THE PHGLM PROCEDURE FOR TIME-DEPENDENT COVARIATES AND THE CASE-COHORT DESIGN

DAVID YU-WU PEE, IMS, INC.
SHOLOM WACHOLDER, NCI

Abstract

The PHGLM procedure is a well documented, efficient and widely used program for fitting the proportional hazards model in survival studies. We have developed some new extensions of the procedure which are useful in some applications to chronic disease epidemiology. One extension allows covariates to vary over time and allows left truncation in addition to the standard right censoring. The other application allows estimation of the hazard ratio from a case-cohort design. Standard variance estimation methods do not apply to this design, but we suggest a bootstrap procedure to estimate the variance of the hazard ratio and obtain confidence intervals. The proposed extensions are illustrated by worked example from two data sets.

Introduction

The Cox proportional hazards model \([1,8]\) is widely used to analyze survival time data. The model assumes that the failure hazard at time \(t\) for a subject with observed covariate value of \(Z\) is,

\[
\lambda(t; Z) = \lambda_0(t) \exp(Z\beta) \tag{1}
\]

where \(\lambda_0(t)\) is an arbitrary unspecified baseline hazard function and \(\beta\) is the unknown regression parameter of interest. Note that \(Z\) and \(\beta\) can be vectors. Furthermore, \(Z\) is allowed to vary over time.

Cox suggested that in the absence of knowledge about \(\lambda_0(t)\), the regression parameter \(\beta\) could be estimated by standard likelihood techniques applied to the partial likelihood [Cox,Z]. The partial likelihood is given as:

\[
L_C(\beta) = \prod_{i=1}^{N} \left[ \frac{\exp(Z_i \beta)}{\sum_{j \in R_i} \exp(Z_j \beta)} \right]^\delta_i \tag{2}
\]

where the \(i\)th individual is followed until time \(t_i\) and has covariate value \(Z_i\), \(R_i\) denotes the subjects at risk at \(t_i\) and \(\delta_i = 1\) if an event like death occurs at \(t_i\) and 0 otherwise. We assume that \(t_1 \leq \ldots \leq t_N\) and no ties are observed for the \(d\) events. All covariates \(Z_i\) are evaluated at the times \(t_i\) corresponding to risk set \(R_i\).

Two Simplifying Conditions

In maximizing the partial likelihood, the following quantities are needed in the calculations of the contribution to the total score and sample information for each event time. The total score and sample information are then obtained by summing over their corresponding contributions from all of the events.

Two simplifying conditions for this maximization problem are:

(a) The covariate \(Z\) is a constant in time.

(b) The risk sets \(R_i\) are proper subsets of prior risk sets.

When these two conditions are satisfied, efficiency in the design and execution of the Cox regression software is obtained by observing that the quantities in (3) consists of telescoping sequences with respect to reverse time order. Working backward in time, the sums in (3) are calculated for the last event, the sums for other events do not have to be recalculated but can be obtained by recursive addition of the contribution of the members of the current risk set who are not at risk at the time of the previous event. Most Cox regression software, and presumably, the PHGLM procedure [Harrell,3], are coded assuming these two conditions.

Time-dependent Covariates

When condition (a) is relaxed such that one or more of the measured covariates are allowed to vary over time, one is then faced with time-dependent covariates. Even for the case when (b) is satisfied, such that the risk sets \(R_i\) consists of proper subsets of
prior risk sets, the quantities in (3) do not possess their usual telescoping properties. Thus, in a concise program for time-dependent covariates, the sums in (3) need to be reaccumulated for every event time. Usual Cox regression programs, require modification when standard one record per subject file structure is used, since they were designed to take advantage of assuming (a) and (b).

Instead of program modification, we propose file structure modification and call this complete risk set assembly. This method requires identification of each member of the risk sets for every unique event time. The ensuing file will be made up of groups of records, where the first group of records will be made up of members of the first risk set, the next group of records will be made up of members of the second risk set and so forth until all risk sets are exhausted. All records from the same risk set are indexed by the same unique risk set identifier. For each risk set, the event indicator, $s$, is set to 1 for the subject with event occurring at the event time associated with the risk set. All other event indicators are set to zero.

Covariate information for time-dependent covariates must be evaluated at the event time corresponding to the risk set, for all members of the risk set.

The PHGLM procedure will handle this data set as if it came from a matched case-control study. Conditional logistic regression analysis can be performed by a stratified Cox analysis procedure [Breslow et al., 4]. An example using PHGLM is given in Harrell [3]. The partial likelihood (2) is identical in form to a conditional likelihood, where, the events acts as "cases" and the respective risk sets, excluding the referent event, act as "matched controls". All time dependent covariates are evaluated in each term of the "conditional" likelihood at the time of the event. Note that some of the subjects may act as "controls" many times. Thus, as suggested by Harrell [5], we can, use PHGLM and the complete risk set assembly method to make inference for time-dependent survival data.

