Question #1:

Solution 1. *For part I:*

For part (A):

we fit a proportional model considering the interaction between sex and weeks(before or after 15 *weeks). Namely, the model is:*

$$h(t) = h_0(t) \cdot \exp\left(\beta_1 RX + \beta_2 LOGWBC + \beta_3 SEX + \beta_4 SEXT\right)$$

where

$$SEXT = SEX \cdot I(WEEKS > 15)$$

The SAS code is given as:

```
dm 'log; clear; output; clear;';

libname data "C:\akira\data";

/*HW4*/

/*Question 1*/

/*part A*/
/*account for different proportional hazard for age
before and after week 15*/

proc phreg data=data.leukemiab;
    model weeks*relapse(0) = RX LOGWBC SEX SEXT/TIES = exact risklimits;
    SEXT = SEX*(weeks > 15);
    BEFORE15: TEST SEX =0;
    AFTER15: TEST SEX+SEXT = 0;
run;
```

and the output is:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
RX	1	1.52479	0.47819	10.1675	0.0014294	4.594	1.800	11.729
LOGWBC	1	1.70262	0.34874	23.8354	0.0000010	5.488	2.771	10.872
SEX	1	0.34951	0.49052	0.5077	0.4761406	1.418	0.542	3.710
SEXT	1	-0.30928	1.30134	0.0565	0.8121434	0.734	0.057	9.406

Linear Hypotheses Testing Results

	Label	Wald Chi-Square	DF	Pr > ChiSq
	BEFORE15	0.5077	1	0.4761406
	AFTER15	0.0010	1	0.9742076

It appears that the interaction between SEX and WEEKS are **NOT** *significant. Also the test suggets that there are not significant effects from SEX either before or after WEEK* 15.

Controlled for SEX and LOGWBC, the treatment effect RX is significant (p value 0.0014*), and a* 95% *hazard ratio confidence interval is* $(1.800, 11.729)$*, so the* 95% *confidence interval for the treatment coefficient* β_1 *is:* $(\log(1.800), \log(11.729)) = (0.588, 2.462)$

For part (B)*:*

we consider a new variable called 'period' to represent the time interval, up to the event or censoring time for each subject.

We also create a new variable called 'response' as the binary dependent variable.

The model now becomes:

$$\log\Big(-\log(1-h_{ij})\Big) = \alpha_j + \beta_1 RX + \beta_2 LOGWBC + \beta_3 SEX + \beta_4 SEXT + \beta_5 PERIOD$$

Here h_{ij} *is the discrete-time hazard rate for subject* i *at time interval* j*.*

The period is modeled as a linear effect instead of categorical(unrestricted effect) to deal with issues of empty time intervals. We also modeld the interaction between SEX and the WEEKS of before and after 15*.*

The SAS code is:

```
/*part B*/
/*fitting a complementary log-log proportional hazard model*/
/*fitting a linear effect of time, since we have empty intervals
by listing the variable on the model statement alone, instead of using
a class statement*/
/*also consider a time dependent variable for sex before and after 15 weeks*/
data leukemiab_extend;
    set data.leukemiab;
    do period = 1 to weeks;
        IF period = weeks and relapse = 1 then response = 1;
        else response = 0;
        sext = sex*(period > 15);
        output;
    end;
run;

proc logistic data=leukemiab_extend;
    model response(event = '1') = RX LOGWBC SEX SEXT period/link = cloglog CLPARM=BOTH;
run;
```

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.9929	1.2291	53.5342	<.0000001
RX	1	1.4379	0.4681	9.4359	0.0021278
LOGWBC	1	1.4765	0.3007	24.1115	0.0000009
SEX	1	0.0926	0.4557	0.0413	0.8389851
sext	1	-0.5964	0.9367	0.4054	0.5243130
period	1	0.1245	0.0406	9.4077	0.0021608

The confidence interval for treatment effect is given by both profile likelihood and wald tests:

