Model Diagnostics[1]

STA312 Spring 2019

[1] See last slide for copyright information.

Background Reading

- Chapter 7 in *Applied Survival Analysis Using R* by Dirk Moore
- *Modeling Survival Data: Extending the Cox Model* (2000) by Terry Thereau and Patricia Grambsch

Overview

1. Stochastic processes

2. Residuals

What could go wrong?

- Proportional hazards assumption could be incorrect. The log-normal model is an example.
- Relationships might not be straight-line. For example,

$$h(t) = h_0(t)\ \exp\{\beta_1 \cos(\beta_2 x)\}$$

- Some individual observations may have too much influence on the results.
- Look at residuals.
- *Martingale* residuals?

Stochastic Processes

- A *stochastic process* is an infinite collection of random variables.
- A *counting process* $N(t)$ counts the number of events up to and including time t.
- Let $N_i(t)$ be the number of deaths for patient i, in the interval $(0, t]$
- This means more general counts are possible (and useful).
  - Number of heart attacks.
  - Number of major auto repairs.
  - Number of admissions to hospital.
  - Number of lectures missed.
  - Number of times a sexually transmitted disease was diagnosed (for one person).
- These all are in the category of *recurrent risks*.
- Being at risk is also a stochastic process that can turn on or off.

Stochastic processes formulation for survival analysis

The pair (T_i, δ_i) is replaced by

- $N_i(t)$: Number of observed events in $(0, t]$ for unit i.
- $Y_i(t) = \begin{cases} 1 & \text{if unit } i \text{ is at risk at time } t \\ 0 & \text{otherwise} \end{cases}$.
  This is called the *risk process*.

And the probability distribution is determined by the hazard function

$$h_i(t) = h_0(t) e^{\mathbf{x}_i(t)^\top \boldsymbol{\beta}}$$

Note this is a conditional model, in which $\mathbf{x}_i$ is a fixed function of t.

Martingales

A *discrete-time martingale* is a sequence of random variables $X_1, X_2, \ldots$ that satisfies

- $E(|X_n|) < \infty$
- $E(X_{n+1}|X_1, \ldots, X_n) = X_n$

Examples:

- An unbiased random walk.
- A gambler's current fortune if the game is fair.

Martingale sequence with respect to another sequence

Still discrete time

The sequence $Y_1, Y_2, \ldots$ is a martingale with respect to $X_1, X_2, \ldots$ if

- $E(|Y_n|) < \infty$
- $E(Y_{n+1}|X_1, \ldots, X_n) = Y_n$

Example: Likelihood ratio. Let $L_n = \prod_{i=1}^{n} \frac{g(X_i)}{f(X_i)}$. If $X_1, X_2, \ldots$ are independent with density $f(x)$, then $\{L_1, L_2, \ldots\}$ is a martingale with respect to $\{X_1, X_2, \ldots\}$.

Continuous time martingale

A stochastic process $Y(t)$ is said to be a martingale with respect to the stochastic process $X(t)$ if for all t,

- $E(|Y(t)|) < \infty$
- $E(Y(t)|\{X(\tau) : \tau \leq s\}) = Y(s)$

Example: If $\widehat{S}(t)$ is the Kaplan-Meier estimate, then under mild technical conditions, $\sqrt{D}(\widehat{S}(t) - S(t))$ is a continuous time martingale.

Martingale convergence theorems

There are many versions

Let X_n be a martingale satisfying $\sup_{t>0} E(|X|^p < \infty)$ for some $p > 1$.

Then there exists a random variable X such that

$$P(\lim_{n\to\infty} X_n = X) = 1$$

Martingale Central Limit Theorems

Again there are quite a few versions

Under some technical conditions, sums of (normalized) independent martingales converge to a Brownian motion process $B(t)$, for which

- $B(0) = 0$.
- $E(B(t)) = 0$ for all t.
- Independent increments: $B(t) - B(u)$ is independent of $B(u)$ for any $0 \leq u \leq t$.
- Gaussian process: For any positive integer n and time points $t_1, \ldots, t_n$, the joint distribution of $B(t_1), \ldots, B(t_n)$ is multivariate normal.

Doob-Meyer decomposition Theorem

Any counting process $N_i(t)$ can be decomposed into

$$N(t) = \Lambda(t) + M(t),$$

where $M(t)$ is a martingale and $\Lambda(t)$ is a "predictable" stochastic process.

"Predictable" has an intense mathematical definition, but the idea is that the distribution of $\Lambda_{n+1}(t)$ depends on the distribution of $\Lambda_1(t), \ldots, \Lambda_n(t)$.

Decomposition for the Proportional Hazards Model

Special case of survival (one event) and right censored data

Let $N_i(t) = 1$ if unit i failed in $(0, t]$, and zero otherwise.

$$N_i(t) = H_i(t) + M_i(t),$$

where $H_i(t) = \int_0^y h_i(s)\,ds$ is the cumulative hazard.

Martingale Residuals

Based on $N_i(t) = H_i(t) + M_i(t)$

$$\widehat{M}_i(t) = N_i(t) - \widehat{H}_i(t)$$

Evaluated at t_i, the *estimated* martingale residual is

$$\begin{aligned} \widehat{M}_i(t_i) &= \delta_i - \widehat{H}_i(t) \\ &= \delta_i + e^{\mathbf{x}_i(t)^\top \widehat{\boldsymbol{\beta}}} \log\left(\widehat{S}_0(t_i)\right) \end{aligned}$$

- Martingale residuals are martingales.
- Add to zero.
- Large values need investigation.
- Plots against x variables can reveal the functional form of the dependence of survival time on x.

Schoenfeld residuals

We have already seen

$$\sum_{i=1}^{D}\left(x_{(i)} - \sum_{j\in R_i} x_j \frac{e^{\widehat{\beta} x_j}}{\sum_{k\in R_j} e^{\widehat{\beta} x_k}}\right) = 0$$

- The terms that add to zero are called the Schoenfeld residuals
- There is one set for each explanatory variable.
- Unusually large or small values are worthy of investigatoin.
- They can be approximately standardized, which helps.
- They can be used to form a chi-squared test of H_0 : Proportional hazards. (Thereau and Grambsch, Chapter 6).

Case Deletion Residuals

- Let $\widehat{\boldsymbol{\beta}}_{(i)}$ denote the partial MLE of $\boldsymbol{\beta}$ with case i deleted.
- Calculate $\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}$.
- There will be p differences.
- These are called `dfbeta`.
- They can be standardized.
- The standardized versions are called `dfbetas`.
- They can reveal observations that are overly influential.

Copyright Information