Soutrik BANERJEE

SDA homework 3 - 6th June 2006, 12:40 pm

1 (b)

Backward selection - I considered the **full model** with the variables (*sans* interactions) :

age ivdrug cd4lt25 white basehgb proph karnof log_cfu ce cr

The parameter estimates (standard errors) in the **final model** are given below :

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

 cd4lt25      1       0.96445       0.42138        5.2387        0.0221
 karnof       1      -0.05573       0.01039       28.7828        <.0001
 ce           1       1.22298       0.33832       13.0671        0.0003
 cr           1       0.90165       0.33264        7.3473        0.0067

                     Hazard
       Variable      Ratio    Variable Label

       cd4lt25       2.623    Baseline CD4<25 (1=yes,0=no)
       karnof        0.946    Karnofsky Status Score
       ce            3.397    Clarithromycin + Ethambutol
       cr            2.464    Clarithromycin + Rifabutin
```

1 (a)

Forward selection - the results obtained were exactly same as in backward selection, *although SAS uses different methods for backward and forward selections.*

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

 cd4lt25      1       0.96445       0.42138        5.2387        0.0221
 karnof       1      -0.05573       0.01039       28.7828        <.0001
 ce           1       1.22298       0.33832       13.0671        0.0003
 cr           1       0.90165       0.33264        7.3473        0.0067

                     Hazard
       Variable      Ratio    Variable Label

       cd4lt25       2.623    Baseline CD4<25 (1=yes,0=no)
       karnof        0.946    Karnofsky Status Score
       ce            3.397    Clarithromycin + Ethambutol
       cr            2.464    Clarithromycin + Rifabutin
```

1(c)

Stepwise selection - the results obtained were exactly same as in backward (& forward) selection.

Conclusion : in the final model, Karnofsky score and CD4+ count are *protective* in terms of hazard ratios for overall survival as the endpoint. For the variables *ce* and *cr*, the interpretation is more complicated, since both variables simultaneously = 0 means 3 drug regimen. To note here that all patients received treatment with 2 or 3 drugs. Hence, the 3 drug regimen is associated with lower hazard than with each of the 2 drug regimens with the hazard ratios given in the tables above. All the hazard ratios are based on the presence of other cofactors in the final models.

1 (d)

Score option (using best = 3 option) :

Model (variables)	***q***	X^2	***difference in df (from previous model)***	***difference in X^2 (~ difference in -2loglikelihood)***	***AIC***	***P***
log_cfu cd4lt25 white karnof ce cr	6	41.48			566.6	
cd4lt25 white karnof ce cr	5	40.16	1	1.32	565.4	.25
cd4lt25 karnof ce cr	4	38.18	1	1.98	564.1	.16
karnof ce cr	3	32.86	1	5.32	567.1	**.02**

Summary : The model with 4 variables couldn't be reduced to model with 3 variables by both AIC and -2loglikelihood. By this approach, the final model is same as found with selection procedures above.

2 (**Collett's approach to modelling PH model**)

(a) Initial screening with all the univariate variables at slstay = 0.25 resulted in 5 variables being significant, *viz.*, *karnof*, *log_cfu*, *cd4lt25*, *white* and *ce* as shown in the table below.

Z	***Parameter estimates***	***Std. errors***	***P***	***-2loglikelihood***	***AIC***
karnof	-0.04	0.01	<.0001	572.00	575.00
log_cfu	0.12	0.10	.25	589.08	592.08
cd4lt25	0.47	0.40	.24	588.86	591.86
white	0.32	0.25	.21	588.87	591.87
ce	0.50	0.26	0.06	586.74	589.74

(b) I used these 5 variables as main effects in my "full model" to reduce the model by backward selection. Slstay = 0.10 for main effects was chosen. At the end of the procedure, the variables *log_cfu* and *white* were removed.

