Question #1:

Solution 1. *For part A, we have the following SAS code:*

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

libname data "C:\akira\data";

proc format;
    value treatment 0 = "CPVM"
                    1 = "BCG";
run;

/*Q1 part a*/
proc lifetest data=data.Melanoma plots=s;
    time surv*dead(0);
    strata trt;
run;
```

The table of KM estiamtes for group CPVM is (the SURV column shows the event/censoring time, censored time denoted by $*$. The survival column shows estimate of survival time):

Stratum 1: 1=BCG/0=CPVM = 0

Product-Limit Survival Estimates

SURV		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000		1.0000	0	0	0	11
3.9000		0.9091	0.0909	0.0867	1	10
5.4000		0.8182	0.1818	0.1163	2	9
6.9000		0.7273	0.2727	0.1343	3	8
7.9000		0.6364	0.3636	0.1450	4	7
8.2000	*	.	.	.	4	6
8.2000	*	.	.	.	4	5
10.5000		0.5091	0.4909	0.1625	5	4
16.6000	*	.	.	.	5	3
16.9000	*	.	.	.	5	2
17.1000	*	.	.	.	5	1
23.8000	*	0.5091	0.4909	.	5	0

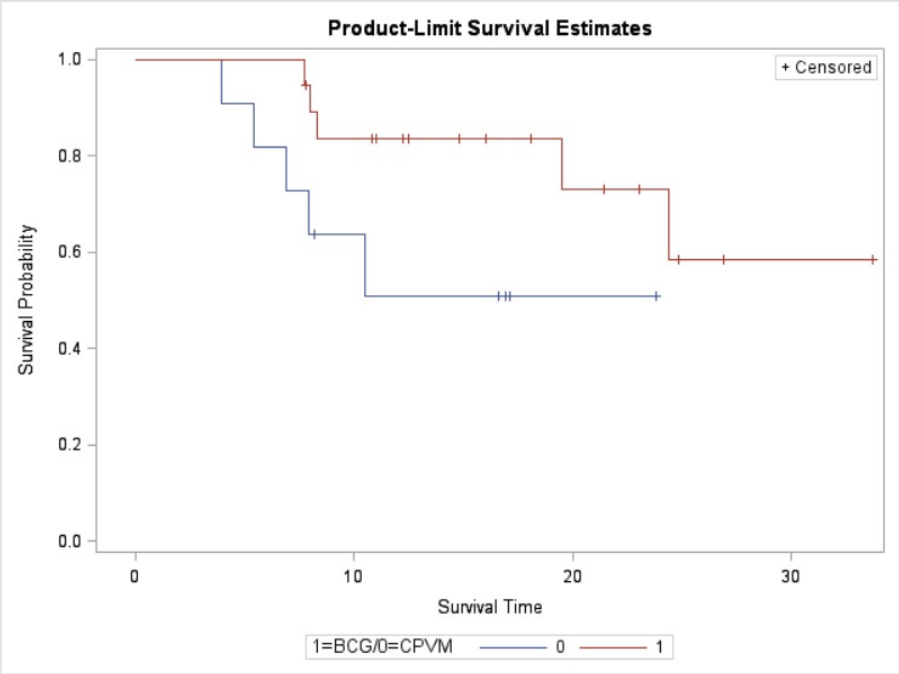
The table of KM estimates for group BCG is:

Stratum 2: 1=BCG/0=CPVM = 1

Product-Limit Survival Estimates

SURV		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000		1.0000	0	0	0	19
7.7000		0.9474	0.0526	0.0512	1	18
7.8000	*	.	.	.	1	17
8.0000		0.8916	0.1084	0.0724	2	16
8.3000		0.8359	0.1641	0.0867	3	15
10.8000	*	.	.	.	3	14
11.0000	*	.	.	.	3	13
12.2000	*	.	.	.	3	12
12.5000	*	.	.	.	3	11
14.8000	*	.	.	.	3	10
16.0000	*	.	.	.	3	9
18.1000	*	.	.	.	3	8
19.5000		0.7314	0.2686	0.1237	4	7
21.4000	*	.	.	.	4	6
23.0000	*	.	.	.	4	5
24.4000		0.5851	0.4149	0.1641	5	4
24.8000	*	.	.	.	5	3
26.9000	*	.	.	.	5	2
33.7000	*	.	.	.	5	1
33.7000	*	0.5851	0.4149	.	5	0

and the plot of survival function by treatment group is:



It appears that subjects in the group of BCG has a longer survival time than CPVM group.

Now to estimate the quartiles, remember by definition the pth quantile is defined as:

$$x_p = \inf\{t : S(t) \leq 1 - p\}$$

an estimate is:

$$\hat{x}_p = \inf\{t : \hat{S}(t) \leq 1 - p\}$$

So the 25th percentile of the survival time for group BCG is(according to the table) 19.5*, and the 25th percentile of the survival time for group CPVM is* 6.9*, which also matches with the summary of the SAS output:*

Summary Statistics for Time Variable SURV

Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	24.4000	.
50	.	LOGLOG	19.5000	.
25	19.5000	LOGLOG	7.7000	.

Summary Statistics for Time Variable SURV

Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	10.5000	.
50	.	LOGLOG	5.4000	.
25	6.9000	LOGLOG	3.9000	.

