Lecture ten: Model Checking (I)

1. Assumptions of Proportional Hazards Model

(a) Independent observations

(b) Random/Independent censoring

(c) Proportional hazard

2. Extensions of the Cox Model (Therneau and Grambsch, 2000)

(a) Correlated survival times (multivariate survival times)

(b) Time-dependent covariates

(c) Time-varying coefficients (VCM)

(d) stratification

3. Review of Classic Regression and Residuals

4. Objectives of Model Diagnosis

In chapter 3, the focus was on estimating and testing covariate effects assuming the model was correctly chosen. We are interested in examining four aspects of the proportional hazards model.

(a) Detecting nonlinearity in relationship between the log hazard and the covariates. For a given covariate, what is the best functional form to explain the influence of the covariate on survival, adjusting other covariates? (martingale residuals) Examples: log(x) or a binary covariate based on x - ASSESS statement of PROC PHREG in SAS.

(b) The adequacy of the PH assumption (for both categorical and continuous covariates) - ASSESS statement of PROC PHREG

i. Test: Construct time-dependent covariate or VCM (eg. cox.zph() in Splus).

ii. A graphical check (Schoenfeld residuals) or plots based on estimates of cumulative hazard from a stratified model (categorical covariates).

(c) The accuracy for predicting the survival of a given subject. patients who died either too early or too late comparing to what the fitted model predicts. This will tell us the potential outliers (score residuals).

(d) The influence or leverage each subject has on the model fit (score residuals). This will also give us some influence on possible outliers.

5. Cox-Snell Residuals

Suppose $t_1, \ldots, t_n$ are survival times, r death times and $n - r$ right censored. $x_1, \ldots, x_p$ are covariates. The Cox-Snell residual for the i'th individual is

$$r_{Ci} = exp(\hat{\beta}'\mathbf{x}_i)\hat{H}_0(t_i),$$

which is $\hat{H}_i(t_i)$, the estimate of cumulative hazard, and $\hat{\beta}'\mathbf{x}_i = \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi}$.

(a) useful for checking overall fit of the final model.

(b) If T is r.v. associated with the survival time, then $Y = -logS(T)$ has an exponential distribution with unit mean.

(c) If the model fitted to the observed data is well, then a model-based estimate $\hat{S}_i(t_i)$ will be close to the true value $S_i(t_i)$. Which means $\hat{S}_i(t_i)$ will have properties similar to $S_i(t_i)$.

(d) $S_i(t_i)$ is right censored if t_i is right censored (if $t_i < t_i^*$, then $-log\, S_i(t_i^*) > -log\, S_i(t_i)$). Thus, the residuals is a censored sample from the unit exponential distribution.

(e) If we fit the data (i.e. $(-log\hat{S}_i(t_i)$, $\delta_i)$ by KM method (why KM?), then log cumulative hazard vs $logt$ (what is t in this case?) is a straight line with unit slope and zero intercept.

(f) r_{Ci} is undefined if the largest survival time is an event.

(g) Modified Cox-Snell residuals: Take censoring into account.

   i. lack of memory property of exponential distribution: $P(\xi > t_1 + t_0 | \xi > t_0) = exp(-\lambda(t_1))$

ii. modified Cox-Snell residual:

$$r'_{Ci} = \begin{cases} r_{Ci} & \text{for uncensored obs} \\ r_{Ci} + \Delta & \text{for censored obs} \end{cases} \qquad (1)$$

where $\Delta = 1$ (mean) or $\Delta = log2$ (median).

6. **Martingale Residuals**

The martingale residual for i'th individual is

$$r_{Mi} = \delta_i - r_{Ci}.$$

(a) useful for determining the functional form of a covariate to be included in the PH model.

(b) The modern mathematical definition involves counting process and martingale theory.

(c) the observed number of deaths minus expected number of deaths

(d) take values between $-\infty$ and 1

(e) Martingale residuals are uncorrelated (not precisely, since the residuals must sum to zero).

(f) not symmetrically distributed about 0.

7. **Deviance Residuals**

Deviance residual:

$$r_{Di} = sgn(r_{Mi})[-2\{r_{Mi} + \delta_i log(\delta_i - r_{Mi})\}]^{1/2},$$

(a) A normalized transformation of martigale residual, more symmetrically distributed about 0.

(b) dominated by r_{Mi} if r_{Mi} is in $(-\infty, 0)$ and dominated by the logarithmic term, if r_{Mi} is in $(0, 1)$.

(c) though have an interesting theoretical justification, they have not proven very useful in practice.

8. **Partial residual or Schoenfeld residual**

The Schoenfeld residual is

$$r_{Schji} = \delta_i\{x_{ji} - a_{ji}\}$$

where

$$a_{ji} = \frac{\sum_{l \in R(t_i)} x_{jl} exp(\hat{\beta}' \mathbf{x}_l)}{\sum_{l \in R(t_i)} exp(\hat{\beta}' \mathbf{x}_l)}$$

(a) The Schoenfeld residuals are defined as a matrix with one row per death time and one column per covariate.

