)$ contains the open neighborhood of radius $\lambda_n\delta$ about $\frac{1}{n}\hat{U}(\beta_0)$ with probability going to one.

Since $\frac{1}{n}\hat{U}(\beta_0)$ goes to zero in probability, we can see that $0 \in \hat{U}(\mathbf{U}_\delta)$ with probability going to 1. Also we have that $|\frac{1}{n}\hat{U}(\beta_0) - 0| < \frac{\lambda\delta}{2}$ with probability going to one as $n \to \infty$.

Consider the inverse function $\frac{1}{n}\hat{U}^{-1} : \frac{1}{n}\hat{U}(\mathbf{U}_\delta) \to \mathbf{U}_\delta$. It is well defined when $\frac{1}{n}\hat{U}$ is one-to-one, i.e. with probability going to one, since $0 \in \hat{U}(\mathbf{U}_\delta)$ with probabilty going

to one we may conclude:

1. the root, $\frac{1}{n}\hat{U}^{-1}(0)$, of the estimating equation exists in $\mathbf{U}_\delta$ with probability going to one;

2. since δ may be taken arbitrarily small, $\frac{1}{n}\hat{U}^{-1}(0)$ converges in probability to β_0;

3. by one-to-oneness of $\frac{1}{n}\hat{U}$ on $\mathbf{U}_\delta$, any other sequence $\{\bar{\beta}\}$ of roots to $\hat{U}(\beta) = 0$ necessarily lies outside of $\mathcal{U}_\delta$ with probability going to 1 which implies that the sequence does not converge to β_0.

Therefore $\hat{\beta}_{EPL} = \frac{1}{n}\hat{U}^{-1}(0)$ is a unique consistent estimator of β_0. □

OUTLINE PROOF OF NORMALITY: The asymptotic normality of $\hat{\beta}_{EPL}$ is obtained as a results of two steps. The first step is showing that $n^{-\frac{1}{2}}\hat{U}(\beta_0, 1)$ converges to a normal distribution as n goes to infinity. It can be shown(Zhou, 1992) that

$$\begin{aligned} & n^{-\frac{1}{2}}\hat{U}(\beta_0, 1) \\ = \; & n^{-\frac{1}{2}} \int_0^1 \sum_{i=1}^{\bar{v}} \left(\frac{\bar{r}_i^{(1)}(\beta_0, \omega)}{\bar{r}_i(\beta_0, \omega)} - \frac{s^{(1)}(\beta_0, \omega)}{s^{(0)}(\beta_0, \omega)} \right) dM_i(\omega) \\ & + n^{-\frac{1}{2}} \sum_{j=1}^{v} \left[\int_0^1 \left(\frac{r_j^{(1)}(\beta_0, \omega)}{r_j(\beta_0, \omega)} - \frac{s^{(1)}(\beta_0, \omega)}{s^{(0)}(\beta_0, \omega)} \right) dM_i(\omega) - \frac{\bar{v}}{v} Q_j \right] + o_p(1) \end{aligned} \quad (12)$$

The first term and the second term on the right hand side are independent. By the martingale central limit theorem(Fleming & Harrison, 1991), the first term in (12) converges weakly to a continuous normal process. The covariance process of this normal process evaluated at $t = 1$ is $(1 - \rho)\Sigma_1$. i.e.

$$\begin{aligned} < W > (t) \quad &= \quad \int_0^t n^{-1} \sum_{i=1}^{\bar{v}} \left(\frac{\bar{r}_i^{(1)}(\beta_0, \omega)}{\bar{r}_i(\beta_0, \omega)} - \frac{s^{(1)}(\beta_0, \omega)}{s^{(0)}(\beta_0, \omega)} \right)^2 Y_i(\omega) \bar{r}_i(\beta_0, \omega) \lambda_0(\omega) d\omega \\ &\xrightarrow{p} \quad (1 - \rho) \int_0^t \left[E\left(\frac{\bar{r}_i^{(1)}(\beta_0, \omega)^{\otimes 2}}{\bar{r}_i(\beta_0, \omega)} Y_i(\omega) \right) - \frac{s^{(1)}(\beta_0, \omega)^{\otimes 2}}{s^{(0)}(\beta_0, \omega)} \right] \lambda_0(\omega) d\omega \\ < W > (1) \quad &\equiv \quad (1 - \rho)\Sigma_1 \end{aligned}$$

The second term in $n^{-\frac{1}{2}}\hat{U}(\beta_0, 1)$ is also a summation of iid terms from subjects in the validation sample. By the Central Limit Theorem, it converges to a normal distribution with mean

$$E\left\{\int_0^1 \left(\frac{r_j^{(1)}(\beta_0,\omega)}{r_j(\beta_0,\omega)} - \frac{s^{(1)}(\beta_0,\omega)}{s^{(0)}(\beta_0,\omega)}\right) dM_j(\omega) - \frac{\bar{v}}{v}Q_j\right\} \tag{13}$$

and covariance

$$\rho Var\left\{\int_0^1 \left(\frac{r_j^{(1)}(\beta_0,\omega)}{r_j(\beta_0,\omega)} - \frac{s^{(1)}(\beta_0,\omega)}{s^{(0)}(\beta_0,\omega)}\right) dM_j(\omega) - \frac{1-\rho}{\rho}Q_j\right\} \tag{14}$$

