Statistics for Biology and Health
Chapter 10 Regression Diagnostics

Qi Guo

July 26,2019

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

Contents

Introduction

Introduction

- Commonly we are interested in examining four aspects of the proportional hazards model.
- First, for a given covariate, see the best functional form to explain the influence of the covariate on survival, adjusting for other covariates. For example, for a given covariate Z, is its influence on survival best modeled by $h_0(t)\exp(\beta Z)$, by $h_0(t)\exp(\beta \log Z)$, or, by a binary covariate defined by 1 if $Z \geq Z_0$; 0 of $Z < Z_0$.
- The second aspect of the model to be checked is the adequacy of the proportional hazards assumption.

Introduction

- The third aspect of the model to be checked is its accuracy for predicting the survival of a given subject.
- The final aspect of the model to be examined is the influence or leverage each subject has on the model fit. This will also give us some information on possible outliers.
- In the usual linear regression setup, it is quite easy to define a residual for the fitted regression model.

Cox-Snell Residuals for Assessing the Fit of a Cox Model

Introduction

- The Cox and Snell residuals can be used to assess the fit of a model based on the Cox proportional hazards model.
- Suppose a Cox model was fitted to data $(T_j, \delta_j, Z_j), j = 1, \cdots, n$, assume that $Z_j = (Z_{j1}, \cdots, Z_{jp})'$ are all fixed-time covariates. Suppose that the proportional hazards model $h(t|Z_j) = h_0(t)\exp(\sum \beta_k Z_{jk})$ has been fitted to the model.
- If the estimates of the β's from the postulated model are $b = (b_1, \cdots, b_p)'$, then, the Cox-Snell residuals are defined as:

$$r_j = \hat{H}_0(T_j)\exp\left(\sum_{k=1}^{p} Z_{jk} b_k\right), j = 1, \cdots, n$$

Introduction

- Here, $\hat{H}_0(t)$ is Breslow's estimator of the baseline hazard rate defined in Chapter 7. If the model is correct and the b's are close to the true values of β then, the r_j's should look like a censored sample from a unit exponential distribution.
- To check it, we compute the NA estimator of the cumulative hazard rate of r_j's. If the unit exponential distribution fits the data, then, this estimator should be approximately equal to the cumulative hazard rate of the unit exponential $H_E(t) = t$.
- Thus, a plot of the estimated cumulative hazard rate of the r_j's,$\hat{H}_r(r_j)$, versus r_j should be a straight line through the origin with a slope of 1.

Determining the Functional Form of a Covariate: Martingale Residuals

Introduction

- Now examine the problem of determining the functional form to be used for a given covariate to best explain its effect on survival through a Cox proportional hazards model, such as $\log Z$, Z^2, or $Z \log Z$, etc.
- The residual we shall use here, called a martingale residual, is a slight modification of the Cox-Snell residual.
- Suppose that for the jth individual in the sample, we have a vector $Z_j(t)$ of possible time-dependent covariates, Let $N_j(t)$ have a value of 1 at time t if this individual has experienced the event of interest, $Y_j(t)$ is the indicator that individual j is under study at a time just prior to time t, b is the vector of regression coefficients and $\hat{H}_0(t)$ the Breslow estimator of the cumulative baseline hazard rate.

Martingale Residuals

- The martingale residual is defined as;

$$\hat{M}_j = N_j(\infty) - \int_0^{\infty} Y_j(t) \exp[b' Z_j(t)] d\hat{H}_0(t),\ j = 1, \cdots, n$$

- When the data is right-censored and all the covariates are fixed at the start of the study, then the martingale residual reduces to

$$\hat{M}_j = \delta_j - \hat{H}_0(T_j) \exp\left(\sum_{k=1}^{p} Z_{jk} b_k\right) = \delta_j - r_j,\ j = 1, \cdots, n.$$

- The residuals have the property $\sum_{j=1}^{n} \hat{M}_j = 0$.

