Lecture twelve: Model Checking (III)

Identification of influential observations

In the assessment of model adequacy it is important to check the influence of individual observations. Were hazard function, coefficoients $\hat{\beta}$ and other estimates substantially affected by any of the observations?

1. **Checking for outliers following a Cox regression analysis**

   The optimal means is to refit the model by omitting one observation in the study and see the changes to the estimates.

   (a) similarity to Jackknife, cross evaluation (CV).

   (b) Huge task in terms of time and space when sample size is moderate or large.

   (c) More difficult in survival analysis than other linear models.

   (d) Must use some kind of approximation when sample size is not small.

2. **Influence on a parameter estimate**

   Let (t_i, δ_i, x_i) be the usual triple for the survival times, $i = 1, 2, \ldots, n$. Define

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

   and let $\mathbf{r}_{Si}, i = 1, 2, \ldots, n$ be the ith score residual (a $p \times 1$ vector) whose jth component is

   $$r_{Sji} = \delta_i (x_{ji} - a_{ji}) + exp(\hat{\beta}' \mathbf{x}_i) \sum_{t_r \leq t_i} \delta_r \frac{(a_{jr} - x_{ji})}{\sum_{l \in R(t_r)} exp(\hat{\beta}' \mathbf{x}_l)},$$

   for $j = 1, 2, \ldots, p$.

   (a) It can be shown (ref. Biometrics, vol. 40: 493 - 499, 1984 by Cain and Lange) that an approximation to $\hat{\beta}_j - \hat{\beta}_{j(i)}$, the change in $\hat{\beta}_j$ on omitting the ith observation, is the jth component of the vector (weighted transformation of score residual)

   $$\mathbf{r}'_{Si} \mathbf{V}(\hat{\beta}),$$

where $\mathbf{V}$ is the $p \times p$ variance-covariance matrix of $\hat{\beta}$.

This quantity, which is called a *delta-beta*, will be denoted by $\Delta_i \hat{\beta}_j$, so that $\Delta_i \hat{\beta}_j \approx \hat{\beta}_j - \hat{\beta}_{j(i)}$.

(b) *Standardized delta-beta*: $\Delta_i \hat{\beta}_j$ divided by the standard error of $\hat{\beta}_j$.

(c) Index plots of the *delta-beta's* for each covariate: plots of the *delta-beta's* for each covariate against the rank order of the survival times.

(d) Standard?: Is the change less than one standard error?

(e) Example 4.6: Infection in patients on dialysis (cont.)
SAS output:

```
Obs       DBAGE        DBSEX         LD         LMAX

  1     0.001969     -0.19767     0.03285     0.16101
  2     0.000406      0.54326     0.33875     0.30927
  3    -0.001056      0.07414     0.00463     0.06766
  4    -0.011880      0.59430     0.33785     0.62061
  5     0.004903      0.01386     0.05001     0.10416
  6    -0.000542     -0.11922     0.01938     0.05750
  7    -0.009462      0.12695     0.13570     0.29117
  8    -0.003241     -0.03455     0.02692     0.05397
  9    -0.007271     -0.07335     0.13335     0.12352
 10     0.003233     -0.20226     0.03532     0.19266
 11     0.005979     -0.21584     0.06108     0.26352
 12     0.004800     -0.19394     0.04318     0.22366
 13     0.012162     -0.31568     0.21903     0.46368
```

Figures generated from Splus: Program is available at the course website (dial3.s)

SAS program:

```
options ls=80;
libname fu '../../sdata';
data work;
        set fu.dialysis;
proc phreg;
        model infectt*censor(0)=age sex;
        output out=outp dfbeta=dbage dbsex ld=ld lmax = lmax;
proc rank data=outp out=fu.infdial;
        var infectt;
        ranks strank;
proc print;
        var dbage dbsex ld lmax;
filename gsasfile 'infdial.gsf';
goptions reset=all gunit=pct border ftext=swissb htitle=6 htext=2.5
gaccess=gsasfile ROTATE=LANDSCAPE gsfmode=append device=ps;
proc gplot data=fu.infdial;
        plot dbage*strank;
        plot dbsex*strank;
run;
```

3. **Influence of observations on the set of parameter estimates**

(a) Likelihood displacement (LD):

i. The influence of each observation on the overall fit of the model can be measured by

$$2\{log\, L(\hat{\beta}) - log\, L(\hat{\beta}_{(i)})\}$$

ii. Can we do it without refitting (n times) model? Pettitt and Bin Daud (Applied Statistics, vol. 38: 313-329, 1989) show

$$LD_i \approx \mathbf{r}'_{Si}\mathbf{V}(\hat{\beta})\mathbf{r}_{Si},$$

iii. plots of LD vs rank of survival time: Observations that have relatively large values of the diagnostic are influential.