The first term in the mean expression (14) is a local martingale and expected value of a local martingale is zero. The second term is also zero, since

$$\begin{aligned}
& E\left\{-\frac{\bar{v}}{v}Q_j\right\} \\
= & -\frac{\bar{v}}{v}E\int_0^1 Y_j(\omega)(r_j(\beta_0,\omega) - \bar{r}_j(\beta_0,\omega))\left(\frac{\bar{r}_j^{(1)}(\beta_0,\omega)}{\bar{r}_j(\beta_0,\omega)} - \frac{s^{(1)}(\beta_0,\omega)}{s^{(0)}(\beta_0,\omega)}\right)\lambda_0(\omega)d\omega \\
= & -\frac{\bar{v}}{v}\int_0^1 E\left\{\left(\frac{\bar{r}_j^{(1)}(\beta_0,\omega)}{\bar{r}_j(\beta_0,\omega)} - \frac{s^{(1)}(\beta_0,\omega)}{s^{(0)}(\beta_0,\omega)}\right)\right. \\
& \left.\times E\left[Y_j(\omega)r_j(\beta_0,\omega) - Y_j(\omega)\bar{r}_j(\beta_0,\omega)|_{Y_j(\omega)=1,Z_j=z_j}\right]\right\}\lambda_0(\omega)d\omega \\
\equiv & \ 0
\end{aligned}$$

The covariance matrix can be expressed as

$$\rho\Sigma_2 \equiv \rho E\left\{\int_0^1 \left(\frac{r_j^{(1)}(\beta_0,\omega)}{r_j(\beta_0,\omega)} - \frac{s^{(1)}(\beta_0,\omega)}{s^{(0)}(\beta_0,\omega)}\right) dM_j(\omega) - \frac{1-\rho}{\rho}Q_j\right\}^{\otimes 2}.$$

Since the two terms in $n^{-\frac{1}{2}}\hat{U}(\beta_0,\omega)$ are from the validation and non-validation sets respectively the limiting distribution of $n^{-\frac{1}{2}}\hat{U}(\beta_0,\omega)$ is normal with mean zero and covariance matrix $(1-\rho)\Sigma_1 + \rho\Sigma_2$, where Σ_1 and Σ_2 are as defined in (8).

The second step in proving the asymptotic nomality of $n^{\frac{1}{2}}(\hat{\beta}_{EPL} - \beta_0)$ is to show that $-n^{-1}\frac{\partial}{\partial\beta}\hat{U}(\beta,\cdot)|_{\beta=\beta^*}$ converges to some positive definite quantity. Recall that $-n^{-1}\frac{\partial}{\partial\beta}\hat{U}(\beta,1)\xrightarrow{p}\Sigma(\beta)$ for any $\beta\in\mathcal{B}$ and that $\Sigma(\beta_0)$ is positive definite, where

$$\Sigma(\beta_0) = \int_0^1 \left[\frac{s^{(2)}(\beta_0,\omega)}{s^{(0)}(\beta_0,\omega)} - \left(\frac{s^{(1)}(\beta_0,\omega)}{s^{(0)}(\beta_0,\omega)}\right)^{\otimes 2}\right] s^{(0)}(\beta_0,\omega)\lambda_0(\omega)d\omega$$

Since $\Sigma(\beta)$ is continuous in β(Zhou 1992), we have

$$\begin{aligned} & \left| -\frac{1}{n}\frac{\partial}{\partial\beta}\hat{U}(\beta,1)|_{\beta=\beta^*} - \Sigma(\beta_0) \right| \\ \leq & \left| -\frac{1}{n}\frac{\partial}{\partial\beta}\hat{U}(\beta,1)|_{\beta=\beta^*} - \Sigma(\beta^*) \right| + |\Sigma(\beta^*) - \Sigma(\beta_0)| \end{aligned}$$

The first term on the right hand side of the inequality converges to zero in probability as $n \to \infty$. The second term converges to zero as well by the continuity of Σ and the fact that β^* is between $\hat{\beta}_{EPL}$ and β_0, and that $\hat{\beta}_{EPL}$ is consistent for β_0. Therefore

$$-n^{-1}\frac{\partial}{\partial\beta}\hat{U}(\beta,1)|_{\beta=\beta^*} \xrightarrow{p} \Sigma(\beta_0) \text{ as } n \to \infty$$

where $\Sigma(\beta_0)$ is positive definite. □

Table 1. Asymptotic Relative Efficiency of EPL Method Relative to the Validation Anlysis and Fully Parametric Analysis Under the Expnontial Failure Time Model and 50% Censoring Percentage. The distribution of failure Time T is of form $f_{\lambda_0,\beta}(T;X) = \lambda_0 e^{\beta X} e^{-T\lambda_0 e^{\beta X}}$

