Cox's proportional hazards/regression model - model assessment

Rasmus Waagepetersen

October 12, 2022

Topics:

- Plots based on estimated cumulative hazards
- Cox-Snell residuals: overall check of fit
- Martingale residuals: assessment of functional form of covariate
- Deviance residuals: detection of outliers
- Score-process residual: check of proportional hazards for each covariate
- Detection of influential observations.

Why not just proceed as for linear normal models ?

Issues:

- censoring.
- for Cox ph model we do not have a fully specified model - thus we do not know distribution of residuals.

Generally, residual analysis is a bit tricky not only for survival data but for non-normal data in general - residuals tend to look 'ugly' even if the model is correct.

Model with one factor

Suppose we have observations (t_{kj}, δ_{kj}) $k = 1, \ldots, K$ and model for the kth group

$$h_k(t) = h_0(t) \exp(\beta_k)$$

Compute a cumulative hazard estimate $\hat{H}_k$ for each group.

Recall

$$H_k(t) = H_0(t) \exp(\beta_k) \Leftrightarrow \log H_k(t) = \log H_0(t) + \beta_k$$

Various types of plots can be considered

1. $\log \hat{H}_k(t)$'s against t
2. $\log \hat{H}_k$ vs $\log \hat{H}_j$
3. $\hat{H}_k$ vs $\hat{H}_j$ (Andersen plot)
4. $\log \hat{H}_k(t) - \log \hat{H}_1(t)$'s vs t.

Alternatives 2.-4. require a bit of programming since the estimates are not obtained for the same ts.

Stratified Cox process

Suppose we have several covariates and the first is a factor dividing subjects into K groups. Then a stratified Cox model is specified by

$$h_k(t|z_{-1}) = h_{0k}(t)\exp(z_{-1}\beta_{-1})$$

where $h_k(\cdot|z_{-1})$ is the hazard for a subject in the kth group with remaining covariate vector $z_{-1} = (z_2, \ldots, z_p)$. That is, a separate baseline hazard h_{0k} for each group/strata.

If proportional hazards holds for the factor used for stratification then

$$H_{0k}(t) = H_0(t)\exp(\beta_k).$$

So we can make plots similar to those on the previous slide to assess proportional hazards for the factor considered.

If we want to assess PH for a quantitative covariate then we can initially discretize it into a factor variable.

Martingale residuals

Notation alert: regarding residuals, covariate vector for ith time is $z_i = (z_{i1}, \ldots, z_{ip})$.

Martingale residuals:

$$r_i^M = \delta_i - \hat{H}_0(t_i) \exp(z_i \hat{\beta})$$

Martingale residuals very skewed with values in interval $]-\infty, 1]$. Not useful for detecting outliers.

Reason for martingale terminology will be more clear when we later on discuss counting processes and martingales.

May be used for assessing functional form of covariate $z_{\cdot l} = (z_{1l}, \ldots, z_{nl})^{\mathsf{T}}$ by computing r_i^M for model without covariate and plotting r_i^M against the omitted covariate z_{il}, $i = 1, \ldots, n$.

Assume true model is

$$h(t_i) = h_0(t_i)\exp[f(z_{i1})]\exp[z_{i,-1}\beta_{-1}]$$

and we fit Cox PH model without $z_{\cdot 1}$. Then (KM page 362)

$$r_i^M \approx a + bf(z_{i1})$$

for coefficients a and b.

Curve fitted to scatter plot may give indication of possible transformation of covariate.

If points (z_{i1}, r_i^M) scattered around straight line then no need for transformation.

Cox-Snell

Cox-Snell residuals based on results for continuous random variable X with survivor function S and cumulative hazard and H:

$$S(X) \sim \text{Unif}(]0,1[) \quad H(X) \sim \text{Exp}(1).$$

Cox-Snell residual:

$$r_i^C = \hat{H}_0(t_i)\exp(z_i\hat{\beta}) = \delta_i - r_i^M$$

Cox-Snell residuals should look like censored sample of unit-rate exponential random variables which have $H(t) = t$.