Parameter Estimates and Profile-Likelihood Confidence Intervals			
Parameter	Estimate	95% Confidence Limits	
Intercept	-8.9929	-11.6075	-6.7045
RX	1.4379	0.5635	2.4339
LOGWBC	1.4765	0.8965	2.1034
SEX	0.0926	-0.8325	0.9846
sext	-0.5964	-2.4681	1.2563
period	0.1245	0.0432	0.2031

Parameter Estimates and Wald Confidence Intervals			
Parameter	Estimate	95% Confidence Limits	
Intercept	-8.9929	-11.4019	-6.5839
RX	1.4379	0.5205	2.3554
LOGWBC	1.4765	0.8871	2.0658
SEX	0.0926	-0.8006	0.9858
sext	-0.5964	-2.4324	1.2395
period	0.1245	0.0449	0.2040

In the complementary log-log model, the treatment effect is also significant(p value 0.002*).*

For part (C):

*Compare the proportional hazard model with the complementary log-log model, the parameter estimates for treatment effect RX(*1.52 *and* 1.44*) and LOGWBC(*1.70 *and* 1.48*) are similar. This is because the regression coeffecients in the log-log model are identical to the coefficients in the underlying proportional hazards model.*

The estimates for sex(0.3495 *and* 0.0926*) and sext(*-0.3093 *and* -0.5964*) are different because in the complementary log-log model, we involved the time interval as a covariate (linear effect here), and also our interaction between sex and the time interval is considered. This is differently handled in the proportional hazard model (where the interaction is only between sex and the event time).*

For part II:

For part (A)*:*

I confirm that I have read page $\#236 - 240$ *of Allison's book and understood the analysis of the job duration data using the logit model.*

To compare later, we print out the output of the logit model here:

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ed	1	6.8392	0.0089
prestige	1	46.5794	<.0001
salary	1	6.6791	0.0098
year	4	23.2529	0.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	3.4443	1.1821	8.4896	0.0036
ed		1	0.2249	0.0860	6.8392	0.0089
prestige		1	-0.1235	0.0181	46.5794	<.0001
salary		1	-0.0268	0.0104	6.6791	0.0098
year	1	1	-2.6875	0.8327	10.4174	0.0012
year	2	1	-1.4475	0.7671	3.5605	0.0592
year	3	1	-0.0130	0.7272	0.0003	0.9857
year	4	1	0.2355	0.7779	0.0916	0.7621
year	5	0	0	.	.	.

	AIC	**SC**
Unrestricted	215.67	244.21
Linear	216.28	234.31
Quadratic	212.19	233.82
Logarithmic	211.98	230.00

For part (B)*:*

we just need to change the link function in order to fit the complementary log-log model. We first use the class statement for the unrestricted model. then consider separately the linear, quadratic and logarithm effect of the year in the model.

```
/*Question 1 part II*/
/*part B*/
/*fit the complementary log-log model to the job duration data*/
data jobyrs;
    set data.jobdur;
    do year = 1 to dur;
        IF year = dur and event = 1 then response = 1;
        else response = 0;
        log_year = log(year);
        output;
    end;
run;

proc logistic data=jobyrs;
    class year/param = glm;
    model response(event = '1') = ed prestige salary year/link = cloglog;
run;

proc logistic data=jobyrs;
    model response(event = '1') = ed prestige salary year/link = cloglog;
run;

proc logistic data=jobyrs;
    model response(event = '1') = ed prestige salary year year*year/link = cloglog;
run;

proc logistic data=jobyrs;
    model response(event = '1') = ed prestige salary log_year/link = cloglog;
run;
```

The output of the unrestricted model is the following:

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ed	1	6.8470	0.0088789
prestige	1	53.8990	<.0000001
salary	1	6.5623	0.0104163
year	4	26.4563	0.0000256

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	2.1952	0.8868	6.1270	0.0133136
ed		1	0.1655	0.0632	6.8470	0.0088789
prestige		1	-0.0926	0.0126	53.8990	<.0000001
salary		1	-0.0229	0.00895	6.5623	0.0104163
year	1	1	-2.0952	0.6600	10.0777	0.0015007
year	2	1	-1.1097	0.6156	3.2489	0.0714700
year	3	1	0.0489	0.5840	0.0070	0.9333091
year	4	1	0.2131	0.6287	0.1149	0.7346025
year	5	0	0	.	.	.

