Lecture Eight: Cox Propotional Hazards Models (III)

Model building (fitting and checking) is a complex process. The final model does not have to contain every covariates in the dataset (*parsimounious*). Model assumptions must be checked (eg.: PH, linearity, independence, random censoring, etc). Issues like interaction (low order, high order) and multicollinearity among covrariates are also important. There are may be several models that explain the data well.

1. Comparing alternative models

   (a) *null model*

   Do covariates explain variation of the data? which covariates should be in the model?

   (b) comparing *nested models*

   i. The smaller the value of $-2log\,\hat{L}$, the better the model.

   ii. The log-likelihood ratio statistic

   $$-2\{log\,\hat{L}(1) - log\,\hat{L}(2)\}$$

   has an asymtotically $\chi^2(q)$ distribution under $H_0 : \beta_{p+1} = \beta_{p+2} = \ldots = \beta_{p+q} = 0$ (see the two models at page 78).

   iii. Example 3.3. Breast cancer (output from ex31.sas)

```
         Model Fit Statistics


                          Without          With
           Criterion     Covariates     Covariates


           -2 LOG L         173.968        170.096
           AIC              173.968        172.096
           SBC              173.968        173.354
```

   Comparison: likelihood ratio test (p = 0.049) and log-rank test (p = 0.061). The hazard functions for the two groups (staining positive, negative) of women are different.

   iv. Example 3.4: Hypernephroma (kidney cancer)

   The data (Table 3.6): treatment: chemo+immunotherapy with/without nephrectomy.

```
age group: 1 (<60); 2 (60-70); 3 (> 70)
Nephrectomy: Yes or No
survival time: months (after treatment)
censoring: 1 event, 0 censored
```

The SAS program for five models:

```
options ls = 80 nodate;
libname fu ’../../sdata’;
data work;
        set fu.kidney;
        if age = 2 then A2 = 1; else A2 = 0;
        if age = 3 then A3 = 1; else A3 = 0;
        A2N=A2*neph;
        A3N=A3*neph;
proc phreg;
        model survt*censor(0)= A2 A3 / covb rl;
proc phreg;
        model survt*censor(0)= neph / covb rl;
proc phreg;
        model survt*censor(0)= neph A2 A3 / covb rl;
proc phreg;
        model survt*censor(0)= neph A2 A3 A2N A3N / covb rl;
run;
```

Table 1: example 3.4

Terms in model	variables in model	$-2log\,\hat{L}$
null model (1)	none	177.667
α_j (2)	A2, A3	172.172
ν_k (3)	N	170.247
$\alpha_j + \nu_k$ (4)	A2, A3, N	165.508
$\alpha_j + \nu_k + (\alpha\nu)_{jk}$ (5)	A2, A3, N, A2N, A3N	162.497

Comparing model (4) and (5); determining the effects of age and nephrectomy on hazard.

Interpretation of parameters: adjusted and unadjusted; overestimated and underestimated.

2. Model selection strategy

(a) *hierachic principle*

(b) *Akaike's information criterion* (AIC)

When comparing models which may not be nested, several criteria were introdued, such as, AIC, BIC or SBC. Basically, penalty is given to models with more covariates. The AIC is defined as

$$AIC = -2log\,\hat{L} + \alpha q,$$

where q is the number of unknown β-parameters in the model and α is a pre-determined constant.

(c) Forward selection, backward elimination, stepwise selection and score (best subset selection)

In SAS, all three procedures have been implemented in PROC PHREG. You need to specify the α level for entry and stay. There are some drawbacks for those automatic variable selection procedures (see sections at p84-88 and SAS manual).

(d) LASSO - the Least Absolute Shrinkage and Selection Operator (Tibshirani, 1996, 1997) and penalized Cox regression

(e) Example 3.5: Multiple myeloma study

  i. SAS program, see it at the course website.

  ii. Log-likelihood for various models

  iii. Summary

    A. univariate analysis without treatment (Reference 1, Klein et al.: keep it in at the very beginning).

    B. The treatment effect is then included in the model.

    C. Check interaction between the treatment and other covariate.

(f) Example 3.6: Prostatic cancer

  i. SAS program: ex36.sas, ex36a.sas, ex36b.sas (posted at the course website)