This full flexibility of case and control definition is often needed in studies of time-dependent covariates because data on covariates may be missing at various times causing people to enter and leave various risk sets. As an example, suppose serial Carcinomaembryonic Antigen (CEA) measurements are collected to detect the recurrence of colorectal cancer following resection. Care must be taken that the schedule for CEA measurements is independent of observed survival status. The time-dependent covariate of interest is defined as the most recent CEA measurement falling within a time window for each event time (See Gail [6] for complete details). As often occurs, patients might not have a valid CEA measurement which satisfies the time window definition for various event times. Thus patients can drop in and out of risk sets due to "intermittent censoring" of time-dependent covariate values, thereby violating assumption (b). This situation is handled easily by the proposed method since the risk sets are constructed at each event time. Furthermore, thoughtful systems design coupled with the power of the SAS system allows one to consider even extremely complex time-dependent covariates, censoring or left truncation.

As an example of the analysis of time-dependent covariates, we consider the Stanford Heart Transplant data set of Crowley and Hu [7], and reproduced in the 1985 SHOP Manual [9, pages 585]. The model presented entertains transplant status, age X transplant status interaction and mismatch X transplant status interaction as covariates. A complete listing of the SAS codes used for this analysis is given in the appendix. Table 1 compares the estimates of $\beta$ and their estimated variance as reported in SHOP with those obtained by the PHGLM procedure. Excellent agreement was achieved.

### Table 1

<table>
<thead>
<tr>
<th>Covariate</th>
<th>BMDP</th>
<th>PHGLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transplant</td>
<td>-3.1780</td>
<td>-3.1781</td>
</tr>
<tr>
<td>Status</td>
<td>(1.1861)</td>
<td>(1.1861)</td>
</tr>
<tr>
<td>Age</td>
<td>0.0552</td>
<td>0.0552</td>
</tr>
<tr>
<td>Transplant</td>
<td>(.0226)</td>
<td>(.0226)</td>
</tr>
<tr>
<td>Interaction</td>
<td>0.4442</td>
<td>0.4442</td>
</tr>
<tr>
<td>Transplant</td>
<td>(.2803)</td>
<td>(.2803)</td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-275.9557</td>
<td>-275.955</td>
</tr>
<tr>
<td>Score Test</td>
<td>9.01</td>
<td>9.01</td>
</tr>
</tbody>
</table>

*Parameter estimate above, estimate of variance in parenthesis.*

---

Case-Cohort Design

The case-cohort design proposed by Prentice [10] is an economical alternative to standard cohort designs in epidemiology when the event of interest is rare. The economy is realized by restricting the need for assembling covariate history to the events plus a small
Table 2

Schematic of Case-Cohort Design

<table>
<thead>
<tr>
<th>Events</th>
<th>Non Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subcohort</td>
<td>q</td>
</tr>
<tr>
<td>Other</td>
<td>d-q</td>
</tr>
</tbody>
</table>

N = Size of cohort
M = Size of subcohort
d = # of observed events

subcohort of randomly selected subjects (Table 2). All members of the cohort are followed until "event" or censoring, while only the events and members of the subcohort are used for the estimation of \( \beta \). Due to the concentration of effort, better quality data might result from such a design. Furthermore, the same subcohort can be used as the comparison group for various definitions of "event". Finally, the subcohort provides a handy means by which the study can be monitored. In depth discussion of the case-cohort design can be found in [10].

One important difference between the case-cohort design and a full cohort design is that the risk set at each event time for the case-cohort design consists only of the subject with the event for that particular event time and members in the subcohort who are at risk at the event time. This definition of risk set causes the violation of condition (b). More importantly, Prentice [10] showed that the components of the total score for different event times are often correlated. This correlation invalidates the usual estimate of the covariance matrix for \( \beta \). He also gave a closed form expression for the asymptotic variance of \( \beta \). This expression is difficult to code and time consuming to process.

Case-cohort estimation can also be handled in the PHGLM procedure by assembling each risk set and mimicking a matched case control analysis. Here, the "controls" for the jth event consists of the members of the subcohort at risk at \( t_j \); if the jth event is in the subcohort, he/she may act as a "control" for the other events, otherwise he/she never appears as a "control". Note that if case-cohort data is analyzed in this fashion with the PHGLM procedure, the resulting estimate of \( \beta \) will be unbiased, but the variance estimate will be incorrect due to the correlations in the scores. As an alternative to the Prentice variance, Wacholder et al. [11] suggest a bootstrap sampling scheme to obtain an estimate of the variance of \( \beta \). This sampling scheme is a generalization of the one described by Efron and Tibshirani [12].