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

cd4lt25        1        0.96337       0.42250         5.1991         0.0226
karnof         1       -0.05300       0.01035        26.2417         <.0001
ce             1        0.74690       0.26546         7.9164         0.0049
```

```
                 Hazard
Variable         Ratio     Variable Label

cd4lt25          2.621     Baseline CD4<25 (1=yes,0=no)
karnof           0.948     Karnofsky Status Score
ce               2.110     Clarithromycin + Ethambutol
```

AIC = 568.69 ; -2loglikelihood = 559.69 ; q = 3 (α = 3).

(c) Now, by doing forward selection in this model with other non-significant variables (of the univariate procedure) at slentry = 0.10. At the end of the procedure, the variable *cr* was added to the model 2 (b).

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

 cd4lt25       1         0.96445        0.42138         5.2387         0.0221
 karnof        1        -0.05573        0.01039        28.7828         <.0001
 ce            1         1.22298        0.33832        13.0671         0.0003
 cr            1         0.90165        0.33264         7.3473         0.0067

                         Hazard
       Variable          Ratio     Variable Label

       cd4lt25           2.623     Baseline CD4<25 (1=yes,0=no)
       karnof            0.946     Karnofsky Status Score
       ce                3.397     Clarithromycin + Ethambutol
       cr                2.464     Clarithromycin + Rifabutin
```

AIC = 564.07 ; -2loglikelihood = 552.07 ; q = 4 (α = 3).

(d) Now, I enter the stepwise reduction procedure with the model in 2 (c) with slentry and slstay, both, equal to 0.10 along with 5 interactions between cd4lt25 & ce, cd4lt25 & cr, karnof & ce, karnof & cr, cd4lt25 & karnof. The SAS codes are given below (the first 3 main effects are forced into the model, because karnof stays always in the final model) :

```
/*changing 0 in categorical variables to 2*/
data mactrt2;
set mactrt;
if cd4lt25=0 then cd4lt25=2;
else cd4lt25=1;
if ce=0 then ce=2;
else ce=1;
if cr=0 then cr=2;
else cr=1;
run;

/*creating interaction variables*/
data mactrt3;
set mactrt2;
if cd4lt25=2 & ce=2 then cd4ce=0;
if cd4lt25=2 & ce=1 then cd4ce=1;
if cd4lt25=1 & ce=2 then cd4ce=2;
if cd4lt25=1 & ce=1 then cd4ce=3;
if cd4lt25=2 & cr=2 then cd4cr=0;
if cd4lt25=2 & cr=1 then cd4cr=1;
if cd4lt25=1 & cr=2 then cd4cr=2;
if cd4lt25=1 & cr=1 then cd4cr=3;
karn_ce=karnof*ce;
karn_cr=karnof*cr;
cd4_karn=cd4lt25*karnof;
run;

title 'Collett stepwise selection';
proc phreg data=mactrt3;
  model survtime*survstat(0)=cd4lt25 ce cr karnof cd4ce cd4cr karn_ce karn_cr
```

```
cd4_karn / include=3 selection=stepwise slentry=0.1 slstay=0.1;
run;
```

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

 cd4lt25       1     -0.96445       0.42138        5.2387        0.0221
 ce            1     -1.22298       0.33832       13.0671        0.0003
 cr            1     -0.90165       0.33264        7.3473        0.0067
 karnof        1     -0.05573       0.01039       28.7828        <.0001