We can see that it is pretty similar to the results of the logit model, and we also get a significant

effect of the year from the Wald chi-square test.

We can also compare the AIC and SC of all four models:

	AIC	*SC*
Unrestricted	216.209	245.055
Linear	218.111	236.140
Quadratic	213.273	234.908
Logarithmic	213.310	231.339

The model with quadratic effect of year gives the smallest AIC value, but very close to the model with logarithm of year . The model with logarithm of year gives the smallest SC value, which is consistent with the logit model comparison.

So both logit and complementary log-log models suggest that the model with logarithmic of year is preferrable.

Question #2.

Solution 2. *For part I:*

For part (A):

If we compute by hand, then we get:

$$\begin{aligned} \#events &= \frac{c(\alpha, P)}{\pi_1(1-\pi_1)\beta_*^2} \\ &= \frac{c(0.05, 0.9)}{0.5 \cdot 0.5 \cdot \Big[\log(1.38)\Big]^2} = \frac{10.51}{0.25 \cdot \Big[\log(1.38)\Big]^2} \\ &\simeq 405.25 \end{aligned}$$

So we need 406 *events.*

If we assume as in part B, that the accrual time is three years and the follow up is one year, then we can do it with SAS as well:

```
/*Question 2*/
/*part I*/

/*part A*/
/*we assume also 3 years accrual time
and 1 year follow up, then it is doable with SAS*/
proc power;
    twosamplesurvival test = logrank
    curve("control") = (1):(0.93)
    refsurvival = "control"
    hazardratio = 1.38
    accrualtime = 3
    followuptime = 1
    eventstotal = .
    power= 0.9
    ;
run;
```

Computed Ceiling Event Total		
Fractional Event Total	Actual Power	Ceiling Event Total
397.182275	0.901	398

So we need 398 *total events, which is slightly smaller than the hand calculation, but with extra information of accrual time and follow up time.*

For part (B)*:*

We just simply replace the SAS line above 'eventstotal' with 'ntotal', we get:

```
/*part B*/
proc power;
    twosamplesurvival test = logrank
    curve("control") = (1):(0.93)
    refsurvival = "control"
    hazardratio = 1.38
    accrualtime = 3
    followuptime = 1
    ntotal = .
    power= 0.9
    ;
run;
```

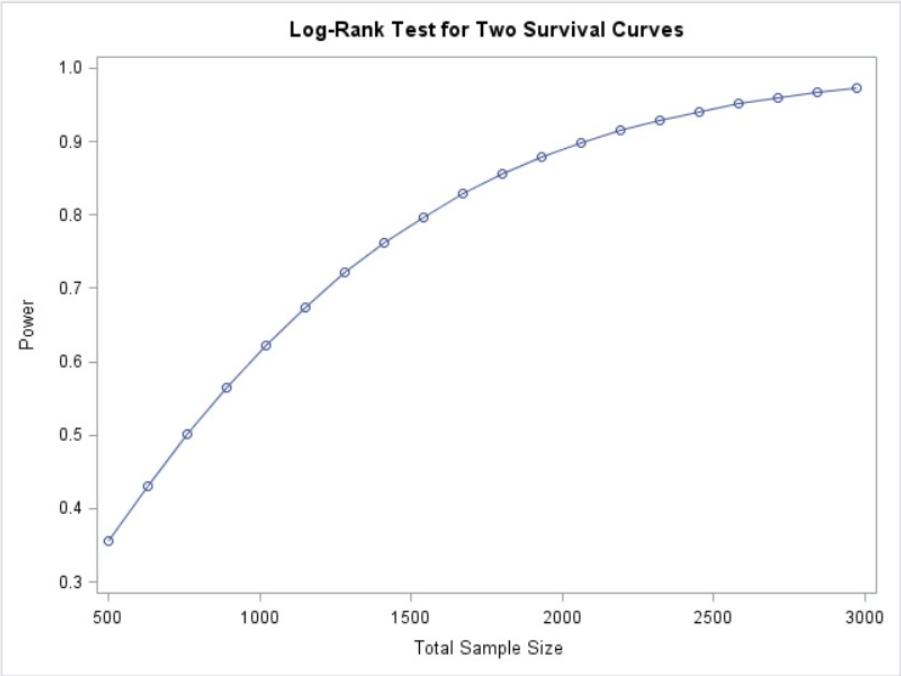
Computed N Total	
Actual Power	N Total
0.900	2076

