Extending the Cox Model

Terry M. Therneau

ABSTRACT Since its introduction, the proportional hazards model proposed by Cox has become the workhorse of regression analysis for censored data. In the last several years, the theoretical basis for the model has been solidified by connecting it to the study of counting processes and martingale theory. These developments have, in turn, led to the introduction of several new extensions to the original model. These include the analysis of residuals, time dependent coefficients, multiple/correlated observations, multiple time scales and time dependent strata.

These new techniques have not, however, been easily available to the practictioner, having as of yet not appeared as options in the usual statistical packages. The aim of this monograph is to show how many of these techniques can be approached using standard statistical software, in particular the S-Plus and SAS packages. As such, it should be a bridge between the statistical journals and actual practice.

1 Introduction

Since its introduction, the proportional hazards model proposed by Cox [8] has become the workhorse of regression analysis for censored data. In the last several years, the theoretical basis for the model has been solidified by connecting it to the study of counting processes and martingale theory, as discussed in the books of Fleming and Harrington [13] and of Andersen et. al. [2]. These developments have, in turn, led to the introduction of several new extensions to the original model. These include the analysis of residuals, time dependent coefficients, multiple/correlated observations, multiple time scales,and time dependent strata.

The aim of this monograph is to show how many of these can be approached using standard statistical software, in particular the S-Plus and SAS packages. As such, it should be a bridge between the statistical journals and actual practice. The focus on SAS and S-Plus is based largely on the author's familiarity with these two packages, and should not be taken as evidence against the abilities of other software products. The text uses the labels 'S' and 'S-Plus' interchangeably, the former is the package developed by Bell Laboratories and the latter is the commercial version of the same. Since nearly every installation of "S" consists of the latter this

D. Y. Lin et al. (eds.), *Proceedings of the First Seattle Symposium in Biostatistics*

shorthand notation should cause no harm. All the examples given in the text actually refer to S-Plus.

Sections 2 and 3 lay a foundation for our methods. In section 2, we discuss the *counting process* formulation of a Cox model, the software implementation of this model, and the flexibility that it allows. Section 3 defines a set of residuals for the Cox model, based on the counting process and the mathematical formalism of a martingale.

Sections 4 and 5 use residuals to test the two basic assumptions of a Cox model: proportional hazards and the functional form of the covariates. The counting process formulation allows us to extend these methods to time-dependent covariate models as well.

Section 6 discusses leverage residuals and their use as a basis for robust/resistant estimates of variance. Such estimates are particularly needed for correlated data.

2 The counting process formulation of a Cox model

The Andersen-Gill formulation of the proportional hazards model as a counting process has proven very useful in theoretical development. Represent the ith subject as a counting process where

- $N_i(t)$ is the cumulative number of events to date for the subject.
- $Y_i(t)$ is an indicator function, $Y_i(t) = 1$ iff the subject is at risk and under observation at time t.

From a data analysis viewpoint, each subject is treated as an observation of a (very slow) Poisson process. A censored subject is thought of not as "incomplete data", but as one whose event count is still zero. Time dependent covariates effect the rate for upcoming events, and can depend in any way on past observation of the subject. Intervals of observation need not be contiguous.

To cast a data analysis in this framework has advantages as well. In the computer data set, each subject i is represented by a set of observations: s_{ij}, t_{ij}, δ_{ij}, x_{ij}, k_{ij}, $j = 1, \ldots, n_i$; where $(s_{ij}, t_{ij}]$ is an interval of risk, open on the left and closed on the right, $\delta_{ij} = 1$ if the subject had an event at time t_{ij} and 0 if the subject did not have an event, x_{ij} is the covariate vector over the interval, and k_{ij} is the stratum during the interval. Data sets like this are easy to construct with a package such as SAS or S-Plus.

2.1 Particular cases

Multiple events

The original motivation for adding the counting process form to the S-Plus coxph function was a study of a calcium channel blocker, diltiazem, in post