FITTING COX'S PROPORTIONAL HAZARDS MODEL USING GROUPED SURVIVAL DATA

IAN W. MCKEAGUE AND MEI-JIE ZHANG

Florida State University and Medical College of Wisconsin

Cox's proportional hazard model is often fit to grouped survival data (i.e., occurrence and exposure data over various specified time-intervals and covariate bins), as opposed to continuous data. The practical limits to using such data for inference in the Cox model are investigated. A large sample theory, allowing the bins and time-intervals to shrink as the sample size increase, is developed. It turns out that the usual estimator of the regression parameter is asymptotically biased under optimal rates of convergence. The asymptotic bias is found, and an assessment of the effect on inference is given.

1. Introduction

The purpose of this paper is to study grouped data based inference for Cox's (1972) proportional hazards model. This popular model specifies the conditional hazard function of the survival time of an individual to be $\alpha(t,z) = \lambda_0(t)\exp\{\beta_0 z\}$, where z is a covariate, λ_0 is an unknown baseline hazard function and β_0 is an unknown regression coefficient. (For notational simplicity we assume that the covariate is one-dimensional.) It is important to know whether the convenience of analyzing grouped data from a given actuarial life table is overshadowed by biases that arise when the grouping is coarse. There exist many numerical studies comparing the grouped and continuous Cox model analyses for specific data sets, see the references in Hoem (1987, p. 137). All these studies have found that the two approaches give quite similar results. Breslow (1986), considering data on cancer mortality among Montana smelter workers, found that the estimated regression coefficients from the grouped data analysis were within one standard error of those from the continuous data analysis. This is to be expected when the variation in the baseline hazard λ_0 is moderate over the follow-up period and the covariate effect is mild. Nevertheless, it would be useful to have a theoretical underpinning for these empirical studies.

Theoretical results for *continuous* data are well developed. Corresponding results for grouped data are available only in special cases. The histogram sieve results of Pons and Turckheim (1987) apply to grouped data (when the covariate takes at most finitely many values and is non-time dependent), but the asymptotic bias is not identified. As far as we know, asymptotic bias arising from grouped data under the Cox model has not been studied in the literature.

Our aim here is to obtain the asymptotic bias of the regression coefficient estimator and to indicate how it can be estimated consistently.

2. Fitting the Cox model to grouped data

2.1 The estimator

Let (X, C, Z) be random variables such that the survival time X and the censoring time C are conditionally independent given the covariate Z. The follow-up period and the range of the covariate are taken to be $[0, 1]$. Denote $\delta = I\{X \le C\}$ and $T = X \wedge C$. The ungrouped data consist of n independent replicates (T_i, δ_i, Z_i) of (T, δ, Z).

Let the cells into which the data are grouped be denoted $\mathcal{C}_{rj} = \mathcal{T}_r \times \mathcal{I}_j$, where $\mathcal{T}_1, \ldots, \mathcal{T}_{L_n}$ and $\mathcal{I}_1, \ldots, \mathcal{I}_{J_n}$ are the respective calendar periods (time intervals) and covariate strata. For simplicity, the time intervals are taken to be of equal length $l_n = 1/L_n$ and the covariate strata are taken to have equal width $w_n = 1/J_n$. Grouped data consist of the total number of failures and the total time at risk (exposure) in each cell $\mathcal{C}_{rj}$, given by N_{rj} and Y_{rj}, respectively. In terms of the counting processes $N_i(t) = I\{T_i \le t, \delta_i = 1\}$, and allowing the covariates Z_i to be time dependent,

$$N_{rj} = \sum_i \int_{\mathcal{T}_r} I\{Z_i(t) \in \mathcal{I}_j\}\, dN_i(t) \quad \text{and} \quad Y_{rj} = \sum_i \int_{\mathcal{T}_r} I\{Z_i(t) \in \mathcal{I}_j\} Y_i(t)\, dt,$$

where $Y_i(t) = I\{T_i \ge t\}$.

All our estimators are based on such data.