A Bootstrap Sampling Scheme for the Case-Cohort Design

The proposed bootstrap scheme consists of two steps. The first step in each bootstrap replication, is to obtain a bootstrap sample of d events by obtaining a simple random sample with replacement from the observed d events. The first q sampled events will be assigned to be members of the subcohort. The remaining sampled events will necessarily fall outside of the subcohort. The second step consists of obtaining simple random sample with replacement of size M-q from the original M-q non-events in the subcohort. These sampled M-q subjects will constitute the non-events in the bootstrap subcohort. Notice that this procedure preserves the cell counts and marginals in the informative part of Table 2. With the assembled bootstrap realization of the case-cohort data, a bootstrap estimate of \( \beta \) can be obtained. If this procedure is repeated B times, the bootstrap variance estimator is obtained from

\[
\sigma^{2}_{\text{Boot}} = \frac{\sum_{b} (\hat{\beta}_b - \bar{\beta})^2}{(B-1)}
\]  

where \( \bar{\beta} \) is the average of the B bootstrap estimates of \( \beta \), and \( \hat{\beta}_b \) is the bth bootstrap estimator.

A convenient method to obtain the bootstrap variance is demonstrated by a SAS program available from the first author. The SAS program invokes a macro procedure to perform the repetitive parts of the bootstrap scheme.

In order to study the operating characteristic of the case-cohort procedure as well as two alternative variance estimators, a FORTRAN program (hereafter referred to as FORTRAN) was developed by Wacholder et al. [11] to simulate data from a hypothetical case-cohort setting. The program proceeds to analyze the generated data. Complete detail can be obtained from [11]. We present a small simulation for demonstration purposes as well as to obtain a data set to document the methodology. The parameters of the simulation were selected so that the cohort was of size 2000 and included 200 exposed individuals. The relative hazard
of exposed subjects with respect to unexposed subjects was two. The subcohort was of size 100. A constant competing risk for all subjects was selected with a hazard 3 times greater than the failure hazard in the unexposed group. All subjects were assumed to be accrued at the beginning of the study, and the trial time was selected to realize 150 expected events [13]. The data were simulated 100 times and we choose to use data from the 29th simulation. This simulation achieved 152 events with 126 events seen in the unexposed group and 26 events seen in the exposed groups. Expected numbers of events for unexposed and exposed groups are 124 and 26, respectively. Table 3 presents the comparison of the analyses by FORTRAN and PHGLM. *Full cohort* refers to the analysis of the full cohort data set by standard Cox regression methods. *Case-cohort* uses only subjects from the subcohort and events. In FORTRAN we apply the method of Prentice [10]. In PHGLM we used the method of complete risk set assembly. Bootstrap displays the bootstrap variance for $\beta$ with repetition of 200. Finally, *Naive Cox* displays the results of the inappropriate analysis of case-cohort data by standard Cox methods. It is seen that for the full cohort design both FORTRAN and PHGLM give identical estimates for $\beta$ and $\sigma^2$. Moreover, while the case-cohort analysis again provides identical $\beta$, the estimates of $\sigma^2$ are quite different, as was expected. The differences in the corresponding bootstrap moments can be explained by the differences in random number generators used by the two different systems. The naive Cox analysis is provided to illustrate the dangers of applying the wrong type of analysis to case-cohort data.

### Table 3

<table>
<thead>
<tr>
<th>Design</th>
<th>PRENTICE** METHOD</th>
<th>COMPLETE*** METHOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Cohort</td>
<td>0.6872 (.0450)</td>
<td>0.6872 (.0450)</td>
</tr>
<tr>
<td>Case Cohort</td>
<td>0.7658 (.1628)</td>
<td>0.7658 (.0452)</td>
</tr>
<tr>
<td>Bootstrap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance B=200</td>
<td>0.1746 (.1465)</td>
<td></td>
</tr>
<tr>
<td>Naive Cox</td>
<td>0.3567 (.0452)</td>
<td>0.3567 (.0452)</td>
</tr>
</tbody>
</table>

**Parameter estimate above, estimate of variance in parenthesis.

**FORTRAN program based on [11].

***PHGLM using method of complete risk set assembly.

Finally, Table 4 provides the results for all 100 simulations as performed by FORTRAN. There is good agreement between asymptotic theory and simulations. The proposed bootstrap case-cohort variance is also seen to be a viable alternative to the Prentice variance. Furthermore, bootstrap
replications as low as 50 seems to do as well as bootstrap replications of 200. Similar results based on a larger simulation are found in [11].