However as we notice that there is no estimate for the median and 75th percentile, since at the largest event time, the estimate of survival time for each group is still larger than $1 - 0.5 = 0.5$*(we get* $\hat{S}(24.4) = 0.5851$ *for BCG group and* $\hat{S}(23.8) = 0.5091$ *for CPVM group.).*

For part B:

The statistic for log rank test is:

$$T = \sum_{j=1}^{r} W(t_j)(e_{ij} - E_{ij})$$

with $W(t_j) = 1$ *for all* t_j*.*

Here we have:

$$E_{1j} = \frac{n_{1j}}{n_{1j} + n_{2j}}(e_{1j} + e_{2j})$$
$$E_{2j} = \frac{n_{2j}}{n_{1j} + n_{2j}}(e_{1j} + e_{2j})$$

So we make the following table:

(the actual procedure is that we sort the data by survival time in the ascending order, and we pick out the survival time where there is death instead of censoring. Based on the data we can easily compute e_{1j}, e_{2j}, n_{1j} *and* n_{2j} *and hence further compute* E_{1j} *and* E_{2j}*).*

t_j	e_{1j}	e_{2j}	n_{1j}	n_{2j}	E_{1j}	E_{2j}	$e_{1j} - E_{1j}$	$e_{2j} - E_{2j}$
3.9	1	0	11	19	$\frac{11}{30} \times 1$	$\frac{19}{30} \times 1$	$1 - \frac{11}{30}$	$0 - \frac{19}{30}$
5.4	1	0	10	19	$\frac{10}{29} \times 1$	$\frac{19}{29} \times 1$	$1 - \frac{10}{29}$	$0 - \frac{19}{29}$
6.9	1	0	9	19	$\frac{9}{28} \times 1$	$\frac{19}{28} \times 1$	$1 - \frac{9}{28}$	$0 - \frac{19}{28}$
7.7	0	1	8	19	$\frac{8}{27} \times 1$	$\frac{19}{27} \times 1$	$0 - \frac{8}{27}$	$1 - \frac{19}{27}$
7.9	1	0	8	17	$\frac{8}{25} \times 1$	$\frac{17}{25} \times 1$	$1 - \frac{8}{25}$	$0 - \frac{17}{25}$
8.0	0	1	7	17	$\frac{7}{24}$	$\frac{17}{24}$	$0 - \frac{7}{24}$	$1 - \frac{17}{24}$
8.3	0	1	5	16	$\frac{5}{21} \times 1$	$\frac{16}{21} \times 1$	$0 - \frac{5}{21}$	$1 - \frac{16}{21}$
10.5	1	0	5	15	$\frac{5}{20} \times 1$	$\frac{15}{20} \times 1$	$1 - \frac{5}{20}$	$0 - \frac{15}{20}$
19.5	0	1	1	8	$\frac{1}{9} \times 1$	$8\frac{8}{9} \times 1$	$0 - \frac{1}{9}$	$1 - \frac{8}{9}$
24.4	0	1	0	5	$\frac{0}{5} \times 1$	$\frac{5}{5} \times 1$	0	0

From the table above, we can compute the log rank-statistic as:

$$\begin{aligned} T &= \sum_{j=1}^{r}(e_{1j} - E_{1j}) \\ &= (1 - \frac{11}{30}) + (1 - \frac{10}{29}) + (1 - \frac{9}{28}) + (0 - \frac{8}{27}) + (1 - \frac{8}{25}) + (0 - \frac{7}{24}) \\ &\qquad + (0 - \frac{5}{21}) + (1 - \frac{5}{20}) + (0 - \frac{1}{9}) \\ &= 2.459908 \end{aligned}$$

Given $\hat{Var}(T) = 1.78$*, we have:*

$$\chi^2 = \frac{T^2}{\hat{Var}(T)} = \frac{2.459908^2}{1.78} = 3.399521$$

(SAS gives the value as 3.4060*, this is just due to rounding error.)*

We have

$$p = P(\chi_1^2 \geq 3.399521) = 0.065(\text{ same } p \text{ value as the one given by SAS})$$

wich is not significant at $\alpha = 0.05$ *significance level. Hence we fail to reject* $H_0 : S_{BCG}(t) = S_{CPVM}(t)$*.*

For part (C):

We have the following SAS code:

```
proc lifetest data= data.Melanoma;
     time surv*dead(0);
     strata trt/test = (logrank wilcoxon);
run;
```

and the output is:

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	3.4060	1	0.0649609
Wilcoxon	4.0561	1	0.0440110

The log-rank test has a p value 0.065 *which is consistent with our hand computation in part B, and we will fail to reject the null hypothesis that the two treatment group has the same survival curve (or hazard rate), while the p value for wilcoxon test is* 0.044 *which is significant at the significance level* $\alpha = 0.05$*, and we wil reject and conclude that the hazard rate (or survival curve) is different between the two treatment groups.*

Although the two tests give different conclusions,from the sketch of KM survival curves(which we already showed in part A). it seems that group BCG seems to have a longer survival time than

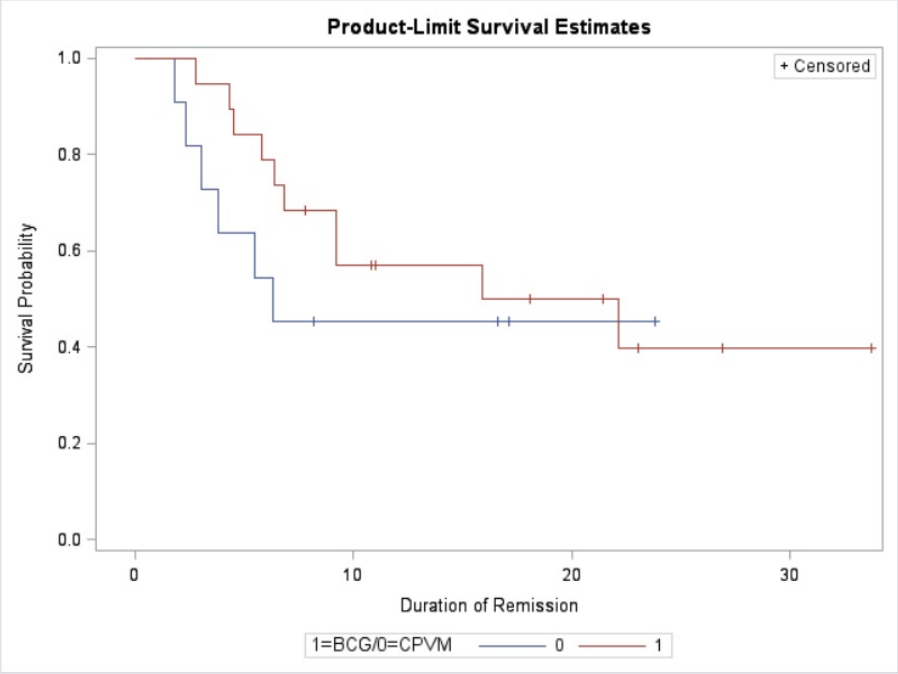
group CPVM. Both 0.044 *and* 0.065 *are actually pretty close to* 0.05*, and we may think that the p value for logrank test is also marginally significant.*

For part (D):

We have the following SAS code:

```
proc lifetest data= data.Melanoma;
     time remdur*recur(0);
     strata trt/test = (logrank wilcoxon);
run;
```

The KM survival curve is:



From the graph we see that the survival curves actually cross.

The result of logrank and wilcoxon test is:

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.7520	1	0.3858301
Wilcoxon	1.8893	1	0.1692787

*Both p values are non-significant (*0.39 *and* 0.17*) and hence we fail to reject and conclude there is not enough evidence showing the difference of hazard rate (or survival curves).*

So it appears that when we consider survival time as outcome, our tests are inclined to conclude that there is a difference of survival time between groups, while considering time to remission as outcome, the tests do not show enough evidence to show the difference of remission period between groups.

For part (E):

The test statement in lifetest proeceudre automatically execute a forward stepwise selection, the code is:

