Statistics for Biology and Health
Chapter 7 Semiparametric Proportional Hazards Regression with Fixed Covariates

Qi Guo

July 23,2019

THE UNIVERSITY OF TEXAS AT DALLAS
School of Natural Sciences and Mathematics

Contents

Introduction

Introduction

- If the groups are similar, except for the treatment under study, then, the non-parametric methods in last chapter may be used.
- If the subjects in the groups have some additional characteristics that may affect their outcome, like gender, race, etc, we will consider in more detail the widely used multiplicative hazards model due to Cox (1972), often called the proportional hazards model.
- And we introduce it in notes before, here just briefly talk about it.

Semiparametric model

- Let $h(t|Z)$ be the hazard rate at time t for an individual with risk vector Z. The basic model due to Cox is as follows:

$$h(t|Z) = h_0(t)c(\beta' Z)$$

  where $h_0(t)$ is an arbitrary baseline hazard rate, $\beta = (\beta_1, \cdots, \beta_p)'$ is a parameter vector, and $c(\beta' Z)$ is a known function. This is called the semiparametric model because a parametric form is assumed only for the covariate effect. The baseline hazard rate is treated nonparametrically.
- Commonly the $h(t|Z)$ is positive, so choose a model for $c(\beta' Z)$ is:

$$c(\beta' Z) = \exp(\beta' Z) = \exp\left(\sum_{k=1}^{p} \beta_k Z_k\right)$$

Example

- The cox model is called the proportional hazard proportional model because the ratio of two groups of their hazard rates is a constant, the baseline function is cancel out.
- If Z_1 indicates the treatment effect($Z_1 = 1$ if treatment and $Z_1 = 0$ if placebo) and all other covariates have the same value, then, $h(t|Z)/h(t|Z^*) = exp(\beta_1)$, is the risk of having the event if the individual received the treatment relative to the risk of having the event should the individual have received the placebo.

Coding Covariates

Coding Covariates

- In general regression analyses one may have either quantitative or qualitative independent variables, and in this chapter we talk about the independent variables are known at the start of the study, and they are called fixed time covariates.
- For dichotomous variables(indicator variable), like gender, the obvious way is to code one of the genders as 1, the other as 0, for example, if we code the gender variable as $Z_1 = 1$,if male, 0 if female, the hazard rate for males will be $h(t|Z) = h_0(t)\exp(\beta_1)$, and for females will be $h(t|Z) = h_0(t)\exp(0) = h_0(t)$.
- For more categories, we can find it in notes before.

Partial Likelihoods When Ties Are Present

Introduction

- Commonly the partial likelihood for the proportional hazards regression problem when there are no ties between the event times, when the ties exists, there are alternate partial likelihoods have been provided by a variety of authors when there are ties between event times.
- Let $t_1 < t_2 < \cdots < t_D$ denote the D distinct, ordered, event times. Let d_i be the number of deaths at t_i and D_i the set of all individuals who die at time t_i. Let s_i be the sum of the vectors Z_j over all individuals who die at t_i. That is $s_i = \sum_{j \in D_i} Z_j$, and let R_i be the set of all individuals at risk just prior to t_i.
- There are several suggestions for constructing the partial likelihood when there are ties among the event times.

The partial likelihood when ties are present

1. The first is Breslow, which arises naturally from the profile likelihood construction, and it's expressed as:

$$L_1(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' s_i)}{[\sum_{j \in R_i} \exp(\beta' Z_j)]^{d_i}}$$

   And when there are few ties, this approximation works quite well.

2. Efron suggests a partial likelihood of:

$$L_2(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' s_i)}{\prod_{j=1}^{d_i} [\sum_{k \in R_i} \exp(\beta' Z_k) - \frac{j-1}{d_i} \sum_{k \in D_i} \exp(\beta' Z_k)]}$$

   which is closer to the correct partial likelihood based on a discrete hazard model than Breslow's likelihood. When the number of ties is small, Efron's and Breslow's likelihoods are quite close.

The partial likelihood when ties are present

3. The third partial likelihood due to Cox (1972) is based on a discrete-time, hazard-rate model.

- This likelihood is constructed by assuming a logistic model for the hazard rate, that is, if we let $h(t|Z)$ denote the conditional death probability in the interval $(t, t+1)$ given survival to the start of the interval and if we assume:

$$\frac{h(t|Z)}{1-h(t|Z)} = \frac{h_0(t)}{1-h_0(t)} \exp(\beta' Z)$$

  then, this likelihood is the proper partial likelihood.