So we need a total sample size of 2076*, namely* 1038 *in each of the control and treatmet group.*

For part (C)*:*

We slightly modify the above code:

```
/*part C*/
proc power;
    twosamplesurvival test = logrank
    curve("control") = (1):(0.93)
    refsurvival = "control"
    hazardratio = 1.38
    accrualtime = 3
    followuptime = 1
    power= .
    ntotal = 2000;
    plot x = n min =500 max = 3000;
run;
```



For part (D):

To compute the number of events by hand, we have:

$$\begin{aligned} \#events &= \frac{c(\alpha, P)}{\pi_1(1-\pi_1)\beta_*^2} \\ &= \frac{c(0.05, 0.9)}{0.25 \cdot 0.75 \cdot \Big[\log(1.38)\Big]^2} = \frac{10.51}{0.25 \cdot 0.75 \cdot \Big[\log(1.38)\Big]^2} \\ &\simeq 540.34 \end{aligned}$$

So we need a total of 541 *events.*

If we want to compute it by SAS, assuming accrual period as three years and follow up as 1 year, then:

```
/*part D*/
proc power;
    twosamplesurvival test = logrank
    curve("control") = (1):(0.93)
    refsurvival = "control"
    hazardratio = 1.38
    accrualtime = 3
    followuptime = 1
    groupweights = (1 3)
    eventstotal = .
    power= 0.9
    ;
run;
```

Computed Ceiling Event Total		
Fractional Event Total	Actual Power	Ceiling Event Total
607.238272	0.900	608

so we need 608 *total events according to SAS.*

To compute the total sample size, we have:

```
/*total sample size*/
proc power;
    twosamplesurvival test = logrank
    curve("control") = (1):(0.93)
    refsurvival = "control"
    hazardratio = 1.38
    accrualtime = 3
    followuptime = 1
    groupweights = (1 3)
    ntotal = .
    power= 0.9
    ;
run;
```

Computed N Total	
Actual Power	N Total
0.900	2964

So we need 2964 *total sample size. It is not surprising to see that for unbalanced study it requires more sample size to maintain the same power.*

For part II:

Change the AFT model to be with

$$\exp(\beta_{AFT}) = 2 = S^{-1}_{TRT}(p)/S^{-1}_{CTL}(p)$$

We get the following SAS code:

```
/*part II of Question 2*/
proc power;
    twosamplesurvival test = logrank
    curve("control") = (0.25 0.5 0.75 1 1.2 1.5 2 2.75 6)
                       :(0.95 0.9 0.65 0.5 0.375 0.225 0.15 0.075 0.05)
    curve("treatment") = (0.5 1 1.5 2 2.4 3 4 5.5 12)
                       :(0.95 0.9 0.65 0.5 0.375 0.225 0.15 0.075 0.05)
    gsurv = "control"|"treatment"
    accrualtime = 3
    followuptime = 2
    npergroup = .
    power = 0.8
    alpha = 0.1
    sides = 1;
run;
```

Computed N per Group	
Actual Power	N per Group
0.810	14

So we need 14 *sample size in each group, namely 28 in total. Also accouting for an O/N ratio* $\simeq 0.90$*, so we need* $14/0.9 \simeq 16$ *in each arm, so the total sample size will be 32.*