                        Hazard
          Variable       Ratio    Variable Label

          cd4lt25        0.381    Baseline CD4<25 (1=yes,0=no)
          ce             0.294    Clarithromycin + Ethambutol
          cr             0.406    Clarithromycin + Rifabutin
          karnof         0.946    Karnofsky Status Score
```

AIC = 564.07 ; -2loglikelihood = 552.07 ; q = 4 (α = 3).

Summary :

The hazard ratios are changed from the first model, although the variables are the same at the end of model building by Collett's method. [This is because of the recoding of 0 values in the data step into 2 in order to avoid multiplication by 0 for the interaction variables.] One also finds that no interaction is retained in the final stepwise model by Collett's method. [Since no interaction was significant in the final model, the parameter estimates obtained by fitting the mactrt (original) data can be used in doing further analysis.]

3

Residual analysis :

The fit of the model (by Collett's method) was evaluated by plotting C-S and deviance residuals.

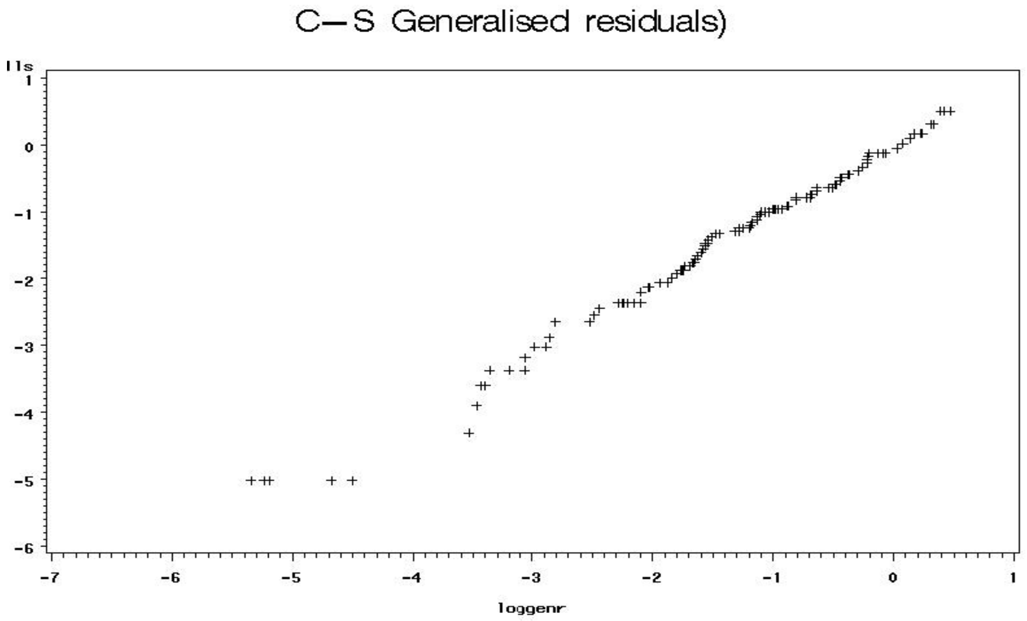

Figure 1 showing Cox-Snell generalised residuals. X-axis shows log(survival), which are log(C-S residuals) or "pseudo/transformed" failure times ; Y-axis shows log(-log(S(t))), which is log(cumulative hazard). This should be a straight line passing through origin with a slope = 1. Apparently, it looks straight, except for the bottom right part.