(b) r_{Schji} is an estimate of i'th component of the efficient score for the j'th covariate in the model. From (3.5), i.e.

$$logL(\beta) = \sum_{i=1}^{n} \delta_i\{\beta' \mathbf{x}_i - log \sum_{l \in R(t_i)} exp(\beta' \mathbf{x}_l)\},$$

we have

$$\frac{\partial logL(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i\{x_{ji} - \frac{\sum_l x_{jl} exp(\beta' \mathbf{x}_l)}{\sum_l exp(\beta' \mathbf{x}_l)}\}.$$

(c) The Schoenfeld score residuals must sum to 0 (from score equations).

(d) In large samples, $E(r_{Schji}) = 0$ and they are uncorrelated with each other (?)

(e) The **weighted Schoenfeld residuals**, which is defined as

$$\mathbf{r}^*_{Schi} = rvar(\hat{\boldsymbol{\beta}})\mathbf{r}_{Schi},$$

where r is the number of deaths, and $\mathbf{r}_{Schi} = (r_{Sch1i}, ..., r_{Schpi})'$ is Schoenfeld residuals for the ith individual, are useful in assessing time trends or lack of proportionality in one of the coefficients of the model since under the proportional hazards assumption, the Schoenfeld residuals have the sample path of a random walk.

9. Score residual

The score residual is defined, for each subject and each variable (an $n \times p$ matrix), as the sum of the score process over time. The score residuals are a decomposition of the first partial derivative of the log likelihood. The score residual is defined as (page 386, Venables & Ripley, 2nd edition)

$$\begin{aligned} r_{Scoji} = & \quad [x_{ji} - a_{ji}(t_i)]\delta_i - \int_0^{t_i}[x_{ji} - a_{ji}(s)]\hat{h}(s)ds \\ = & \quad [(x_{ji} - a_{ji})]\delta_i + exp(\hat{\beta}'\mathbf{x}_i)\sum_{t_r \le t_i} \delta_r \frac{(a_{jr} - x_{ji})}{\sum_{l \in R(t_r)} exp(\hat{\beta}'\mathbf{x}_l)}, \end{aligned}$$

(a) The score residuals play a role in assessing influential or leverage data points.

(b) They also play an important role in the computation of the robust sandwich variance estimators.

(c) The score residuals must sum to 0 (from score equations).

(d) Intuitively, Schoenfeld and score residuals can be described as following:

At each unique event time, there is a contribution to the log-likelihood, and of course also to the first deriviative of the log-likelihood, or score.

Let U_{jki} be the contribution of subject i at unique event time k from covariate j. The score residuals are one per subject,

$$r_{Scoji} = \sum_k U_{jki}.$$

The Schoenfeld residuals are one per time point

$$r_{Schjk} = \sum_i U_{jki}.$$

10. Example 4.1: Infection in patients on dialysis

```
                    Analysis of Maximum Likelihood Estimates
                Parameter     Standard                                Hazard
Variable DF Estimate          Error     Chi-Square    Pr>ChiSq      Ratio

AGE         1     0.03037         0.02624        1.3400       0.2470      1.031
```

```
  SEX       1   -2.71076      1.09590       6.1184     0.0134     0.066

     I                                       R        R        R        R
     N C                                     E        E        E        E
     F E          N                          S        S        S        S
     E N          R        M                 S        S        S        S
 O   C S  A S     E        A        D        C        C        C        C
 b   T O  G E     S        R        E        H        H        O        O
 s   T R  E X     C        T        V        1        2        1        2

 1   8 1 28 1 -0.32857  0.71996  1.05152  -1.0850 -0.24163  -0.7811 -0.173
 2  15 1 44 2 -0.07847  0.92769  1.84341  14.4930  0.66438  13.4322  0.613
 3  22 1 32 1 -1.43313 -0.21393 -0.20034   3.1291 -0.30646  -0.3223  0.057
 4  24 1 16 2 -0.09385  0.91573  1.76519 -10.2216  0.43406  -9.2144  0.384
 5  30 1 10 1 -1.77362 -0.50602 -0.43943 -16.5882 -0.55037   9.8333  0.129
 6  54 0 42 2 -0.31165 -0.26463 -0.72751   .        .       -3.8255 -0.145
 7 119 1 22 2 -0.26553  0.76453  1.16760 -17.8286 -0.00000 -15.4014 -0.079
 8 141 1 34 2 -0.53856  0.51632  0.64809  -7.6201 -0.00000  -7.0914 -0.113
 9 185 1 60 2 -1.65233 -0.43792 -0.38658  17.0910 -0.00000 -15.8114 -0.250
10 292 1 43 2 -1.42341 -0.21235 -0.19894  10.2390 -0.00000   1.5643 -0.149
11 402 1 30 2 -1.42071 -0.18662 -0.17613   2.8575  0.00000   6.5754 -0.100
12 447 1 31 2 -2.39270 -0.82793 -0.67044   5.5338  0.00000   4.7967 -0.103
13 536 1 17 2   .      -1.19482 -0.90412   0.0000  0.00000  16.2456 -0.067
```

```
options ls=80;
libname fu '../sdata';
data fu.dialysis;
infile '../data/dialysis.dat';
input id infectt censor age sex;
proc phreg;
model infectt*censor(0)=age sex;
output out=fu.dialoutp logsurv=nresc resmart=mart resdev=dev
               ressco = ressco1 ressco2 ressch = ressch1 ressch2
               wtressch = wtrsch1 wtrsch2;
proc print data=fu.dialoutp;
run;
```