In the continuous data case the regression coefficient β_0 is estimated by maximizing Cox's partial likelihood function which has logarithm

$$C(\beta) = \sum_i \int_0^1 \beta Z_i(u)\, dN_i(u) - \int_0^1 \log\Big(\sum_i Y_i(u) e^{\beta Z_i(u)}\Big)\, dN^{(n)}(u),$$

where $N^{(n)} = \sum_i N_i$. Pons and Turckheim (1987) estimate β_0 by maximizing a histogram-type Cox's partial likelihood function that has logarithm

$$C_h(\beta) = \sum_r \sum_i \int_{\mathcal{T}_r} \beta Z_i(u)\, dN_i(u) - \sum_r \log\Big(\sum_i \int_{\mathcal{T}_r} e^{\beta Z_i(u)} Y_i(u)\, du\Big) \int_{\mathcal{T}_r} dN^{(n)}(u).$$

In the grouped data case neither $C(\beta)$ nor $C_h(\beta)$ is observable. In fact $C_h(\beta)$ is observable with grouped data only when the covariate process Z takes at most finitely many values and is non-time dependent.

For the general grouped data case we need to consider

$$C_g(\beta) = \sum_{r,j} \beta z_j N_{rj} - \sum_r \log\Big(\sum_j Y_{rj}^{(n)} e^{\beta z_j}\Big) N_r,$$

where $N_r = \sum_{j=1}^{J_n} N_{rj}$ is the number of failures in the rth calendar period, and z_j is the midpoint of the jth covariate stratum. The estimator $\hat{\beta}_g$ is defined as a solution to $U_g(\beta) = 0$, where U_g is the derivative of C_g. This estimator has been studied by Kalbfleisch and Prentice (1973), Holford (1976), Prentice and Gloeckler (1978), Breslow (1986), Hoem (1987), Selmer (1990), and Huet and Kaddour (1994). It can be interpreted as the maximum likelihood estimator in a Poisson regression model, see Laird and Olivier (1981).

2.2 Asymptotic results

As in Andersen and Gill (1982), we denote $S^{(k)}(\beta,t)=\frac{1}{n}\sum_i Z_i^k(t)Y_i(t)e^{\beta Z_i(t)}$ and $s^{(k)}(\beta,t)=ES^{(k)}(\beta,t)$ for $k=0,1,2$, where $0^0=1$. We need the following mild conditions:

(C1) There exists a compact neighborhood $\mathcal{B}$ of β_0 such that, for all t and $\beta\in\mathcal{B}$,

$$s^{(1)}(\beta,t)=\frac{\partial}{\partial\beta}s^{(0)}(\beta,t),\quad s^{(2)}(\beta,t)=\frac{\partial^2}{\partial\beta^2}s^{(0)}(\beta,t).$$

(C2) The functions $s^{(k)}$ are Lipschitz, $s^{(0)}$ is bounded away from zero on $\mathcal{B}\times[0,1]$, and

$$V^{-1}=\int_0^1 v(\beta_0,t)s^{(0)}(\beta_0,t)\lambda_0(t)\,dt$$

is positive, where $v=s^{(2)}/s^{(0)}-(s^{(1)}/s^{(0)})^2$.

Here we state the main results.

Therorem 2.1 (Consistency of $\hat\beta_g$). If $w_n\to 0$ and $l_n\to 0$, then

$$\hat\beta_g\xrightarrow{P}\beta_0.$$

Theorem 2.2 (Asymptotic normality of $\hat\beta_g$). If $l_n\sim w_n\sim n^{-1/4}$, then

$$\sqrt{n}(\hat\beta_g-\beta_0)\xrightarrow{\mathcal{D}}N(\mu,V),$$

where the asymptotic bias

$$\mu=\tfrac{V}{12}\iint e^{\beta_0 z}\{z-\bar z(\beta_0,t)\}\{\dot\lambda_0(t)\dot F'(t,z)+\beta_0\lambda_0(t)F''(t,z)\}\,dtdz,$$

the double integral is over the region covered by the cells used in grouping the data, $\bar z=s^{(1)}/s^{(0)}$ and $F(t,z)=P(T\ge t,Z\le z)$. Here $\dot F,F'$ denote the partial derivatives of F with respect to t and z, respectively. The various derivatives implicit in μ are assumed to exist and to be continuous.

