Semiparametric Survival Analysis in R

In the previous computing handout we discussed nonparametric survival analysis procedures such as the Kaplan-Meier curve and log rank test. In this handout we consider the semiparametric proportional hazards (PH) model. We discuss how to fit survival data to the PH model in the R software package. Survival data are given by \( D = \{(y_i, x_i, \delta_i) : i = 1, 2, 3, \ldots, n\} \), where \( y_i = \min(T_i, C_i) \) is the minimum of the failure time \( T_i \) and censoring time \( C_i \) for subject \( i \), \( x_i \) denotes a collection of covariate information obtained on subject \( i \), and \( \delta_i \) is the event indicator for subject \( i \). Here \( \delta_i = 1 \) if the event was observed and \( \delta_i = 0 \) otherwise. The PH model specifies the hazard function for subject \( i \) as

\[
h_i(t) = e^{\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}} h_0(t).
\]

Inferences focus on HRs comparing individuals with covariate combination \( x^* \) to those with covariate combination \( \tilde{x} \), namely

\[
HR(x^*, \tilde{x}) = \frac{e^{\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}} h_0(t)}{e^{\beta_1 \tilde{x}_{1i} + \beta_2 \tilde{x}_{2i} + \cdots + \beta_k \tilde{x}_{ki}} h_0(t)} = e^{\beta_1 (x^*_1 - \tilde{x}_{1i}) + \beta_2 (x^*_2 - \tilde{x}_{2i}) + \cdots + \beta_k (x^*_k - \tilde{x}_{ki})}.
\]

The R code presented here was used to obtain the results in the handout Survival Analysis Notes 3. Therefore this computing handout is a supplement of that set of notes.

1 The Burn Data

We consider data from a study designed to assess the effect of a new wound cleansing treatment on the time to infection of burn patients. The analyses presented are for illustrative purposes only. The covariates included in the data set that we will consider include (i) treatment (1 = new, 0 = standard), (ii) female (1 = female, 0 = male), (iii) white (1 = white, 0 = other), (iv) surface area burned, and (v) burn type (1 = chemical, 2 = scald, 3 = electric, 4 = flame).
2 PH Model with Binary Exposure for the Burn Data

The PH model using treatment group as the single binary predictor is

\[
h_i(t) = e^{New_i \beta} h_0(t) \text{ for all } t
\]

(1)

where \( New_i = 1 \) if patient \( i \) received the new treatment and \( New_i = 0 \) otherwise. The R code used to fit this model (1) is

```r
PHmodel.1 <- coxph(Surv(time, status) ~ newtrt)
summary(PHmodel.1)
```

The resulting output was

```
Call:
coxph(formula = Surv(time, status) ~ newtrt)

n= 154

coef exp(coef) se(coef)     z      p
newtrt -0.561  0.57   0.293  -1.91 0.056

exp(coef) exp(-coef) lower .95 upper .95
newtrt    0.57  1.75  0.321  1.01

Rsquare= 0.024   (max possible= 0.942)
Likelihood ratio test= 3.73 on 1 df,  p=0.0535
Wald test = 3.66  on 1 df,   p=0.0557
Score (logrank) test = 3.76 on 1 df, p=0.0526
```

3 PH model With Multiple Exposure Levels for the Burn Data

The case of two levels of a categorical covariate such as \( E \) and \( \bar{E} \) can be extended to a categorical covariate with \( L \geq 3 \) levels. A PH model for the
burn data using burn type as the only covariate, where the 4 types of burns considered were chemical, scald, electric, and flame, specifies the hazard for subject $i$ as

$$h_i(t) = e^{\beta_1 s_i + \beta_2 e_i + \beta_3 f_i} h_0(t). \quad (2)$$

Here $s_i = 1$ if subject $i$ suffered a scalding burn and $s_i = 0$ otherwise, $e_i = 1$ if subject $i$ experienced an electrical burn and $e_i = 0$ otherwise, and $f_i = 1$ if subject $i$ was burned by a flame and $f_i = 0$ otherwise. The model can be written as

$$h_i(t) = \begin{cases} 
  e^{\beta_1} h_0(t), & \text{subject } i \text{ experienced a scalding burn} \\
  e^{\beta_2} h_0(t), & \text{subject } i \text{ experienced an electrical burn} \\
  e^{\beta_3} h_0(t), & \text{subject } i \text{ burned by flame} \\
  h_0(t), & \text{subject } i \text{ experienced a chemical burn}
\end{cases}$$