Martingale Residuals

- Suppose that the covariate vector Z is partitioned into a vector Z^*, for which we know the proper functional form of the Cox model, and a single covariate Z_1 for which we are unsure of what functional form of Z_1 to use.
- Assume that Z_1 is independent of Z^*. Let $f(Z_1)$ be the best function of Z_1 to explain its effect on survival.Our optimal Cox model is, then:

$$H(t|Z^*, Z_1) = H_0(t) \exp(\beta^* Z^*) \exp[f(Z_1)]$$

- To find f, we fit a Cox model to the data based on Z^* and compute the martingale residuals, $\hat{M}_j, j = 1, \cdots, n$. These residuals are plotted against the value of Z_1 for the jth observation.
- If the plot is linear, then, no transformation of Z_1 is needed. If there appears to be a threshold, then, a discretized version of the covariate is indicated.

Graphical Checks of the Proportional Hazards Assumption

Graphical Checks of the Proportional Hazards Assumption

- If we check for proportional hazards for a given covariate Z_1 after adjusting for all other relevant covariates in the model, after we write the full covariate vector as $Z = (Z_1, Z_2')'$ where Z_2 is the vector of the remaining $p - 1$ covariates in the model. We assume that there is no term in the model for interaction between Z_1 and any of the remaining covariates.
- Here introduce two approaches. The first series of plots requires that the covariate Z_1 has only K possible values. For a continuous covariate, we stratify the covariate into K disjoint strata, $G_1, G_2, \cdots, G_K$, whereas, for a discrete covariate, we assume that Z_1 takes only the values $1, 2, \cdots, K$.

Graphical Checks of the Proportional Hazards Assumption

- Fit a Cox model stratified on the discrete values of Z_1,and we let $\hat{H}_{g0}(t)$ be the estimated cumulative hazard rate in the gth stratum. If the proportional hazards model holds, then, the baseline cumulative hazard rates in each of the strata should be a constant multiple of each other.
- To check the proportionality assumption one could plot $\ln[\hat{H}_{10}(t)], \cdots, \ln[\hat{H}_{k0}(t)]$ vs t. If the assumption holds, then, these should be approximately parallel and the constant vertical separation between $\ln[\hat{H}_{g0}(t)]$ and $\ln[\hat{H}_{h0}(t)]$ should give a crude estimate of the factor needed to obtain $\ln[\hat{H}_{h0}(t)]$ from $\ln[\hat{H}_{g0}(t)]$.

Graphical Checks of the Proportional Hazards Assumption

- An alternative approach is to plot $\ln[\hat{H}_{g0}(t)] - \ln[\hat{H}_{10}(t)]$ vs t for $g = 2, \cdots, K$. If the proportional hazards model holds, each curve should be roughly constant. This method has the advantage that we are seeking horizontal lines for each curve rather than comparing parallel curves.
- And we still have another graphical method based on $\hat{H}_{g0}(t)$ is the so-called Andersen (1982) plots, more details in book P368-P377.

Deviance Residuals

The Deviance Residuals

- Now we examine a model for outliers, after a final proportional hazards model has been fit to the data.
- The martingale residual $\hat{M}_j$ is a candidate for the desired residual, which give a measure of the difference between the indicator of whether a given individual experiences the event of interest and the expected number of events the individual would have experienced.
- The deviance residual is used to obtain a residual which has a distribution more normally shaped than the martingale residual, which is defined by:

$$D_j = sign[\hat{M}_j]\{-2[\hat{M}_j + \delta_j \log(\delta_j - \hat{M}_j)]\}^{1/2}$$

The Deviance Residuals

- We construct a plot of the deviance residuals D_j vs the risk scores $\sum_{k=1}^{p} b_k Z_{jk}$.
- When there is light to moderate censoring, the D_j should looks like a sample of normally distributed noise.
- When there is heavy censoring, a large collection of points near zero will distort the normal approximation, in either cases, potential outliers will have deviance residuals whose absolute values are too large.

Checking the Influence of Individual Observations

Introduction

- The optimal means of checking the influence of a given observation on the estimation process is to compare the estimate b one obtains by estimating β from all the data to the estimate $b_{(j)}$ obtained from the data with the given observation deleted from the sample.
- If $b - b_{(j)}$ close to zero, jth observation has little influence.
- To compute $b - b_{(j)}$ directly requires fitting $n + 1$ Cox regression models, one with the complete data and n with a single observation eliminated, but it's not feasible in larger problems.
- And we can use the codes in notes before.