FITTING COX'S PROPOR

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 $\delta = I\{X \leq C\}$ 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 $N_i(t) = I\{T_i \leq t, \delta_i = 1\}$, 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 $Y_i(t) = I\{T_i \geq t\}$.

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 tak y v um likelihood estimator in a Poissonhorregression 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

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

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 (197) 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 $S_g^{(k)}(\beta, t) = n^{-1} \sum_j z_j^k Y_{rj} e^{\beta z_j \mathrm{Z}}$ esed0(the)

The simulation results indicate that

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, 1157–11 5.

Department of Statistics
Florida State University
Tallahassee, Florida 32306

Division of Biostatistics
Medical College of Wisconsin
Milwaukee, Wisconsin 53226