The R code used to fit this model (2) is

```r
PHmodel.2 <- coxph(Surv(time, status) ~ factor(burntype))
summary(PHmodel.2)
```

The resulting output was

Call:
coxph(formula = Surv(time, status) ~ factor(burntype))

```
n= 154

coef exp(coef) se(coef)     z     p
factor(burntype)2  1.26  3.53    1.08  1.166  0.240
factor(burntype)3  2.21  9.12    1.08  2.042  0.041
factor(burntype)4  1.01  2.74    1.02  0.994  0.320

exp(coef) exp(-coef) lower .95 upper .95
factor(burntype)2 3.53 0.284  0.424  29.3
factor(burntype)3 9.12 0.110  1.093  76.1
factor(burntype)4 2.74 0.364  0.375  20.1

Rsquare= 0.047   (max possible= 0.942 )
Likelihood ratio test= 7.44 on 3 df,  p=0.0592
```
Wald test = 8.57 on 3 df, p=0.0356

Score (logrank) test = 9.78 on 3 df, p=0.0206

4 PH Model with a Continuous Covariate for the Burn Data

For purposes of an example, we consider the unadjusted effect of surface area burned on time to infection. The PH model fit is

\[ h_i(t) = e^{SA_i \beta} h_0(t) \]  

where \( SA_i \) denotes the surface area burned for subject \( i \). The R code used to fit this model (3) is

```r
PHmodel.3 <- coxph(Surv(time, status) ~ saburned)
summary(PHmodel.3)
```

The resulting output was

```
Call:
  coxph(formula = Surv(time, status) ~ saburned)

  n= 154

  coef  exp(coef)  se(coef)     z     p
saburned 0.0089     1.01    0.00701 1.27 0.20

  exp(coef) exp(-coef) lower .95 upper .95
saburned    1.01     0.991    0.995    1.02

Rsquare= 0.01   (max possible= 0.942 )
Likelihood ratio test= 1.49 on 1 df,  p=0.222
Wald test = 1.61 on 1 df,  p=0.204
Score (logrank) test = 1.62 on 1 df,  p=0.202
```

4
5 Case 1: Exposure status and continuous covariate

Consider the effect of the new and standard burn treatments adjusted for surface area burned. The PH model fit is

\[ h_i(t) = e^{\beta_1 \text{New}_i + \beta_2 SA_i} h_0(t). \] (4)

The R code used to fit this model (4) is

```r
PHmodel.4 <- coxph(Surv(time, status) ~ newtrt + saburned)
summary(PHmodel.4)
```

The resulting output was

```
Call:
coxph(formula = Surv(time, status) ~ newtrt + saburned)
n= 154

coef exp(coef) se(coef)     z      p
newtrt  -0.52476   0.592 0.29577 -1.77 0.076
saburned  0.00725   1.007 0.00714  1.01 0.310

exp(coef) exp(-coef) lower .95 upper .95
newtrt    0.592       1.690    0.331    1.06
saburned  1.007       0.993    0.993    1.02

Rsquare= 0.03  (max possible= 0.942 )
Likelihood ratio test= 4.7 on 2 df,  p=0.0955
Wald test = 4.72 on 2 df,  p=0.0946
Score (logrank) test = 4.82 on 2 df, p=0.0897
```
6 Case 2: Exposure status and categorical covariate for the burn data

A PH model for the burn data to assess the effect of treatment adjusted for burn type specifies the hazard for subject \( i \) as

\[
h_i(t) = e^{\beta_1 \text{New}_i + \beta_2 s_i + \beta_3 e_i + \beta_4 f_i} h_0(t).
\] (5)

The R code used to fit this model (5) is

```r
PHmodel.5 <- coxph(Surv(time, status) ~ newtrt + factor(burntype))
summary(PHmodel.5)
```

The resulting output was

```
Call:  
coxph(formula = Surv(time, status) ~ newtrt + factor(burntype))

n= 154

 coef exp(coef) se(coef) z p
newtrt -0.596 0.551 0.297 -2.008 0.045
factor(burntype)2 1.133 3.104 1.083 1.046 0.300
factor(burntype)3 2.266 9.641 1.084 2.091 0.037
factor(burntype)4 0.989 2.688 1.016 0.973 0.330

exp(coef) exp(-coef) lower .95 upper .95
newtrt 0.551 1.815 0.308 0.986
factor(burntype)2 3.104 0.322 0.372 25.920
factor(burntype)3 9.641 0.104 1.153 80.635
factor(burntype)4 2.688 0.372 0.367 19.690

Rsquare= 0.072  (max possible= 0.942 )
Likelihood ratio test= 11.5 on 4 df,  p=0.0211
Wald test = 12.6 on 4 df,  p=0.0137
Score (logrank) test = 13.9 on 4 df, p=0.00773
```
7 Case 3: Exposure status and continuous covariate with effect modification for the burn data