Discretizing a Continuous Covariate

Introduction

- Sometimes for continuous covariate, we treat it as a binary covariate by assigning a score of 1 to subjects with large values and 0 otherwise, that's discretizing a continuous covariate.
- In most cases a major problem is determining the value of the cut point between high and low-risk groups.
- Here seek a cut point for the covariate which gives us the largest difference between individuals in the two data-defined groups. That is, for a continuous covariate, X, we seek a binary covariate Z defined by $Z = 1$ if $X \geq C$ and 0 if $X < C$,which makes the outcomes of the groups with $Z = 1$ as different from the group with $Z = 0$ as possible based on some statistic.

Test statistic

- The statistic is the score statistic from the Cox model, and for the procedure we look at all possible cut points; and for each cut point, C_k , we compute the log rank statistic based on the groups defined by X being less than the cut point or greater than the cut point.
- At each event time, t_i ,find the total number of deaths d_i , and the total number at risk, r_i, and also find the total number of deaths and risks with $X \geq C_k$, d_i^+ and r_i^+, and then get the log rank statistic:

$$S_k = \sum_{i=1}^{D} [d_i^+ - d_i \frac{r_i^+}{r_i}]$$

where D is the total number of distinct death times.

Test statistic

- The estimated cut point $\hat{C}$ is the value of C_k which yields the maximum $|S_k|$. At this cut point the Cox regression model is:

$$h(t|X) = h_0(t)\exp(bZ)$$

  where $Z = 1$ if $X \geq \hat{C}$, 0 o.w. The usual tests of $H_0 : b = 0$ can not be used here since we picked the cut point$\hat{C}$, which is most favorable to rejecting H_0.
- First defined the quantity s^2:

$$s^2 = \frac{1}{D-1}\sum_{i=1}^{D}\left\{1 - \sum_{j=1}^{i}\frac{1}{D-j+1}\right\}$$

Test statistic

- The test statistic is then:

$$Q = \frac{max|S_k|}{s\sqrt{D-1}}$$

which under the H_0 has a limiting distribution of the supremum of the absolute value of a Brownian Bridge. For $Q > 1$ the p-value of the test is approximately equal to $2\exp\{-2Q^2\}$.

Estimation of the Survival Function

Introduction

- Once we have obtained estimates of the risk coefficients β from a proportional hazards regression model, we can estimate the survival probability for a new patient with a given set of covariates Z_0.
- First, fit a proportional hazards model to the data and obtain the partial maximum likelihood estimators b and the estimated covariance matrix $\hat{V}(b)$ from the inverse of the information matrix.
- Let $t_1 < t_2 < \cdots < t_D$ denote the distinct death times and d_i be the number of deaths at time t_i, let:

$$W(t_i; b) = \sum_{j \in R(t_i)} \exp\left(\sum_{h=1}^{p} b_h Z_{jh}\right)$$

Estimation of the Survival Function

- The estimator of the cumulative baseline hazard rate is given by:

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{d_i}{W(t_i; b)}$$

  which is a step function with jumps at the observed death times, and it reduces to the NA estimator when there are no covariates present.
- The estimator of the baseline survival function, $S_0(t) = \exp[-H_0(t)]$ is given by:

$$\hat{S}_0(t) = \exp[-\hat{H}_0(t)]$$

- To estimate the survival function for an individual with a covariate vector $Z = Z_0$, use the estimator:

$$\hat{S}(t|Z = Z_0) = \hat{S}_0(t)^{\exp(b'Z_0)}$$

Estimation of the Survival Function

- For fixed t, it has an asymptotic normal distribution with mean $S(t|Z = Z_0)$ and a variance which can be estimated by:

$$\hat{V}[\hat{S}(t|Z = Z_0)] = [\hat{S}(t|Z = Z_0)]^2[Q_1(t) + Q_2(t; Z_0)]$$

Here,

$$Q_1(t) = \sum_{t_i \leq t} \frac{d_i}{W(t_i, b)^2}$$

is an estimator of the variance of $\hat{H}_0(t)$ if b were the true value of β. Here

$$Q_2(t; Z_0) = Q_3(t; Z_0)' \hat{V}(b) Q_3(t; Z_0)$$

Estimation of the Survival Function

- with Q_3 the p-vector whose kth element is defined by

$$Q_3(t, Z_0)_k = \sum_{t_i \leq t} \left[\frac{W^k(t_i; b)}{W(t_i; b)} - Z_{0k}\right]\left[\frac{d_i}{W(t_i, b)}\right], \; k = 1, \cdots, p$$

  where

$$W^k(t_i; b) = \sum_{j \in R(t_i)} Z_{jk} \exp(b' Z_j)$$

- Q_2 reflects the uncertainty in the estimation process added by estimating β. Here, $Q_3(t, Z_0)$ is large when Z_0 is far from the average covariate in the risk set.