The proofs of these asymptotic results can be found in McKeague and Zhang (1994).

2.3 Estimation of μ

Some elementary calculus shows that

$$\mu=\tfrac{V}{12}\left(\int_0^1\tfrac{1}{2}\{\bar z(\beta_0,t)-\bar z(2\beta_0,t)\}s^{(0)}(2\beta_0,t)\lambda_0^2(dt)+\beta_0\{\psi(1)-\psi(0)-P(\delta=1)\}\right),$$

where

$$\psi(z)=\int_0^1\{z-\bar z(\beta_0,t)\}e^{\beta_0 z}\lambda_0(t)F'(t,z)\,dt.$$

If the variation in the baseline hazard λ_0 is moderate over the follow-up period, then a correction for grouping in the time domain would not be necessary. Use Holford's (1976) grouped data based estimator of λ_0:

$$\hat{\lambda}_0(t) = \frac{\sum_j N_{rj}}{\sum_j Y_{rj} e^{\hat{\beta}_g z_j}} \quad \text{for } t \in \mathcal{T}_r.$$

We recommend inspection of a plot of $\hat{\lambda}_0$ to assess the variation in λ_0 over the follow-up period.

A grouped data based estimator of $s^{(k)}(\beta, t)$ is given by $S_g^{(k)}(\beta, t) = n^{-1} \sum_j z_j^k Y_{rj} e^{\beta z_j}$ at $t \in \mathcal{T}_r$. We may estimate $F'(t, z)$, at $(t, z) \in \mathcal{C}_{rj}$, by $Y_{rj}/(n w_n l_n)$. These estimators can be plugged into μ, replacing each integral by a sum of terms. The last term in μ is consistently estimated by $\int_0^1 S_g^{(0)}(\hat{\beta}_g, t) \hat{\lambda}_0(t)\, dt$. A consistent grouped data based estimator of V^{-1} is given by $-n^{-1} \partial U_g(\hat{\beta}_g)/\partial \beta$.

This leads to a consistent estimator $\hat{\mu}$ of μ.

3. Simulation

We have carried out a Monte Carlo study to evaluate the performance of our method of bias correction.

We used $\beta_0 = 3$ and a linear baseline hazard function $\lambda_0(t) = bt$, with $b = 1, 3$. The covariate was uniformly distributed on $[0, 1]$. The censoring time was independent of both the survival time and the covariate, and exponentially distributed with parameter values 0.35 and 0.70, for $b = 1, 3$ respectively. The follow-up intervals were taken as $[0, 1]$ and $[0, .6]$, respectively. In each case, this gave a censoring rate of about 30%, including about 12% that were still at risk at the end of follow-up. We used equal numbers of time periods and covariate strata. There were 1000 samples in each simulation run.

Table 1 contains the results. We report Monte Carlo estimates of the mean bias correction, the (normalized) mean bias correction, and the (normalized) mean difference between $\hat{\beta}$ and $\hat{\beta}_g$, where $\hat{\beta}$ is the regression parameter estimator based on the continuous data. The normalization used here was the 'standard error' $\sigma/\sqrt{n}$, where $\sigma^2 = E\hat{V}_g$. The corrected estimator is given by $\hat{\beta}_c = \hat{\beta}_g + \hat{\Delta}$, where $\hat{\Delta} = -\hat{\mu}/\sqrt{n}$. We also report observed levels of Wald tests of the null hypothesis that $\beta_0 = 3$, based on $\hat{\beta}_g$, $\hat{\beta}_c$, and $\hat{\beta}$, against the two-sided alternative.