The PH model with covariates treatment group, surface area burned, and their interaction is

\[
h_i(t) = e^{\beta_1 New_i + \beta_2 SA_i + \beta_3 (New_i \times SA_i)} h_0(t) = \begin{cases} 
  e^{\beta_1 + (\beta_2 + \beta_3)SA_i} h_0(t), & \text{subject } i \text{ receive new treatment} \\
  e^{\beta_2 SA_i} h_0(t), & \text{subject } i \text{ receive standard treatment} 
\end{cases}
\]

The R code used to fit this model (6) is

```r
PHmodel.6 <- coxph(Surv(time, status) ~ newtrt + saburned + newtrt*saburned)
summary(PHmodel.6)
```

The resulting output was

```
Call:
  coxph(formula = Surv(time, status) ~ newtrt + saburned +
         newtrt*saburned)

  n= 154

  coef  exp(coef)  se(coef)      z     p
newtrt -0.42091   0.656  0.50798 -0.829 0.41
saburned 0.00859   1.009  0.00886  0.970 0.33
newtrt:saburned -0.00367  0.996  0.01468 -0.250 0.80

  exp(coef) exp(-coef) lower .95 upper .95
newtrt   0.656   1.523    0.243    1.78
saburned 1.009   0.991    0.991    1.03
newtrt:saburned 0.996   1.004    0.968    1.03
```

7
Rsquare= 0.03 (max possible= 0.942 )
Likelihood ratio test= 4.76 on 3 df, p=0.190
Wald test = 4.91 on 3 df, p=0.178
Score (logrank) test = 5.1 on 3 df, p=0.164

8 Case 4: Exposure status and categorical covariate with effect modification for the burn data

The PH model with covariates treatment group, sex, and their interaction is

\[
    h_i(t) = e^{\beta_1 \text{New}_i + \beta_2 \text{Female}_i + \beta_3 (\text{New}_i \times \text{Female}_i)} h_0(t)
\]

\[
    h_0(t) =
    \begin{cases} 
    e^{\beta_1 \text{New}_i + \beta_2 \text{Female}_i + \beta_3} h_0(t), & \text{subject } i \text{ female, receive new treatment} \\
    e^{\beta_1} h_0(t), & \text{subject } i \text{ male, receive new treatment} \\
    e^{\beta_2} h_0(t), & \text{subject } i \text{ female, receive standard treatment} \\
    h_0(t), & \text{subject } i \text{ male, receive standard treatment}
    \end{cases}
\]

The R code used to fit this model (7) is

```r
PHmodel.7 <- coxph(Surv(time, status) ~ newtrt + female + newtrt*female)
summary(PHmodel.7)
```

The resulting output was

```
Call:
coxph(formula = Surv(time, status) ~ newtrt + female + newtrt*female)

n= 154
coef exp(coef) se(coef) z  p
newtrt -0.642  0.526  0.321 -2.001 0.045
female -0.736  0.479  0.496 -1.483 0.140
```
newtrt:female  0.219  1.245  0.799  0.274  0.780

exp(coef) exp(-coef) lower .95 upper .95
newtrt  0.526  1.900  0.281  0.987
female  0.479  2.087  0.181  1.267
newtrt:female  1.245  0.803  0.260  5.956

Rsquare= 0.045  (max possible= 0.942 )
Likelihood ratio test= 7.03 on 3 df,  p=0.071
Wald test = 6.88 on 3 df,  p=0.0758
Score (logrank) test = 7.28 on 3 df, p=0.0636

9 Model selection using stepAIC

The stepAIC function can be used for model development. In the following R code the variable surface area burned was removed and no interaction terms were included.

library(MASS)

PHmodelstart <- coxph(Surv(time, status) ~ newtrt + female + white + saburned + factor(burntype))

PHmodel.8 <- stepAIC(PHmodelstart, scope = list(upper = .~. +newtrt*female+newtrt*factor(burntype)+newtrt*saburned, lower = ~newtrt))

summary(PHmodel.8)

The model is given by and the command summary(PHmodel.8) in the above code can be used to extract the fitted model.

