Applied Survival Analysis - January 2016

Lab 6: Model Selection in Survival Analysis

Today, we are going to understand the Collet's approach for model selection within the context of a proportional hazards model and to assess the overall fit of the model by checking the residuals.

(a) **Collet's Approach for Model Selection:** We are going to work with the MAC dataset (*mac.csv*), focusing on the outcome *dthstat* which equals 1 if a patient died and 0 otherwise. The time to death is *dthtime*, and subjects who did not die are censored at their time of study discontinuation. The covariates of interest for the purpose of this lab are:

agecat sex cd4 karnof ivdrug antiret rif clari

This time we are interested in the time to death, and thus, to specify the survival time along with the failure indicator, you have to use `Surv(dthtime,dthstat)`.

Step 1: Fit univariate models to choose candidate predictors. Use criterion of $p \leq 0.15$ in order to identify predictors, and fill in the following table: Is the effect of treatment significant?

Table 1: Univariate models of interest for the Mac study.

Predictor	Estimate	SE	Pvalue	HR
agecat				
sex				
cd4				
karnof				
ivdrug				
antiret				
rif				
clari				

Step 2: (i) Fit a multivariate model with all significant predictors ($p \leq 0.15$) from Step 1.

(ii) Then use backward selection to eliminate non-significant predictors in a multivariate framework using the AIC criterion.

Step 3: Use forward selection to add any variables not significant at Step 1 to the multivariate model obtained at the end of Step 2. Remember to force the variables that were significant at the end of Step 2 into the model. Also, you should examine the significance of a categorical variable with ≥ 2 levels simultaneously. Are there any other variables added to the model?

Step 4: (i) Do final pruning of the main-effects model using forward stepwise regression.

(ii) Then, create all possible 2-way interaction terms based on the main effects of your last model. Add these to a multivariate model and use a backward stepwise selection procedure to eliminate those considered not significant by the *AIC* criterion. Use the hierarchical principle when considering the significance of the interaction terms of your model.

Step 5: Consider alternate coding of the following covariates:

(i) Use *cd4cat* instead of *cd4*.

(ii) Use *age* instead of *agecat*.

Step 6: Among the models we have fitted so far, which one seems best in terms of the *AIC* criterion?

(b) For the purposes of this lab, let's assess the fit of the model that includes the variables: *age*, *sex*, *cd4*, *karnof* and *antiret*, by checking the residuals:

(i) Use the Cox-Snell residuals to examine the overall fit of the model.

(ii) Get the Martingale residuals and plot them versus the predicted $\log HR_i$ (i.e. $\mathbf{z}_i^T\hat{\boldsymbol{\beta}}$).

(iii) One problem with the martingale residuals is that they tend to be asymmetric. A solution is to use deviance residuals. Get the deviance residuals in `R` and plot them against the predicted $\log HR_i$ and the other covariates of your model.

(iv) Check the proportional hazards assumption by plotting the Weighted Schoenfeld residuals (for each regressor) against time.

(v) **Optional (but important):** Investigate the functional form required for CD4. Based on Collet (page 127), the following algorithm can be used to assess this aspect of model adequacy:

- Fit a Cox PH model without CD4 and obtain the martingale residuals.
- Obtain the CD4 residuals after regressing on the other covariates (i.e, *age*, *sex*, *karnof* and *antiret*), using a linear regression model.

- Plot the martingale residuals against the CD4 residuals, superimposing some smoothed curve to help the interpretation of the plot. Then, this plot should display the correct functional form required for CD4.

Comment on the plot. Are happy with the linearity assumption we have made so far? Is there any transformation of CD4 you would suggest based on this graph?