Question #3:

Solution 3. *For part I, we solve exercise* 5.2 *by following the algorithm presented by lecture notes.*

The data is given as the following:

Time (t) After Graduation (in Months)	*Number Who Had Needlestick Injury at Time t*	*Number Who Had Needlestick Injury Prior to Time t*	*Number Who Never Had Needlestick Injury at Time t*
2	3	0	0
4	2	0	0
8	1	0	0
10	2	1	0
12	4	2	0
15	6	2	1
20	3	4	1
24	3	3	2
28	2	3	3
34	1	4	5
41	0	2	3
62	0	3	4
69	0	2	6
75	0	1	6
79	0	2	3
86	0	3	7
Total	27	32	41

Our SAS code is as following:

```
dm 'log; clear; output; clear;';
proc iml;
    /*input our data and set up initial values for important vectors*/
    /*t is the time grid*/
    t = {0, 2, 4, 8, 10, 12, 15, 20, 24, 28, 34, 41, 62, 69, 75, 79, 86};
    /*initialize survival. S[1] is for t = 0 so it is always 1*/
    S = {1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0};
    m = 17;
    /*c is the left censoring number at each time*/
    c = {0, 0, 0, 0, 1, 2, 2, 4, 3, 3, 4, 2, 3, 2, 1, 2, 3};
    /*d is the number of events at each time*/
    d = {0, 3, 2, 1, 2, 4, 6, 3, 3, 2, 1, 0, 0, 0, 0, 0, 0};
    /*r is the number of right censoring at each time*/
    r = {0, 0, 0, 0, 0, 0, 1, 1, 2, 3, 5, 3, 4, 6, 6, 3, 7};

    /*step 0: produce an initial estimate of the survival function S_0(t_j)
 by using product limits estimate (KM)*/
    /*computing risking set, ignore left censoring*/
    Y = {68, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0};
    do i = 2 to m;
        do j = i to m;
            Y[i] = Y[i] + d[j] + r[j];
        end;
    end;
    print y;
    do j = 2 to m;
        S[j] = S[j - 1]*(Y[j] - d[j])/Y[j];
    end;
    /*we checked that we have the same result as from proc lifetest for step 0*/
    print s;


/*k is a fake index here. just to let the same program run 3 times*/
do k = 1 to 3;
    /*Step (k)1- (k)2: estimate p_ij, estimate the number of events at time t_j*/
    do j = 2 to m;
        do i = j to m;
            d[j] = d[j] + c[i]*(s[j - 1]- s[j])/(1 - s[i]);
        end;
    end;
    /*Step (k)3 compute the product limit estimate as in step 0, using the new d*/
    Y = {100, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0};
    do i = 2 to m;
        do j = i to m;
            Y[i] = Y[i] + d[j] + r[j];
        end;
    end;
    do j = 2 to m;
        S[j] = S[j - 1]*(Y[j] - d[j])/Y[j];
    end;
end;
print s;
quit;
```

The output for the survival function estimate after 3 iteration is:

S
1
0.8963466
0.8272443
0.7926932
0.7235909
0.5993282
0.4389609
0.3644963
0.3011747
0.2613354
0.2413775
0.2413775
0.2413775
0.2413775
0.2413775
0.2413775
0.2413775

During the discussion with classmates, I have noticed that a few of those who did this problem with excel have different results than mine. With a closer check, I find that my result after the first iteration is still the same as theirs, also my estimate at step 0 is the same as when I check with proc lifetest. But apprently there is something wrong here, because the survival curve from proc iclifetest later on clearly shows that the survival estimates should stay above 0.40 *at all time. However I could not figure out what is wrong with my code.*

For part II:

We give the following SAS code:

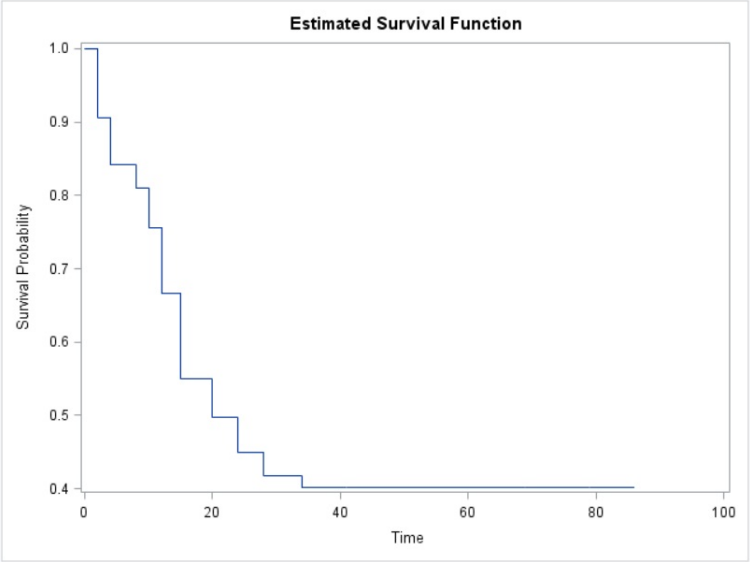
```
/*Question 3 part II*/

/*reformat the data*/
data ex5_2;
    input ltime rtime @@;
    datalines;
    2 2     2 2     2 2     4 4     4 4     8 8
    10 10   10 10   . 10    12 12   12 12   12 12
    12 12   . 12    . 12    15 15   15 15   15 15
    15 15   15 15   15 15   . 15    . 15    15 .
    20 20   20 20   20 20   . 20    . 20    . 20
    . 20    20 .    24 24   24 24   24 24   . 24
    . 24    . 24    24 .    24 .    28 28   28 28
    . 28    . 28    . 28    28 .    28 .    28 .
    34 34   . 34    . 34    . 34    . 34    34 .
    34 .    34 .    34 .    34 .    . 41    . 41
    41 .    41 .    41 .    . 62    . 62    . 62
    62 .    62 .    62 .    62 .    . 69    . 69
    69 .    69 .    69 .    69 .    69 .    69 .
    . 75    75 .    75 .    75 .    75 .    75 .
    75 .    . 79    . 79    79 .    79 .    79 .
    . 86    . 86    . 86    86 .    86 .    86 .
    86 .    86 .    86 .    86 .
    ;
run;

proc iclifetest data=ex5_2 impute(seed = 1234) outsurv = compare;
    time (ltime rtime);
run;

proc print data=compare;
    var leftboundary rightboundary survprob;
run;
```

and the estimate for survival is:



Obs	LeftBoundary	RightBoundary	SurvProb
1	2	2	0.90507
2	4	4	0.84179
3	8	8	0.81014
4	10	10	0.75590
5	12	12	0.66714
6	15	15	0.54965
7	15	20	0.54965
8	20	20	0.49805
9	20	24	0.49805
10	24	24	0.44974
11	24	28	0.44974
12	28	28	0.41805
13	28	34	0.41805
14	34	34	0.40213
15	34	41	0.40213
16	41	62	0.40213
17	62	69	0.40213
18	69	75	0.40213
19	75	79	0.40213
20	79	86	0.40213
21	86	I	0.00000