\[ h_i(t) = e^{\beta_1 New_i + \beta_2 s_i + \beta_3 e_i + \beta_4 f_i + \beta_5 female_i + \beta_6 white_i} h_0(t). \]
10 Estimated survival curves for burn patients

We plot the estimated time to infection functions for patients that are white, male, and were burned by flame (this type of patient was the most common in the data set), and where one patient received the new treatment and the other received the standard treatment. The estimated survival curves are in Figure 1, which was produced using the following R code.

chemical <- burntype==1
scald <- burntype==2
electric <- burntype==3
flame <- burntype==4

PHmodel.9 <- coxph(Surv(time,status)~newtrt+female+white+scald+electric+flame)

PHmodelx1 <- survfit(PHmodel.9, list(newtrt=1, female=0, white=1, scald=0,electric=0,flame=1))

PHmodelx2 <- survfit(PHmodel.9, list(newtrt=0, female=0, white=1, scald=0,electric=0,flame=1))

plot(PHmodelx1$time, PHmodelx1$surv, type="l", ylim=c(0,1), xlab="Time to Infection", ylab="Surv Prob")

lines(PHmodelx2$time, PHmodelx2$surv, lty=2)

legend(2, 0.3, c("New treatment", "Routine Treatment"), lty=c(1,2), bty='n')
Figure 1: Survival curves for patients that are white, male, and burned by flame where one patient received the new treatment and the other received the standard treatment.

11 Median survival times

In some cases we can get inferences for the median time to event. This, of course, is data dependent. For example, note that the estimated survival curve for the new treatment group does not quite reach $S(m) = 0.5$. As a result, we cannot estimate the median time to infection for that type of person. The estimated median time to infection for the standard treatment patient is 42 days (95% CI: 17, NA). The lower endpoint for the 95% CI for the median time to infection for the new treatment patient is 44 days. These results are obtained from the R code

```R
> PHmodelx1
Call: survfit.coxph(object = PHmodel.8, newdata = list(newtrt = 1, newtrt = 1,
```

```R
> PHmodelx2
which resulted in the the following output
```
female = 0, white = 1, scald = 0, electric = 0, flame = 1))

<table>
<thead>
<tr>
<th>n</th>
<th>events</th>
<th>median</th>
<th>0.95LCL</th>
<th>0.95UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>154</td>
<td>48</td>
<td>42</td>
<td>17</td>
<td>Inf</td>
</tr>
</tbody>
</table>

> PHmodelx2
Call: survfit.coxph(object = PHmodel.8, newdata = list(newtrt = 0, female = 0, white = 1, scald = 0, electric = 0, flame = 1))

<table>
<thead>
<tr>
<th>n</th>
<th>events</th>
<th>median</th>
<th>0.95LCL</th>
<th>0.95UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>154</td>
<td>48</td>
<td>42</td>
<td>17</td>
<td>Inf</td>
</tr>
</tbody>
</table>

12 Testing the Proportional Hazards Assumption

Consider the simplest PH model that includes only the treatment effect:

\[ h_i(t) = e^{New_i \beta} h_0(t) \] for all \( t \).

One way to formally test the proportional hazards assumption involves the introduction of a time-dependent covariate (TDC) into the model. TDCs are simply covariates whose values change over time. To test the PH assumption in the simple model we create a TDC given by

\[ x_i(t) = New_i \times \ln(t). \]

Then the model specifies

\[ h_i(t) = e^{New_i \beta_1 + x_i(t) \beta_2} h_0(t) \] for all \( t \),

To test the PH assumption use the code

```r
PHmodel.1 <- coxph(Surv(time, status) ~ newtrt)
cox.zph(PHmodel.1)
```

The output of `cox.zph(PHmodel.1)` was

<table>
<thead>
<tr>
<th>rho</th>
<th>chisq</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>newtrt</td>
<td>-0.101</td>
<td>0.488</td>
</tr>
</tbody>
</table>
13 Testing the Proportional Hazards Assumption with Multiple Predictors