```
/*Q1 part E*/

proc lifetest data=data.Melanoma;
    time surv*dead(0);
    test age  sex  trt;
run;
```

The output is:

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test

Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment	Label
TRT	1	3.6077	0.0575136	3.6077	0.0575136	1=BCG/0=CPVM
SEX	2	5.2017	0.0742089	1.5941	0.2067457	1=Female/0=Male
AGE	3	6.1515	0.1044692	0.9497	0.3297900	Age in Years

*Only the p value for treatment group is marginally significant (*0.058*), which matches with our analysis in part C. However the p values for sex(*0.074*) and p value for age* (0.104) *are not significant and we do not need to include them if we are considering any parametric model in the future.*

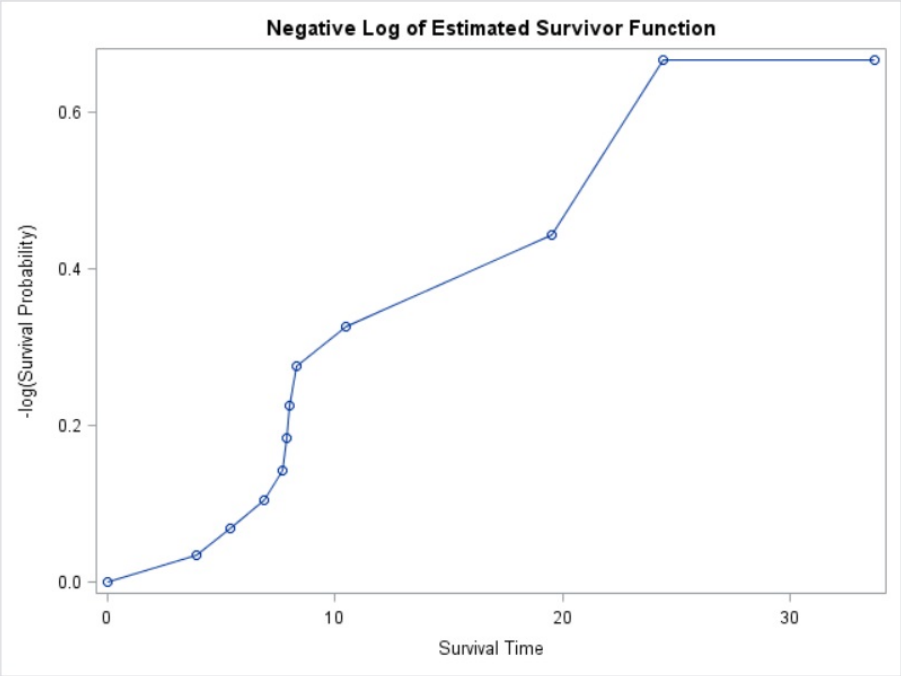
*Of course as we know that the lifetest procedure is very incompetent when it comes to looking at several covariates simultaneously. So further analysis should be conducted as we will do in answering Question #*2*.*

For part (F):

We make graphs about negative log survive versus survival time and log-log survival versus log survival time, the code is following:

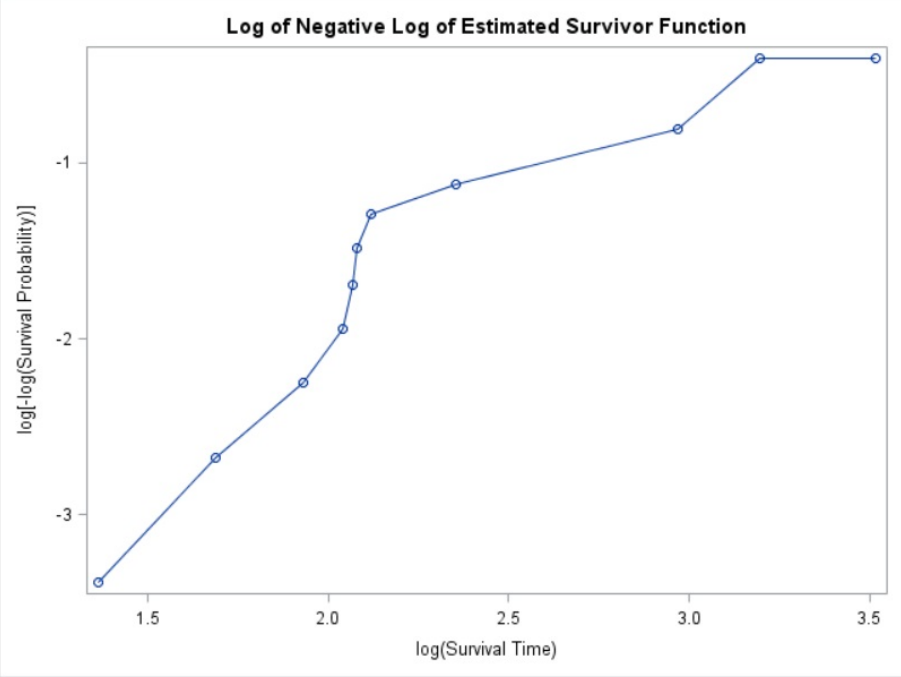
```
/*Q1 part F*/
proc lifetest data=data.Melanoma plots=(ls lls);
    time surv*dead(0);
run;
```

The negative log survival versus survival time is:



It is not quite a straightline, so the hazard function may not be a constant, and hence we may not consider exponential AFT model in the future.

The log-log survival versus log survival time is:



It is more like a straightline compared to the first graph, so the hazard function may be in a multiplicative form, and we may consider the weibull AFT model in the future.

Question #2.

Solution 2. *For part (A):*

Method*:*

Step 1*: To select for the the best model to fit the Melanoma data, we use proc lifereg in SAS to fit expoential, weibull, generalized gamma, lognormal and log-logistic models considering three covariates treatmentgroup, sex and age.*

The SAS code is the following:

```
/*Q2*/
/*exponential AFT*/
proc lifereg data=data.Melanoma;
    class trt;
    model surv*dead(0) = trt sex age/dist = exponential;
run;

/*weibullAFT*/
proc lifereg data=data.Melanoma;
    class trt;
    model surv*dead(0) = trt sex age/dist = weibull;
run;

/*generalized gamma AFT*/
proc lifereg data=data.Melanoma;
    class trt;
    model surv*dead(0) = trt sex age/dist = gamma;
run;

/*log-normal AFT*/
proc lifereg data=data.Melanoma;
    class trt;
    model surv*dead(0) = trt sex age/dist = lnormal;
run;

/*log-logistic AFT*/
proc lifereg data=data.Melanoma;
    class trt;
    model surv*dead(0) = trt sex age/dist = llogistic;
run;
```

Step 2*: We run the likelihood ratio tests among exponential, weibull and generalized gamma models since they are nested models. We also compare AIC among all the above five models to compare between non-nested models. Those with smaller AIC values are more favored.*

From the output in **Step 1** *we have:*

Name of Distribution	Exponential
Log Likelihood	-23.18007426

Name of Distribution	Weibull
Log Likelihood	-21.73695777

Name of Distribution	Gamma
Log Likelihood	-19.14139434

We run the likelihood ratio test with the following code:

```
/*log-logistic AFT*/
proc lifereg data=data.Melanoma;
    class trt;
    model surv*dead(0) = trt sex age/dist = llogistic;
run;

/*goodness of fit for nested models: exp, weibull, generalized gamma*/
data fit;
    pvalue_gamma_vs_exp =1 - probchi(46.360 - 38.283, 2);
    pvalue_gamma_vs_weib= 1 - probchi(43.474 -38.283, 1);
    pvalue_weib_vs_expo = 1 - probchi(46.360 - 43.474, 1);
run;

proc print data=fit;
run;
```

and we got the p values as :

Obs	pvalue_gamma_vs_exp	pvalue_gamma_vs_weib	pvalue_weib_vs_expo
1	0.017624	0.022704	0.089353

The above p values tell that there is a significant lack of adequacy of weibull and exponential AFT model, compared to generalied gamma(p values 0.023 *and* 0.017*). However the exponential AFT model is adequate compared to weibull(p values* 0.089*).*

The AIC values for the above five models are:

distribution	*AIC*
exponential	54.360
weibull	53.474
generalized gamma	50.283
log normal	52.030
log-logistic	53.144

We may be tempted to choose generalized gamma here based on the small AIC value and the likelihood ratio test. However we should not consider generalized gamma as a dependable model in this situation, because SAS gives the following warning for the generalized gamma model:

```
WARNING: The relative gradient convergence criterion of 0.0412540986 is greater than the limit of
         0.0001. The convergence is questionable.
WARNING: The procedure is continuing in spite of the above warning. Results shown are based on
         the last maximum likelihood iteration. Validity of the model fit is questionable.
NOTE: PROCEDURE LIFEREG used (Total process time):
      real time           0.04 seconds
      cpu time            0.01 seconds
```

On the other hand, log-normal has the smallest AIC value(if we take out generalized gamma), so we should consider log-normal as a more favorable candidate.

Between Weibull and exponential, we would still favor weibull over exponential for two reasons. One is that weibull has smaller AIC values, the other is that we reject the test on scale parameter equal to 1 *(p value* 0.0000801*):*

Lagrange Multiplier Statistics		
Parameter	Chi-Square	Pr > ChiSq
Scale	15.5560	0.0000801

In fact the estimate of the scale parameter when fitting the weibull AFT model is: 0.5950. *(Exponential AFT model requires the scale parameter to be* 1 *however).*

Step 3*: We also use proc nlmixed in SAS to fit a Gompertz model. The nlmixed procedure usea different way to compute the likelihood, so we can not directly compare the likelihood from this procedure with those from the lifereg procedure. Accordingto the comments from Dale Mclerran we can compare the likelihood between gompertz and exponential (for large negative value of log gamma) when both are coming from the nlmixed procedure.*

```
title "NLMIXED: Gompertz distribution";
proc nlmixed data=data.Melanoma;
    parms log_gamma -5;
    gamma = exp(log_gamma);
    linp = b0 + b1*trt + b2*sex + b3*age;
    alpha = exp(-linp);
    G_t = exp((alpha/gamma)*(1 - exp(gamma*surv)));
    g = alpha*exp(gamma*surv)*G_t;
    ll = (dead=1)*log(g) + /* ll for observed failures */
        (dead=0)*log(G_t); /* ll for censored failures */
    model ll ~ general(ll);
    estimate "gamma" exp(log_gamma);
run;
```

theoretically when log_gamma is a large negative value, the Gompertz model is close to exponential. During testing though we could not make it go beyond -8 *on our data otherwise there will be the following warning:*

```
NOTE: The parameters b0, b1, b2, b3 are assigned the default starting value of 1.0, because they
      are not assigned initial values with the PARMS statement.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
NOTE: At least one element of the gradient is greater than 1e-3.
NOTE: Moore-Penrose inverse is used in covariance matrix.
WARNING: The final Hessian matrix is full rank but has at least one negative eigenvalue.
         Second-order optimality condition violated.
```

This is not tolerable since with negative eigenvalues in Hessian matrix, the algorithm will fail.

Plugging any value between -2 *and* -8 *into log_gamma will yield same* $-2log$ *likelihood value as* 88.8*, so we consider that there is no significant improvement from exponential to Gompertz.*

Our final conclusion based on the discussion above is that:

1. *we will not choose generalized gamma due to convergence issue*
2. *we will not choose exponential model because the scale is not* 1 *and it has larger AIC value than weibull*
3. *we will not choose Gompertz because it is not an improvement over exponential*

4. *we will not choose weibull because it has larger AIC than log-normal*

5. *we choose log-normal as our final model*

Results:

From fitting the log-normal model into the data, we have the following estimate:

Analysis of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	3.1289	0.6222	1.9094	4.3484	25.29	0.0000005
TRT	BCG	1	1.0669	0.4486	0.1877	1.9460	5.66	0.0173880
TRT	CPVM	0	0.0000	.	.	.	.	.
SEX		1	0.7093	0.4727	-0.2171	1.6357	2.25	0.1334649
AGE		1	-0.0186	0.0114	-0.0409	0.0037	2.67	0.1023202
Scale		1	0.8556	0.2123	0.5261	1.3915		

The output is translated as following:

1. *there is not significant impact on survival time from sex (p value* 0.13*),*

2. *there is not significant impact on survival time from age (p value* 0.10*),*

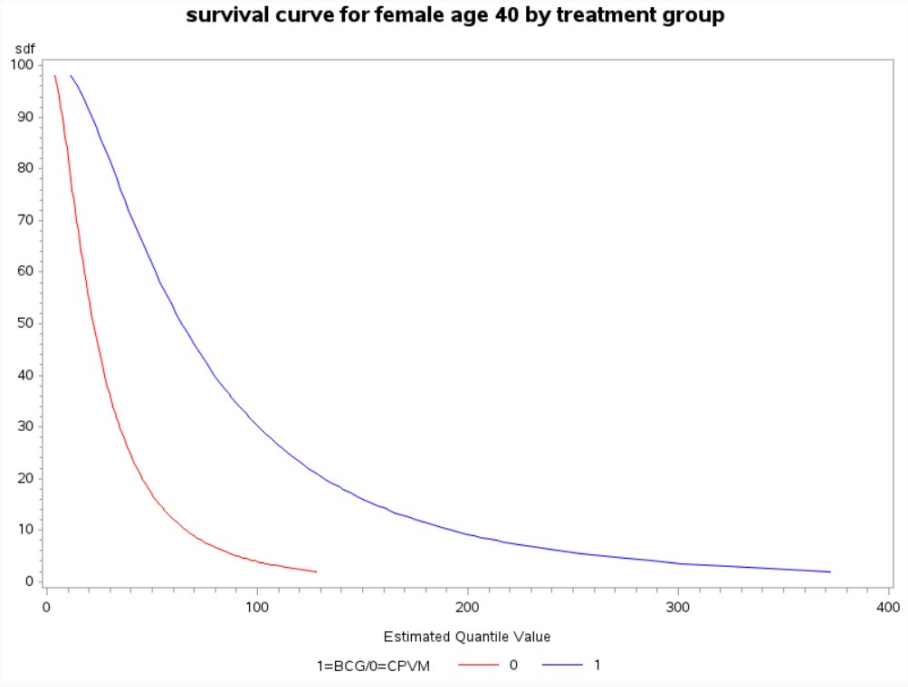
3. *controlling for sex and age, the expected survival time in treatment group BCG is* $exp(1.0669) - 1 = 190\%$ *longer than the treatment group CPVM.*

For part (B)*:*

To graph the survival curve by treatment group for female at age of 40, we have the following code:

```
/*Q2 part B*/
title "survival curve for female age 40 by treatment group"
data one;
    C = 1; trt = 0; age = 40; sex = 1; output;
    C = 1; trt = 1; age = 40; sex = 1; output;
run;

data Melonoma;
    set data.Melanoma one;
run;

proc print data=Melonoma;
run;

proc lifereg data=Melonoma;
    class trt;
    model surv*dead(0) = trt sex age
    /dist = lognormal;
    output out = surv_est quantiles = 0.02 to 0.98 by 0.02
    predicted = pred control = C;
run;

proc print data=surv_est;
run;

data surv_est2;
    set surv_est; sdf = 100*(1 - _prob_);
run;

proc print data=surv_est2;
run;

proc gplot data=surv_est2;
    plot sdf*pred=trt;
    symbol1 I = spline color = red L= 1;
    symbol2 I = spline color = blue L = 1;
run;
```

The output is:



Question #3:

Solution 3. *For part (A):*

For parametric survival models (AFT model), the number of likelihood terms is the same as the number of the subjects. In our case, it is 238.

For Cox-Regression model, when considering partial likelihood for data with ties, suppose at a particular time point the number of observations that tied at this point is m*, then instead of having* m *terms of likelihood we will then have* $m!$ *term involved in computing the likelihood function. So the extra terms involved in the likelihood function is* $m! - m = m((m-1)! - 1)$.

For example, if there is a 2 *tie event, we have two terms coming out of this tie involved in computing likelihood, but if without tie we also have two terms, so the extra terms involved is* $2(1-1) = 0$. *If there is a* 3 *tie event, we have* 6 *terms coming out of this tie, compared to the* 3 *terms without tie, we have extra* $3(2-1) = 3$ *terms involved.*

The following SAS code print out the record of ties:

```
data Addicts;
    set data.Addicts;
    where status = 1;
run;

proc freq data=Addicts noprint;
    table length
    /out = ties;
run;

proc print data=ties;
     where count >= 2;
run;
```

Obs	LENGTH	COUNT	PERCENT
9	35	2	1.33333
11	41	2	1.33333
44	180	2	1.33333
54	212	2	1.33333
55	216	2	1.33333
64	262	2	1.33333
65	268	3	2.00000
105	523	2	1.33333
115	612	2	1.33333
132	821	2	1.33333
138	899	2	1.33333

We only have 1 *event time that is a* 3 *tie event, and the rest are* 2 *tie events or without ties. So a total of extra* 3 *terms are produced out of the* 3 *tie event when computing the partial likelihood.*

From proc freq we also get the following information:

STATUS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	88	36.97	88	36.97
1	150	63.03	238	100.00

We have 150 *event time, so the total number of terms involved in computing partial likelihood function is* $150 + 3 = 153$.

For part (B)*:*

Our proportional hazard model is:

$$h\Big(t, clinic(i), prison(i), dose(i)\Big) = h_0(t) \exp\Big(\beta_1 \cdot clinic(i) + \beta_2 \cdot prison(i) + \beta_3 \cdot dose(i)\Big)$$

For part (C):

The SAS code to fit the model in part (B) is:

```
proc phreg data=data.Addicts;
    model length*status(0) = clinic prison dose/risklimits;
run;
```

part of 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	
CLINIC	1	-0.99359	0.21149	22.0717	0.0000026	0.370	0.245	0.560
PRISON	1	0.32351	0.16793	3.7113	0.0540468	1.382	0.994	1.921
DOSE	1	-0.03473	0.00639	29.5478	<.0000001	0.966	0.954	0.978

The 95% confidence interval for clinic is $(0.245, 0.560)$*. It does not include* 0 *suggesting we should reject the null hypothesis for* $H_0 : \beta_1 = 0$ *(also consistent with p value* $= 0.0000026$*), so we conclude that the effect of clinic is significant at the* 0.05 *significance level.*

For part (D):

We use exact and efron option in the model statement for different ways to handle ties. The exact method is like the one we discussed in part (A)(ii) and lecture notes.