(b) Eigenvector (l_{max}) associated with the largest eigenvalue:

i. The $n \times n$ symmetric matrix

$$\mathbf{B} = \mathbf{\Theta}'\mathbf{V}(\hat{\beta})\mathbf{\Theta},$$

where $\mathbf{\Theta}'$ is the $n \times p$ matrix whose ith row is $\mathbf{r}'_{Si}$.

ii. Eigenvalues and eigenvectors of a square matrix

iii. The absolute values of the elements of the standardized eigenvector corresponding to the largest eigenvalue of the matrix **B**, is a measure of the sensitivity of the fit of the model to each of the n observations in the data set. denoting the eigenvector by l_{max}.

iv. index plots: Plot of $|l_{max}|$ vs rank order of survival times, plots vs covariates can used to assess inluence of each observation.

v. plot $|l_{max}|$ vs covariates will not have a deterministic pattern if the fitted model is correct.

(c) Example 4.7: infection in patients on dialysis

The observations from patients 2, 4 and 13 affect the form of the hazard function to the greatest extent, the four models are

Omitting patient number 2: $0.031Age_i - 3.530Sex_i$,
Omitting patient number 4: $0.045Age_i - 3.529Sex_i$,
Omitting patient number 13: $0.011Age_i - 2.234Sex_i$,

The model based on full data is

$$0.03Age_i - 2.711Sex_i.$$

Illustration: compare the hazard ratio for two age groups; male vs female.

(d) example 4.8: survival of multiple myeloma patients

Figures generated from Splus: Influencial obs. Patient 32, 38 (BUN). patient 32 had very short survival times and the second largest values of BUN. Patient 38 also had short survival. Patient 13 (l_{max}) is an outlier.

Splus program:

```
ex48.s<-function(){
        tmpdf <- importData("../../sdata/infmye.sas7bdat")
        motif()
        par(mfrow=c(2,3))
        attach(tmpdf)
        plot(strank, dbhb, xlab="Rank of survival time",
                ylab="Delta-beta for HB", xlim=c(0,50),
                ylim=c(-0.02,0.03))
        identify(strank, dbhb)
        plot(strank, dbbun, xlab="Rank of survival time",
                ylab="Delta-beta for BUN", xlim=c(0,50),
                ylim=c(-0.002,0.002))
        identify(strank, dbbun)
    plot(strank, lmax, xlab="Rank of survival time",
        ylab="Absolute value of Lmax", xlim=c(0,50), ylim=c(0,0.4))
        identify(strank, lmax)
    plot(hb, lmax, xlab="Value of HB",
        ylab="Absolute value of Lmax", xlim=c(4,15), ylim=c(0,0.4))
    plot(bun, lmax, xlab="Value of BUN",
        ylab="Absolute value of Lmax", xlim=c(5,176), ylim=c(0,0.4))
        detach()
}
```

SAS program:

```
options ls=80;
libname fu '../../sdata';
data work;
        set fu.myeloma;
proc phreg;
        model survt*censor(0)=hb bun;
            output out=outp dfbeta=dbhb dbbun ld=ld lmax = lmax;
proc rank data=outp out=fu.infmye;
        var survt;
        ranks strank;
filename gsasfile 'infmye.gsf';
goptions reset=all gunit=pct border ftext=swissb htitle=6 htext=2.5
gaccess=gsasfile ROTATE=LANDSCAPE gsfmode=append device=ps;
proc gplot data=fu.infmye;
        plot dbhb*strank;
        plot dbbun*strank;
        plot lmax*strank;
        plot lmax*hb;
        plot lmax*bun;
run;
```

4. **What to do about influential observations?**

    (a) check the origin of influential observations (medical charts?), human error?

    (b) report the analysis with and without the influential values.

    (c) delete those observations if they are out of reasonable range.