  ii. AIC for models fitted to the data

(g) Example 3.8: Testing for non-linearity

  i. SAS program

Table 2: example 3.5

variables in model	$-2log\,\hat{L}$
null model	215.94
AGE	215.817
SEX	215.906
BUN	207.453
CA	215.494
HB	211.068
PC	215.875
BJ	213.890
HB+BUN	202.938
HB+BJ	209.829
BUN+BJ	203.641
BUN+HB+BJ	200.503
HB+BUN+AGE	202.669
HB+BUN+SEX	202.553
HB+BUN+CA	202.937
HB+BUN+PC	202.773

```
options ls=80 nodate;
libname fu ’../../sdata’;
data work;
        set fu.myeloma;
        lbun = log(bun);
        hbsquare = hb*hb;
        if 7 < hb <= 10 then hb2 = 1; else hb2 = 0;
        if 10 < hb <= 13 then hb3 = 1; else hb3 = 0;
        if  hb > 13 then hb4 = 1; else hb4 = 0;
/* check non-linearity of HB */
proc phreg;
        model survt*censor(0)= lbun hb / rl;
proc phreg;
        model survt*censor(0)= lbun hb hbsquare / rl;

/* check non-linearity of HB: an alternative approach */
proc phreg;
```

Table 3: example 3.6

variables in model	$-2log\,\hat{L}$	AIC
null model	36.349	36.349
AGE	36.269	39.269
SHB	36.196	39.196
SIZE	29.042	32.042
INDEX	29.127	32.127
AGE+SHB	36.151	42.151
AGE+SIZE	28.854	34.854
AGE+INDEX	28.760	34.760
SHB+SIZE	29.019	35.019
SHB+INDEX	27.981	33.981
SIZE+INDEX	23.533	29.533
AGE+SHB+SIZE	28.852	37.852
AGE+SHB+INDEX	27.893	36.893
AGE+SIZE+INDEX	23.269	32.269
SHB+SIZE+INDEX	23.508	32.508
AGE+SHB+SIZE+INDEX	23.231	35.231

```
        model survt*censor(0)= lbun hb2 hb3 hb4 / rl;
run;
```

ii. lbun was used in 1st edition. lbun? (see chapter 4)

iii. SAS output The $-2log\,\hat{L}$ for the three models are 208.175, 208.32 and 206.755, respectively.

```
**************** part 1 **************
  Analysis of Maximum Likelihood Estimates

                  Parameter    Standard
Variable  DF      Estimate       Error   Chi-Square  Pr > ChiSq

lbun       1       0.51874     0.30207      2.9490      0.0859
HB         1      -0.12711     0.06092      4.3532      0.0369

*************part 2 *****************
           Analysis of Maximum Likelihood Estimates
```

```
                Parameter    Standard
Variable   DF    Estimate       Error   Chi-Square  Pr > ChiSq

lbun        1     0.49440     0.30812       2.5747      0.1086
HB          1     0.04195     0.45457       0.0085      0.9265
hbsquare    1    -0.00880     0.02344       0.1411      0.7072

*************part 3 *********************
          Analysis of Maximum Likelihood Estimates

                Parameter    Standard
Variable  DF    Estimate       Error   Chi-Square  Pr > ChiSq

lbun       1     0.56176     0.32503       2.9871      0.0839
hb2        1     0.13750     0.52255       0.0692      0.7925
hb3        1    -0.81140     0.50760       2.5552      0.1099
hb4        1    -0.56413     0.59792       0.8902      0.3454
```

3. Interpretation of parameter estimates

   (a) Continuous covariate

   i. The interpretation of coefficient

   $$h_i(t) = e^{\beta x_i} h_0(t)$$

   Thus, $\hat{\beta}$ can be interpreted as the logarithm of a hazart ratio when the value of x is increased by one unit.

   ii. how about the hazard ratio and its standard error of r units?

   (b) Factor

   i. Baseline level (0 coefficient)

   ii. Coefficients for other levels, and hazard ratios:

   iii. Example 3.10: hypernephroma (cont.)

```
******************** SAS program *********************************
options ls = 80 nodate;
libname fu '../../sdata';
data work;
        set fu.kidney;
```

```
if neph = 1;/*patients on whom a nephrectomy has beeen performed*/
        if age = 2 then A2 = 1; else A2 = 0;
        if age = 3 then A3 = 1; else A3 = 0;
proc phreg;
        model survt*censor(0)= A2 A3 / covb rl;
run;
******************** SAS output *******************************
       Analysis of Maximum Likelihood Estimates

              Parameter    Standard
Variable  DF   Estimate       Error    Chi-Square    Pr > ChiSq

  A2      1    -0.06457    0.49843      0.0168        0.8969
  A3      1     1.82448    0.68184      7.1600        0.0075

              Analysis of Maximum Likelihood Estimates

                          Hazard      95% Hazard Ratio
            Variable       Ratio      Confidence Limits

              A2             0.937       0.353       2.490
              A3             6.200       1.629      23.591

           The PHREG Procedure

                          Estimated Covariance Matrix

                  Variable                A2                 A3

                  A2            0.2484315618       0.0832117019
                  A3            0.0832117019       0.4649089423
```

iv. The hazard ratio and its variance between other levels:
Calculation based on the existing output or Recoding.