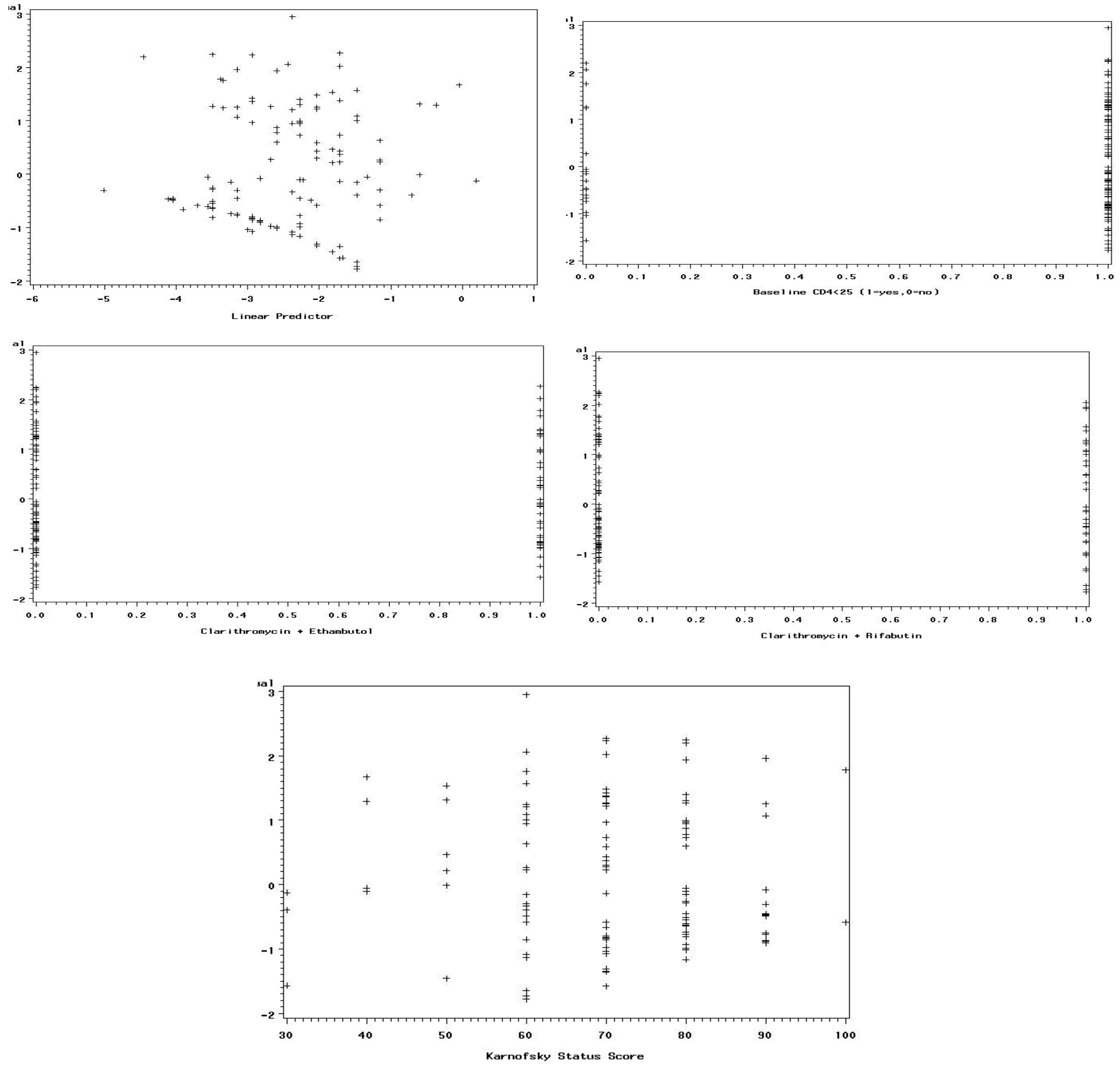

Figure 2 showing deviance residuals against linear predictors (top left) and other covariates. The residuals apparently are distributed symmetrically on both sides of zero. (*The poor quality of images is regretted.*)

Influence diagnostics :

These are obtained by *dfbeta* and *ld* options in SAS. Deletion diagnostics means how much change will occur in the estimated parameter (beta_hat) when the *i*-th person was removed from the sample. The 5 smallest and largest values are provided by SAS (proc univariate). (*The tables are not presented here, but only a brief summary is mentioned*). The subject 140 was most influential in all the 4 covariates in the fitted model. Besides subjects 23 and 135 were also influential in the covariates *ce* and *cr*.

Subject 140 was again found to be most influential to cause change in -2loglikelihood.

4 (a)

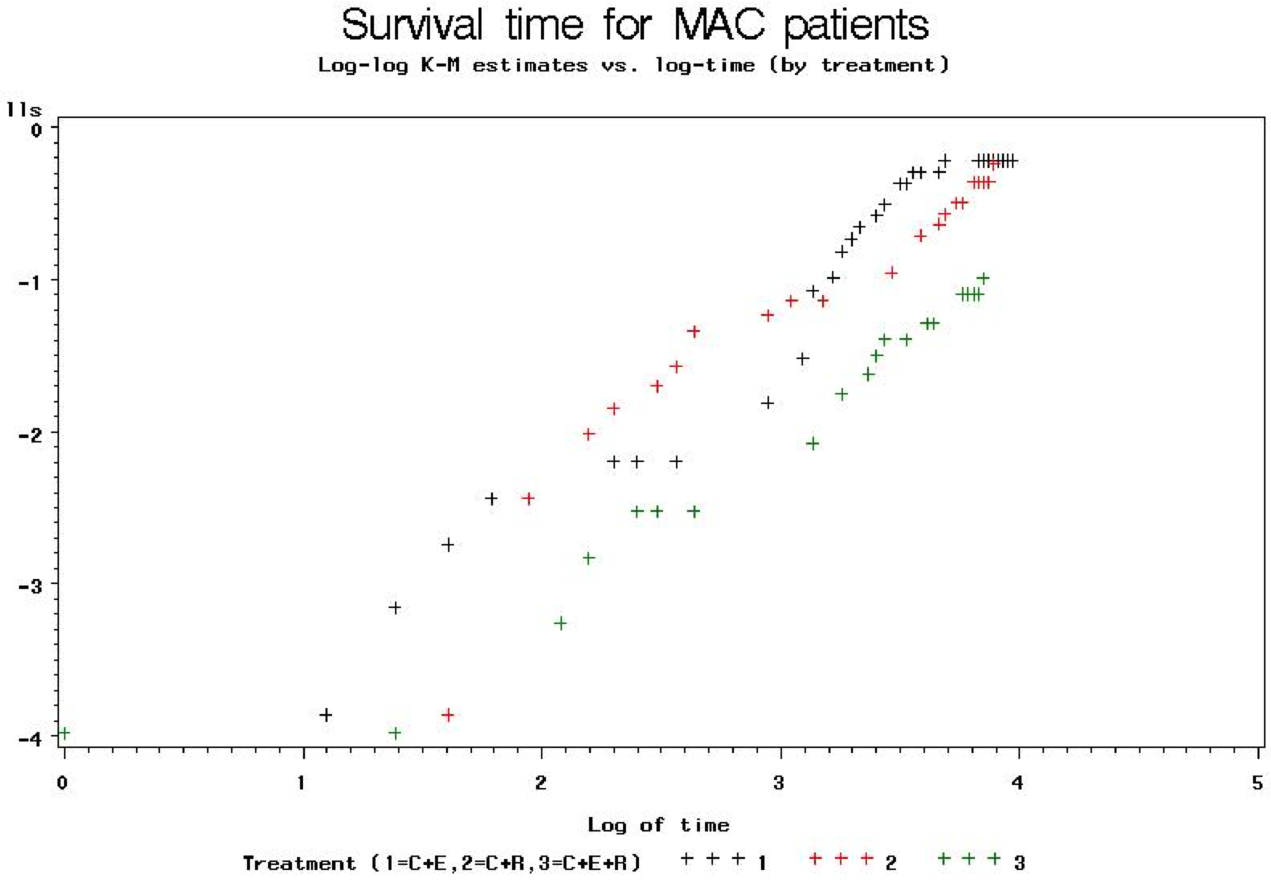

Figure 3 showing log(-log(survival time)) [which is log(cumulative hazard)] vs. log(time) by treatment (3 arms). The curves cross, hence PH model may not hold, but that should be confirmed by a formal statistical test, for example, by "time*covariate" interaction (whether significant or not). This is done in the next exercise.

4 (b)

To test the PH assumption, I included time-varying covariates (already given in the dataset) as *cetime* and *crtime* (*ce* and *cr* varying with time, respectively).

The interaction of treatment and survival time is not significant (table below).

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

 cetime        1      -0.00359        0.00780         0.2120         0.6452
 crtime        1      -0.01090        0.00740         2.1663         0.1411
 cd4lt25       1       0.82203        0.43045         3.6468         0.0562
 karnof        1      -0.04508        0.01020        19.5243         <.0001

                          Hazard
          Variable        Ratio    Variable Label

          cetime          0.996
```

```
crtime          0.989
cd4lt25         2.275    Baseline CD4<25 (1=yes,0=no)
karnof          0.956    Karnofsky Status Score
```

Summary : The hazard ratio is constant (non-significant time-varying covariates) for the 3 treatment arms conditional on the covariates (*cd4lt25* and *karnof*). Therefore, a PH model is plausible in this case.
If we do not control for these covariates, the logrank test (proc lifetest) would be inappropriate for the analysis of the effect of treatment on survival time (as curves cross in Figure 3). One can therefore do a stratified analysis or include treatment by time interactions (extended Cox model).