To test the PH assumption for each predictor in a model we fit separate models where each model includes a time-dependent covariate corresponding to the variable that we want to test. For example, to test whether the PH assumption holds for the new treatment covariate we include into the model a TDC given by \( x_i(t) = \text{New}_i \times \ln(t) \), which results in the model

\[
    h_i(t) = e^{\beta_1 \text{New}_i + \beta_2 s_i + \beta_3 e_i + \beta_4 f_i + \beta_5 \text{female}_i + \beta_6 \text{white}_i + \beta_7 x_i(t)} h_0(t).
\]

and test \( H_0 : \beta_7 = 0 \). The R code to do this is

```r
PHmodel.10 <- coxph(Surv(time,status)~newtrt+female+white+factor(burntype))
cox.zph(PHmodel.10)
```

which resulted in

<table>
<thead>
<tr>
<th></th>
<th>rho</th>
<th>chisq</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>newtrt</td>
<td>-0.1071</td>
<td>0.569</td>
<td>0.4508</td>
</tr>
<tr>
<td>female</td>
<td>0.1663</td>
<td>1.311</td>
<td>0.2522</td>
</tr>
<tr>
<td>white</td>
<td>-0.2595</td>
<td>3.495</td>
<td>0.0616</td>
</tr>
<tr>
<td>factor(burntype)2</td>
<td>-0.2699</td>
<td>3.552</td>
<td>0.0595</td>
</tr>
<tr>
<td>factor(burntype)3</td>
<td>-0.1410</td>
<td>0.907</td>
<td>0.3410</td>
</tr>
<tr>
<td>factor(burntype)4</td>
<td>-0.0918</td>
<td>0.406</td>
<td>0.5238</td>
</tr>
</tbody>
</table>

GLOBAL NA 14.401 0.0255

14 Stratified PH model for the burn data

Given the above results, a stratified PH model where we stratify on race was considered. Thus we consider two strata, namely white (W) and other (Ot).
The stratified PH model specifies

\[
\begin{align*}
  h_{ij}(t) = \begin{cases} 
    e^{\beta_{i1}New_i + \beta_{i2}s_i + \beta_{i3}e_i + \beta_{i4}f_i + \beta_{i5}female_i} h_{0.W}(t), & \text{subject } i \text{ white} \\
    e^{\beta_{i1}New_i + \beta_{i2}s_i + \beta_{i3}e_i + \beta_{i4}f_i + \beta_{i5}female_i} h_{0.Ot}(t), & \text{subject } i \text{ not white}
  \end{cases}
\end{align*}
\] (8)

The R code used to fit this model (8) is

```R
PHmodel.11 <- coxph(Surv(time,status)~newtrt+female+factor(burntype),strata(white))
summary(PHmodel.11)
```

The resulting output was

```
Call:  
  coxph(formula = Surv(time, status) ~ newtrt + female + factor(burntype), 
        data = strata(white))

n= 154

 coef exp(coef) se(coef)     z     p
newtrt -0.644   0.525   0.298 -2.162 0.031
female -0.559   0.572   0.397 -1.409 0.160
factor(burntype)2  1.131   3.099   1.082 1.045 0.300
factor(burntype)3  2.145   8.546   1.087 1.973 0.048
factor(burntype)4  0.983   2.672   1.016 0.967 0.330

exp(coef) exp(-coef) lower .95 upper .95
newtrt 0.525 1.905 0.293 0.942
female 0.572 1.749 0.263 1.244
factor(burntype)2 3.099 0.323 0.372 25.839
factor(burntype)3 8.546 0.117 1.015 71.970
factor(burntype)4 2.672 0.374 0.365 19.582

Rsquare= 0.085  (max possible= 0.942 )
```

14
Likelihood ratio test = 13.7 on 5 df, p=0.0173

Wald test = 14.3 on 5 df, p=0.0139

Score (logrank) test = 15.6 on 5 df, p=0.00801

To test the PH assumption for the stratified model we use

`cox.zph(PHmodel.11)`

which resulted in the following output

<table>
<thead>
<tr>
<th></th>
<th>rho</th>
<th>chisq</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>newtrt</td>
<td>-0.119</td>
<td>0.697</td>
<td>0.404</td>
</tr>
<tr>
<td>female</td>
<td>0.151</td>
<td>1.080</td>
<td>0.299</td>
</tr>
<tr>
<td>factor(burntype)2</td>
<td>-0.251</td>
<td>2.985</td>
<td>0.084</td>
</tr>
<tr>
<td>factor(burntype)3</td>
<td>-0.153</td>
<td>1.054</td>
<td>0.305</td>
</tr>
<tr>
<td>factor(burntype)4</td>
<td>-0.104</td>
<td>0.512</td>
<td>0.474</td>
</tr>
<tr>
<td>GLOBAL</td>
<td>NA</td>
<td>9.492</td>
<td>0.091</td>
</tr>
</tbody>
</table>