Homework III
(Due on 04/22/2014 Tuesday by 11:59pm)

Instructions: While discussion with classmates are allowed and encouraged, please try to work on the project independently and direct your questions to me. Please interpret your analysis results using concise and clear language and focusing on interesting findings. Remember to include your R codes in an Appendix. Either an e-copy or a hard copy for submission is acceptable.

PART I Exercises

1. The following table gives a small data set of survival times and a covariate $z$. The ‘+’ sign indicates censoring.

<table>
<thead>
<tr>
<th>ID</th>
<th>Time</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>9+</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>6</td>
</tr>
</tbody>
</table>

(a) Suppose that a Cox proportional hazards (PH) model is used to study the relationship between $z$ and survival. Write down the partial likelihood $L(\beta)$ of $\beta$.

(b) Suppose the last observation is censored instead, i.e., the value is 10+. Does this alter the partial likelihood?

(c) Plot the log partial likelihood $l(\beta)$ of $\beta$ in $[-8,3]$. Does the function look concave?

(d) Find the maximum partial likelihood estimate (MPLE) $\hat{\beta}$ by setting first derivative equal to 0 and solving for $\beta$. Also calculate the second derivative of $l(\beta)$.

(e) Use R to fit the above Cox PH model. How do your results compare to the R output?

2. We showed in class that the score test for comparing two treatments and two sample log rank test are equivalent when there are no ties in the censored survival times. This equivalence is also true for the situation where there are more than two treatments. In this problem, you are asked to prove the equivalence when there are three treatment groups. Namely, suppose we have the following proportional hazards model

$$h(t, z) = h_0(t) \exp (\beta_1 z_1 + \beta_2 z_2),$$
where $z_1$ an $z_2$ are two dummy variables created for 3 treatment groups so that

$$z_1 = \begin{cases} 
1 & \text{if subject is assigned to the 1st treatment group} \\
0 & \text{otherwise} 
\end{cases}$$

$$z_2 = \begin{cases} 
1 & \text{if subject is assigned to the 2nd treatment group} \\
0 & \text{otherwise} 
\end{cases}$$

Given data $\{(T_i, \delta_i, z_{i1}, z_{i2}) : i = 1, 2, \ldots, n\}$, where there are NO tied death times, show that the score test for testing $H_0 : \beta_1 = \beta_2 = 0$ is identical to the logrank test statistic.

**PART II  Computer Project**

Consider the lung cancer data available from

http://www.mayo.edu/research/documents/lunghome/DOC-10027247

The data set contains the following variables:

- Enrolling institution
- Survival time
- Status 1=alive, 2=dead
- Age
- Sex 1=male 2=female
- ECOG performace score, as judged by physician: 0,1,2,3
- Karnofsky performace score, as judged by physician: 100, 90, ..., 30
- Karnofsky performace score, as judged by the patient (self)
- Daily calories consumed at meals
- Weight loss in the last 30 days (negative number = weight gain)

Detailed information about the data set can be found in Loprinzi et al. (1994, *J. Clinical Oncology*).

To bring in the data, you may use

```r
lung.cancer <- read.table(
  file="http://www.mayo.edu/research/documents/lungdat/DOC-10027697",
  sep=" ", header=F, 
  col.names=c("inst", "time", "status", "age", "sex", "ECOG",
               "Karnofsky.physician", "Karnofsky.patient", "calories",
               "weight.loss"))
head(lung.cancer)
```

With this lung cancer data set, perform the analysis by following the steps listed below.
1. Fit a Cox PH model with all covariates included. For ECOG, you may simply treat it as continuous. For missing values, apply listwise deletion.

2. Concerning the full Cox PH model, test to see if we can drop Karnofsky.physician and Karnofsky.patient simultaneously by using Wald test and likelihood ratio test (LRT). Compare the two testing results and comment.

3. Find the ‘best’ Cox model using a variable selection procedure of your choice.

4. Interpret the final model in terms of hazard ratios. In particular, the 95% confidence intervals should be supplied for the hazard ratios.

5. Suppose that we want to test at significance level $\alpha = 0.05$ if the survival functions are the same between males and females, while adjusting for ECOG performance score. Since ECOG is measured on an ordinal scale, the adjustment could be made in three ways by

(a) Treating ECOG as continuous;

(b) Treating ECOG as categorical by coding it with dummy variables;

(c) Using the stratified Cox PH model approach.

Carry out all three approaches and inspect how different the results are in terms of the (adjusted) gender effect.