This can be checked by considering estimated cumulative hazard for r_i^C.

Cox-Snell residuals may be used for checking overall fit of model - but see reservations in practical notes in KM page 358-359.

Deviance residuals

Deviance residuals are obtained by applying 'symmetrizing' transformation to martingale residuals:

$$r_i^D = \text{sign}(r_i^M)[-2(r_i^M + \delta_i \log(\delta_i - r_i^M))]^{1/2}.$$

These residuals should look (approximately) like a sample of *iid* normal random variables if model correct.

However, if heavy censoring distribution becomes bimodal.

May be useful for spotting outliers.

Schoenfeld residuals and score process

For a time t let R_t denote the random index of the person that dies at t given that persons $R(t)$ are at risk and that a death occurs at time t.

Recall score function $u(\beta)$ for Cox's partial likelihood is a sum of terms (p-dimensional vectors)

$$u_i(\beta) = z_i - \mathbb{E}[z_{R_{t_i}} | H(t_i)]) = z_i - e_i \quad i \in D$$

where $H(t_i)$ is history up to time t_i (determines $R(t_i)$ and that a death occurs at time t_i).

The components of these terms are also known as Schoenfeld residuals (KM page 376).

Assessment of timevarying effects

Suppose that we do not have proportional hazards for the jth covariate in the sense that the true effect of z_j is timevarying:

$$\beta_j(t) = \beta_j + \gamma_j g(t).$$

Let $r_{j,i}^S$ be Schoenfeld residual scaled with the covariance matrix of $\hat{\beta}$. Then the expected value of $r_{j,i}^S$ is approximately equal to $\gamma_j g(t_i)$.

Thus a plot of scaled Schoenfeld residuals versus time may reveal deviations from proportional hazards.

Implemented in the `cox.zph` procedure.

This is not covered in KM. See e.g. book by Collett.

We can define the score *process* (KM page 376) as

$$u(\beta, t) = \sum_{\substack{l \in D: \\ t_l \leq t}} u_l(\beta)$$

By definition $u(\hat{\beta}, t) = 0$ for t greater than the maximal observed death time.

KM suggest to plot score process $u(\hat{\beta}, t)$ against time and compare with 95% boundaries of Brownian bridge process.

Martinussen and Scheike (2006) Dynamic regression models for survival data, suggest to compare with simulations of score process under assumed model.

The score process can also be expressed as

$$u(\beta, t) = \sum_{i=1}^{n} \delta_i (z_i - e_i) - \exp(z_i^{\mathsf{T}} \beta) \sum_{\substack{l \in D: \\ t_l \leq t}} \frac{(z_l - e(l))}{\sum_{k \in R(t_l)} \exp(z_k^{\mathsf{T}} \beta)}$$

(we will see later why, when considering counting processes and martingales).

The score residuals are given by the components of $u(\beta, t_i)$, $i = 1, \ldots, n$ (i.e. in total np residuals).

These are also available from the `residuals` function and can be cumulated to obtain score process.

Influential observations

Do some observations have unusually large influence on estimation of β ?

Let $\hat{\beta}$ and $\hat{\beta}_{-i}$ denote estimates of β based on full data set and data with ith observation omitted. Want to look for i where $\hat{\beta} - \hat{\beta}_{-i}$ is an outlier.

Based on score process residuals it is possible to compute approximation of $\hat{\beta}_{-i}$ - i.e. we do not need to fit Cox model for all datasets obtained by omitting one observation.

The resulting estimates of $\hat{\beta} - \hat{\beta}_{-i}$ are called dfbeta in the `residual` function for `coxph` objects.

Use of formal testing ?

KM note 5 on page 380 advocates use of graphical checks rather than formal tests. This is because we know that any statistical model is just an approximation and thus is bound to be rejected if the sample size is large enough.

Remember the famous quote by Box: 'all models are wrong but some are useful'

Graphical checks may reveal if there are any serious deviations between model and data and possibly also hint to the cause of such deviations.