v. Example 3.11: hypernephroma (cont.)

```
********************** SAS program ****************************
options ls = 80 nodate;
libname fu '../../sdata';
```

```
data work;
        set fu.kidney;
if neph = 1;/*patients on whom a nephrectomy has beeen performed*/
        if age = 1 then A1 = 1; else A1 = 0;
        if age = 3 then A3 = 1; else A3 = 0;
proc phreg;
        model survt*censor(0)= A1 A3 / covb rl;
run;
******************* SAS output *****************************
Analysis of Maximum Likelihood Estimates

                Parameter    Standard
Variable   DF    Estimate       Error   Chi-Square  Pr > ChiSq

 A1          1    0.06457     0.49843      0.0168      0.8969
 A3          1    1.88905     0.73954      6.5247      0.0106

              Analysis of Maximum Likelihood Estimates

                          Hazard      95% Hazard Ratio
            Variable       Ratio      Confidence Limits

            A1             1.067       0.402       2.833
            A3             6.613       1.552      28.177
```

(c) Models with interaction

i. Treatment effect after adjusting other confounding factors.

ii. Example 3.12: prostatic cancer

```
*************** SAS program ************************************
/* Tumor size (SIZE) and Gleason score (INDEX) were important
                      variables (example 3.6) */
options ls=80 nodate;
libname fu '../../sdata';
data work;
        set fu.prostat;
proc phreg;
        model st*censor(0)= treat size index / rl;
proc phreg;
```

```
        model st*censor(0)= treat / rl;
run;
******************** SAS output ***************************
        Analysis of Maximum Likelihood Estimates

                 Parameter    Standard
Variable   DF    Estimate       Error    Chi-Square  Pr > ChiSq

 TREAT      1    -1.11276     1.20312      0.8554      0.3550
 SIZE       1     0.08257     0.04746      3.0273      0.0819
 INDEX      1     0.71022     0.33790      4.4178      0.0356

                              The PHREG Procedure

              Analysis of Maximum Likelihood Estimates

                          Hazard      95% Hazard Ratio
            Variable       Ratio      Confidence Limits

            TREAT          0.329       0.031       3.474
            SIZE           1.086       0.990       1.192
            INDEX          2.034       1.049       3.945
```

The hazard ratio unadjusted for SIZE and INDEX is $exp(-1.978) = 0.138$. How to calculate the hazard ratio of two individuals with different covariates?

iii. Hazard ratio and its variance with interaction term: Example 3.13: hypernephroma

```
*************** SAS program ************************************
options ls = 80 nodate;
libname fu '../../sdata';
data work;
        set fu.kidney;
        if age = 2 then A2 = 1; else A2 = 0;
        if age = 3 then A3 = 1; else A3 = 0;
        A2N=A2*neph;
        A3N=A3*neph;
proc phreg;
```

```
        model survt*censor(0)= neph A2 A3 A2N A3N / covb rl;
run;
****************** SAS output ******************************
          Analysis of Maximum Likelihood Estimates

                Parameter    Standard
Variable  DF     Estimate        Error   Chi-Square  Pr > ChiSq

NEPH       1     -1.94325     0.73052      7.0761      0.0078
A2         1      0.00548     0.83489      0.0000      0.9948
A3         1      0.06513     1.17737      0.0031      0.9559
A2N        1     -0.05114     0.97067      0.0028      0.9580
A3N        1      2.00299     1.34276      2.2251      0.1358

                              The PHREG Procedure

                    Analysis of Maximum Likelihood Estimates

                                Hazard      95% Hazard Ratio
                  Variable       Ratio      Confidence Limits

                  NEPH           0.143       0.034       0.600
                  A2             1.005       0.196       5.165
                  A3             1.067       0.106      10.727
                  A2N            0.950       0.142       6.368
                  A3N            7.411       0.533     103.003
```

Using the estimated covariance matrix to calculate the variance of hazard ratio.

Assignment four: Using Cox model to fit the survival data of patients with multiple myeloma (Table 1.3, page 9). For each possible covariate in Table 1.3, find the final model(s) for this study using forward, backward and stepwise selection procedures by providing appropriate selection criteria. Provide your program and output of your program. Interpret your results and compare the final models selected by the three procedures.