Table 1: Monte Carlo estimates: mean bias correction, mean relative bias correction, and mean relative difference between $\hat{\beta}$ and $\hat{\beta}_g$; observed levels of (nominal 5%) Wald tests of $\beta_0 = 3$ based on $\hat{\beta}_g$, $\hat{\beta}_c$, and $\hat{\beta}$ are labeled P_g, P_c and P_0, respectively.

b	n	L_n, J_n	$E\hat{\Delta}$	$\sqrt{n}E\hat{\Delta}/\sigma$	$\sqrt{n}E(\hat{\beta} - \hat{\beta}_g)/\sigma$	P_g	P_c	P_0
1	100	3, 3	0.210	0.413	0.484	0.082	0.085	0.048
	500	5, 5	0.097	0.436	0.454	0.056	0.057	0.039
	1000	6, 6	0.069	0.445	0.464	0.086	0.070	0.058
3	100	3, 3	0.203	0.397	0.537	0.085	0.067	0.053
	500	5, 5	0.099	0.441	0.466	0.081	0.061	0.055
	1000	6, 6	0.071	0.450	0.497	0.084	0.060	0.050

The simulation results indicate that $\hat{\Delta}$ adequately removes the bias from $\hat{\beta}_g$ (compare the fifth and sixth columns of Table 1). Moreover, it has restored the levels of the hypothesis tests to be much closer to the level of the analogous continuous data tests (compare the last three columns of Table 1). Although the effect of the grouping in this example is modest—less than half a standard error—the bias correction is expected to continue to perform adequately in cases where the bias is more pronounced.

Acknowledgments

This research was supported by Grant 1 RO1 CA54706-03 from the National Cancer Institute, and PO1-CA-40053 from the National Cancer Institute, the National Institute of Allergy and Infectious Diseases and The National Heart, Lung and Blood Institute.

References

Andersen, P. K. and Gill, R. D. (1982), "Cox's regression model for counting processes: a large sample study," *Ann. Statist.*, 10, 1100–1120.

Bickel, P. J. and Wichura, M. J. (1971), "Convergence criteria for multiparameter stochastic processes and some applications," *Ann. Statist.*, 42, 1656–1670.

Breslow, N. E. (1986), "Cohort analysis in epidemiology," in A. C. Atkinson and S. E. Fienberg, eds., *A Celebration of Statistics: the ISI Centenary Volume*, Springer-Verlag, New York, 109–143.

Cox, D. R. (1972), "Regression models and life tables (with discussion)," *J. Roy. Statist. Soc. B*, 34, 187–220.

Hahn, M. G. (1978), "Central limit theorems in $D[0,1]$", *Z. Wahrsch. Verw. Gebiete*, 44, 89–102.

Hoem, J. M. (1987), "Statistical analysis of a multiplicative model and its application to the standardization of vital rates: a review," *Int. Statist. Rev.*, 55, 119–152.

Holford, T. R. (1976), "Life tables with concomitant information," *Biometrics*, 32, 587–597.

Huet, S. and Kaddour, A. (1994), "Maximum likelihood estimation in survival analysis with grouped data on censored individuals and continuous data on failures," *Appl. Statist.*, 43, 325–333.

Kalbfleisch, J. D. and Prentice, R. L. (1973), "Marginal likelihoods based on Cox's regression and life model," *Biometrika*, 60, 267–278.

Laird, N. and Olivier, D. (1981), "Covariance analysis of censored survival data using log-linear analysis techniques," *J. Amer. Statist. Assoc.*, 76, 231–240.

McKeague, I. W. and Zhang, M.J. (1994), "Sheppard's correction for grouping in Cox's Proportional hazards model," Technical Report 5, 1994, Division of Biostatistics, Medical College of Wisconsin.

Pons O. and Turckheim, E. de (1987), "Estimation in Cox's periodic model with a histogram-type estimator for the underlying intensity," *Scand. J. Statist.*, 14, 329–345.

Prentice R. L. and Gloeckler L. A. (1978), "Regression analysis of grouped survival data with application to breast cancer data," *Biometrics*, 34, 57–67.

Selmer, R. (1990), "A comparison of Poisson regression models fitted to multiway summary tables and Cox's survival model using data from a blood pressure screening in the city of Bergen, Norway," *Statistics in Medicine*, 9, 1157–1165.

Department of Statistics
Florida State University
Tallahassee, Florida 32306

Division of Biostatistics
Medical College of Wisconsin
Milwaukee, Wisconsin 53226