Assignment seven: Assume the final Cox model for prostatic cancer patients study (Table1.4, page 10) includes **SIZE, INDEX and TREAT** (see example 3.6 at page 73). Identify the influential observations if any. Write up your comments and interpretation. Provide tables as well as relevant plots to justify your answer.

Figure 1: Example 4.6

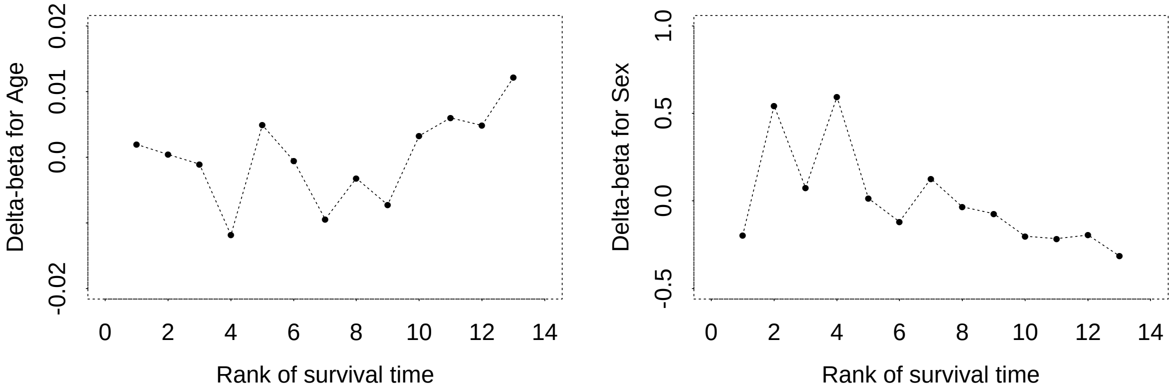
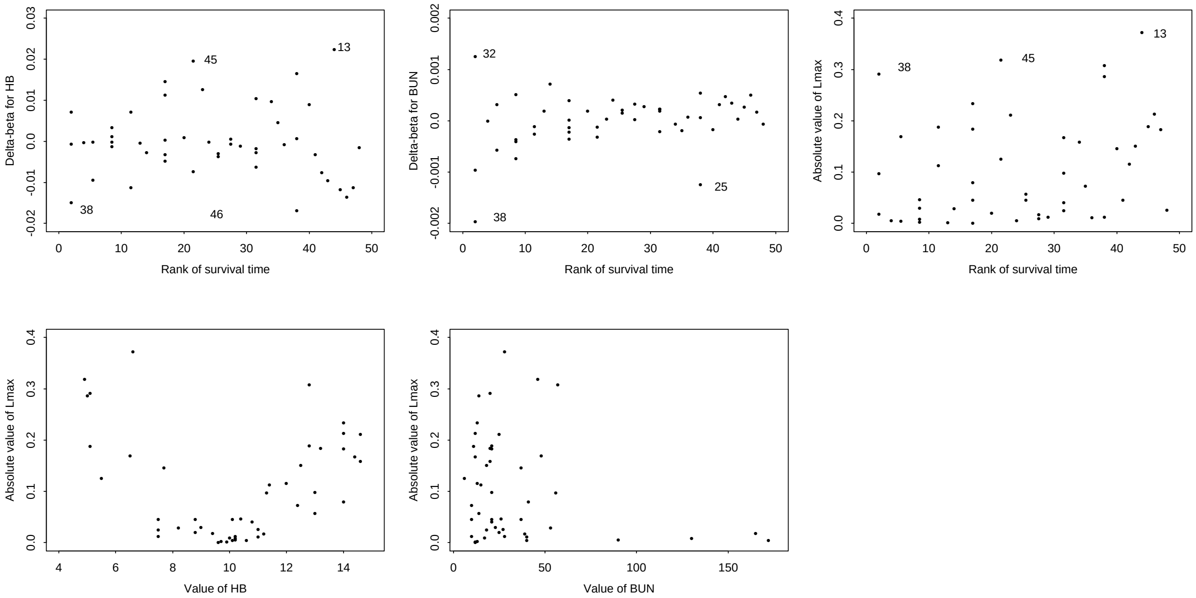


Figure 2: Example 4.8


Delta-beta for HB
Rank of survival time
Delta-beta for BUN
Rank of survival time
Absolute value of Lmax
Rank of survival time
Absolute value of Lmax
Value of HB
Absolute value of Lmax
Value of BUN