Question #4

Solution 4. *The following code read in the given data:*

```
data Q4;
    /*for group, 1 = adult, 0 = children    */
    input group left right @@;
    datalines;
    1 0 4   1 0 24  1 0 24  1 0 36  1 0 36  1 0 36  1 4 8   1 4 8
    1 4 8   1 4 8   1 4 8   1 4 24  1 4 24  1 4 24  1 4 36  1 4 48
    1 8 12  1 8 12  1 8 12  1 8 12  1 8 36  1 12 24 1 12 24 1 12 24
    1 12 24 1 12 24 1 12 48 1 12 48 1 12 48 1 12 48 1 24 36 1 24 36
    1 24 36 1 24 36 1 36 48 1 36 48 1 36 48 1 36 48 1 48 .  1 48 .
    1 48 .  1 48 .  1 48 .  1 48 .  1 48 .  1 48 .   0 0 12  0 0 12
    0 8 12  0 8 36  0 8 36  0 8 48  0 12 24 0 12 48 0 24 36 0 24 36
    0 24 36 0 24 36 0 24 36 0 24 36 0 36 48 0 36 48 0 36 48 0 36 48
    0 36 48 0 36 48 0 36 48 0 48 .  0 48 .  0 48 .  0 48 .  0 48 .
    0 48 .  0 48 .  0 48 .  0 48 .  0 48 .  0 48 .  0 48 .  0 48 .
    0 48 .  0 48 .  0 48 .  0 48 .  0 48 .  0 48 .  0 48 .  0 48 .
    0 48 .  0 48 .  0 48 .  0 48 .  0 48 .
    ;
run;

data cov;
    group = 0; output;
    group = 1; output;
run;
```

and the following code fit piecewise exponential model and cubic spline model separately(the screenshot was made on my laptop without 13.2 patch of SAS 9.4, that is why the color of text looks different. But the code is runnable on the computer in the basement):

```
/*fit the piecewise exponential model*/
proc icphreg data= Q4;
    class group;
    model (left, right) = group/basehaz = piecewiseexponential;
    hazardratio group;
    baseline covariates = cov;
run;

/*fit the cubic spline model*/

proc icphreg data=Q4;
    class group;
    model (left, right) = group/basehaz = splines;
    hazardratio group;
    baseline covariates = cov;
run;
```

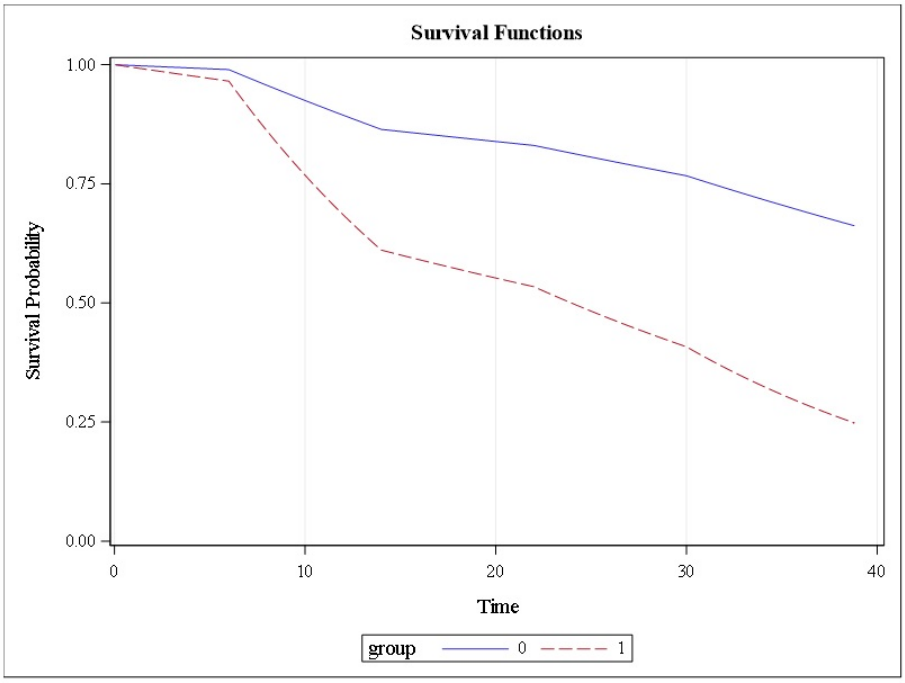
The estimate for the piecewise exponential model is:

Analysis of Maximum Likelihood Parameter Estimates								
Effect	group	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Haz1		1	0.0017	0.0017	0.0000	0.0052		
Haz2		1	0.0169	0.0060	0.0051	0.0287		
Haz3		1	0.0050	0.0054	0.0000	0.0156		
Haz4		1	0.0100	0.0096	0.0000	0.0288		
Haz5		1	0.0167	0.0055	0.0060	0.0274		
group	1	1	1.2172	0.2779	0.6725	1.7619	19.18	<.0001
group	0	0	0.0000					

Hazard Ratios for group			
Description	Point Estimate	95% Wald Confidence Limits	
group 1 vs 0	3.378	1.959	5.824

As we can see that the p value for group effect is significant (< 0.0001)*, so there is a difference between adult and children when it comes to the judgement of shelf life time. The point estimate of the hazard ratio of adult vs children is* 3.378*, indicates that the shelf life from adults' judgement is shorter than that from children.*

The plot of survival curve supports this:



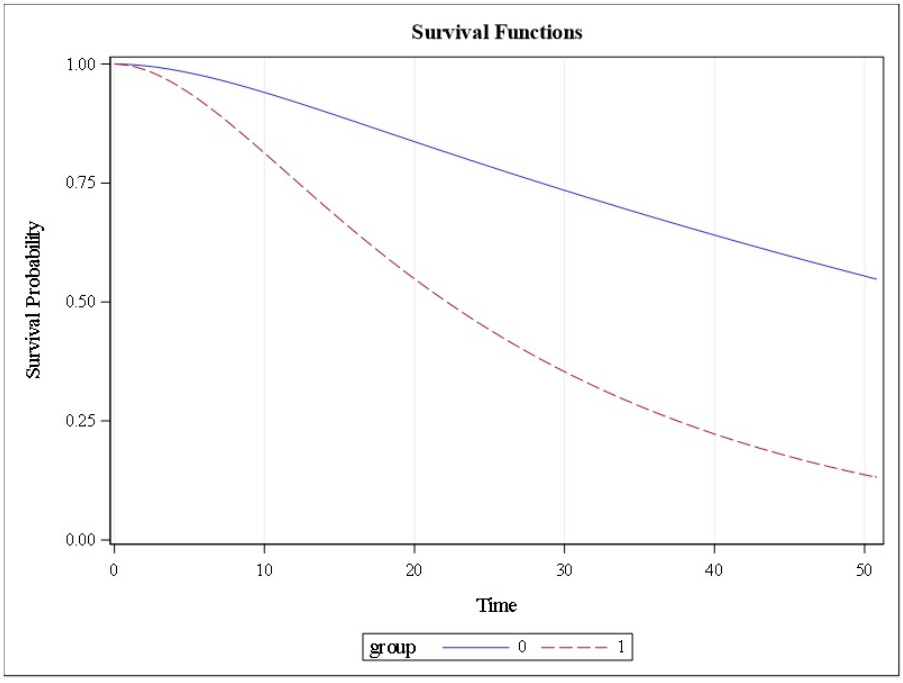
The estimate for the cubic spline model is:

Analysis of Maximum Likelihood Parameter Estimates								
Effect	group	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Coef1		1	-6.9408	1.4333	-9.7500	-4.1316		
Coef2		1	1.8514	0.6146	0.6468	3.0561		
Coef3		1	0.0987	0.1277	-0.1515	0.3489		
group	1	1	1.2157	0.2778	0.6713	1.7602	19.16	<.0001
group	0	0	0.0000					

Hazard Ratios for group			
Description	Point Estimate	95% Wald Confidence Limits	
group 1 vs 0	3.373	1.957	5.813

We also got significant p value(< 0.0001) for the group effect(adults vs children), and the point estimate (3.373) suggests the same conclusion as in piecewise exponential model.

A plot of survival curves is provided too:



Again we see the shelf life time judged by adults are shorter than children.

Question #5

Solution 5. *I confirm that I have read page* 149 *to page* 150 *of the textbook about "Nonparametric estimation of the survival function for right-truncated data" and understood how the calculations are done.*