			$\beta = 0$		$\beta = in2$	
		corr	ARE	ARE	ARE	ARE
ρ	σ^2	(X,Z)	$\beta_{EPL}\|\beta_{VAL}$	$\beta_{EPL}\|\beta_{MLE}$	$\beta_{EPL}\|\beta_{VAL}$	$\beta_{EPL}\|\beta_{MLE}$
20%	0.00001	0.87	4.07	1.00	3.56	0.90
	0.05	0.82	3.71	1.00	3.12	0.87
	0.1	0.78	3.43	1.00	2.82	0.86
	0.2	0.70	2.96	1.00	2.36	0.84
	0.5	0.54	2.19	1.00	1.72	0.81
	1	0.41	1.70	1.00	1.38	0.79
	2	0.31	1.39	1.00	1.19	0.77
	5	0.20	1.16	1.00	1.07	0.77
	10000	0.01	1.00	1.00	1.00	0.77
50%	0.00001	0.87	1.77	1.00	1.69	0.96
	0.05	0.82	1.68	1.00	1.59	0.96
	0.1	0.78	1.61	1.00	1.52	0.95
	0.2	0.70	1.49	1.00	1.40	0.94
	0.5	0.54	1.30	1.00	1.23	0.93
	1	0.41	1.18	1.00	1.13	0.92
	2	0.31	1.10	1.00	1.07	0.92
	5	0.20	1.04	1.00	1.03	0.92
	10000	0.01	1.00	1.00	1.00	0.91
80%	0.00001	0.87	1.19	1.00	1.18	0.98
	0.05	0.82	1.17	1.00	1.15	0.98
	0.1	0.78	1.15	1.00	1.13	0.97
	0.2	0.70	1.12	1.00	1.11	0.97
	0.5	0.54	1.07	1.00	1.06	0.97
	1	0.41	1.04	1.00	1.04	0.97
	2	0.31	1.02	1.00	1.02	0.96
	5	0.20	1.01	1.00	1.01	0.96
	10000	0.01	1.00	1.00	1.00	0.96

NOTE: X is uniform(0,2). Z is a binary variable that takes value 1 if $X + e$ greatter than or equals to 1. The error e is independent of X and has normal(0, σ^2). ρ is the asymptotic validation fraction. β_{VAL} is obtained from the partial likelihood analysis based on the validation set only. β_{MLE} is obtained from the maximum likelihood analysis based on the fully and correctly specified underline distributions.

Table 2. Monte Carlo Study For Comparing EPL Estimates With VAL Estimates When $\beta = ln2$ and 50% Censoring Percentage Under the Same Model in Table 1. 1000 Data Sets Were Generated and the Same Data Set Was Analyzed by Both Methods.

ρ	n	σ^2	Mean $\hat{\beta}$	Var $\hat{\beta}$	Mean $\hat{\Sigma}$	90%$C.I.$ Coverage	$\widehat{ARE}$ $EPL\|VAL$
20%	100	0.01	0.715	0.122	0.111	0.888	4.172
		0.1	0.717	0.124	0.115	0.893	4.106
		1	0.587	0.257	0.280	0.840	1.992
		VAL	0.740	0.512	0.472	0.926	1.000
	200	0.01	0.699	0.054	0.050	0.892	4.023
		0.1	0.698	0.065	0.062	0.904	3.309
		1	0.656	0.118	0.120	0.871	1.838
		VAL	0.739	0.217	0.193	0.892	1.000
	400	0.01	0.698	0.025	0.024	0.886	3.382
		0.1	0.699	0.029	0.029	0.893	2.934
		1	0.673	0.056	0.058	0.898	1.535
		VAL	0.702	0.086	0.086	0.892	1.000
50%	100	0.01	0.705	0.091	0.083	0.881	1.751
		0.1	0.707	0.091	0.084	0.869	1.751
		1	0.675	0.131	0.122	0.883	1.221
		VAL	0.702	0.160	0.147	0.896	1.000
	200	0.01	0.700	0.043	0.040	0.886	1.690
		0.1	0.697	0.045	0.044	0.885	1.618
		1	0.697	0.063	0.058	0.881	1.162
		VAL	0.708	0.073	0.068	0.902	1.000
	400	0.01	0.699	0.020	0.020	0.900	1.544
		0.1	0.699	0.021	0.021	0.899	1.496
		1	0.689	0.028	0.029	0.907	1.122
		VAL	0.692	0.031	0.033	0.918	1.000

NOTE: The small sample simulation is under the same configuration as in Table 1. n is the sample size and ρ is the validation fraction. The rows for $\sigma^2 = 0.01, 0.1, 1$ are cooresponding to the results of EPL. Var(β) is the true sample variance based on 1000 simulation results. Mean($\hat{\Sigma}$) is mean of the variance estimates. $\widehat{ARE}$ is the ratio of the true variance of EPL and VAL.