Summary

An ad hoc procedure which demonstrates the ability to accommodate various experimental designs in survival time data studies has been discussed. When the experimental design generates data in which the covariates are constant in time and the risk sets are proper subsets of all prior risk sets, standard Cox regression programs can be applied directly to the data for estimating and making inferences about \( \beta \). For experimental designs which relax these two assumptions, the proposed method calls for complete assembly of risk sets for all event times and the analysis of the resulting data set by a stratified Cox regression program.

In particular, when the Cox regression program used for the analysis is the **PGCLM** procedure, the following two benefits are seen for this method:

1) Various types of designs can be accommodated within the same software system. This eliminates any overhead associated with going outside of the system to perform analysis.

2) The full capability of the SAS system is always available to the user for pre- and post-processing of the data. Repetitive and often performed routines can be automated by using the SAS macro facility.

Since the procedure is ad hoc in nature, it does suffer from loss of processing efficiency when compared to other dedicated software which might take advantage of task specific shortcuts. Time-dependent Cox regression analysis displays low to moderate loss in processing efficiency while the bootstrap procedure for the case-cohort design can be expensive. Also note that preprocessing to assemble the risk sets is always required and that for large data sets, significant amount of disk storage space must be available.

We feel that the above procedure proves adequate for investigators who only infrequently analyze survival data wherein assumptions (a) and (b) are violated. For those investigators who have constant need to analyze data where either assumption is relaxed, it may prove beneficial to develop customized software which was designed to meet his/her particular needs.

Acknowledgements

This work was supported by NCI Contract No. NCI-CP-71011 for the first author. The authors thank Dr. Mitchell Gail, of NCI, for helpful comments and Mrs. Sandra Kline, of IMS, for excellent secretarial support.

References


Correspondence Address:
David Pee
IMS, Inc.
1400 Spring Street, Suite 500
Silver Spring, MD 20910
(301) 495-0440

SAS and SAS/GRAPH are registered trademarks of SAS Institute, Inc., Cary, N.C., U.S.A.

Appendix

DATA BMDPSTAN;
INFILE DATAIN;
INPUT ID 1-5 SURVTIME 4-8 EVNT_IND 9-10 WAITIME 11-14 TRANSAGE 15-17 MISMATCH 18-22 2;
* FILLING IN THE WAITING TIME OF NON-TRANSPLANT SUBJECTS WITH A NON-INFORMATIVE TIME;
IF (WAITIME EQ .) THEN WAITIME = SURVTIME + 1;

PROC SORT DATA=BMDPSTAN;
BY DESCENDING SURVTIME EVNT_IND;

DATA UNIQUEVN (KEEP=EVNT_NUM EVENTIME);
SET BMDPSTAN;
RETAIN EVNT_NUM EVENTIME;
* CREATING A DATA SET OF UNIQUE EVENT TIMES;
IF (EVNT_IND EQ 1) THEN DO;
IF (SURVTIME NE EVENTIME) THEN DO;
EVNT_NUM = EVNT_NUM + 1;
EVENTIME = SURVTIME;
OUTPUT UNIQUEVN;
END;
END;

DATA BIGSET (DROP = I);
SET BMDPSTAN;
* FOR EACH SUBJECT DUPLICATE HIS HER RECORD AS MANY TIME AS THERE ARE UNIQUE EVENTS;
DO I = 1 TO 60;
EVNTj(UM = I);
OUTPUT;
END J;

PROC SORT DATA=BIGSET;
BY EVNT_NUM DESCENDING SURVTIME EVNT_IND;

DATA RISKSETS (DROP = I);
MERGE BIGSET UNIQUEVN;
BY EVNT_NUM;
* CREATING THE RISK SETS AT EACH UNIQUE EVENT TIME;
IF (SURVTIME GE EVENTIME) THEN DO;
IF (WAITIME LE EVENTIME) THEN DO;
XPLANT = 1;
TRANSAGE = TRANSAGE;
MISMATCH = MISMATCH;
END;
ELSE IF (WAITIME GT EVENTIME) THEN DO;
XPLANT = 0;
TRANSAGE = 0;
MISMATCH = 0;
END;
IF (SURVTIME GT EVENTIME) THEN EVNT_IND = 0.0;
OUTPUT RISKSETS;
END;

PROC PHGLM DATA=RISKSETS PCOV BLOCK;
EVENT EVNT_IND;
MODEL SURVTIME = XPLANT TRANSAGE MISMATCH;
BY DESCENDING EVENTIME;

1417