```
/*part D other tie handling*/
proc phreg data=data.Addicts;
    model length*status(0) = clinic prison dose/ties = exact risklimits;
run;

proc phreg data=data.Addicts;
    model length*status(0) = clinic prison dose/ties = efron risklimits;
run;
```

Output for exact:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
CLINIC	1	-0.99935	0.21179	22.2658	0.0000024	0.368	0.243	0.558
PRISON	1	0.32242	0.16789	3.6882	0.0548003	1.380	0.993	1.918
DOSE	1	-0.03484	0.00639	29.6995	<.0000001	0.966	0.954	0.978

Output for efron:

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
CLINIC	1	-0.99934	0.21179	22.2657	0.0000024	0.368	0.243	0.558
PRISON	1	0.32240	0.16788	3.6879	0.0548094	1.380	0.993	1.918
DOSE	1	-0.03484	0.00639	29.6993	<.0000001	0.966	0.954	0.978

as we can see that the eferon option produce a result more close to the exact method than the breslow. The breslow produce a slightly different estimates but the conclusion for inference is the same.

For part (E):

we make our categorical variable and make the dosage < 60 as the baseline dosage.

```
/*part E make dosage a categorical variable*/
data Addicts;
    set data.Addicts;
    if (dose < 60) then do;
        dose1 = 0;
        dose2 =0;
    end;
    else if (60 <= dose < 80) then do;
        dose1 = 1;
        dose2 = 0;
    end;
    else if (dose >=80) then do;
        dose1 = 0;
        dose2 = 1;
    end;
run;

proc print data=Addicts;
run;

proc phreg data=Addicts;
    class dose1(ref = "0") dose2(ref = "0");
    model length*status(0) = clinic prison dose1 dose2 /risklimits;
run;
```

We got the following output:

Analysis of Maximum Likelihood Estimates									
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
CLINIC		1	-0.93021	0.21602	18.5437	0.0000166	0.394	0.258	0.602	
PRISON		1	0.28808	0.16861	2.9190	0.0875422	1.334	0.958	1.856	
dose1	1	1	-0.56208	0.17782	9.9917	0.0015725	0.570	0.402	0.808	dose1 1
dose2	1	1	-1.55852	0.30751	25.6858	0.0000004	0.210	0.115	0.385	dose2 1

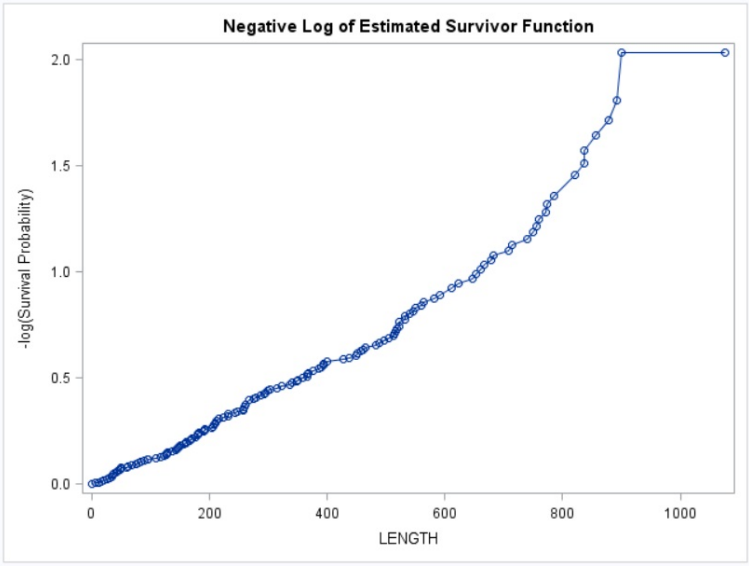
We see that the estimate for the effect of clinic and prison is slightly different, compared to before when we treat dose as a continuous variable. But they are still very close. The interpretaiton is

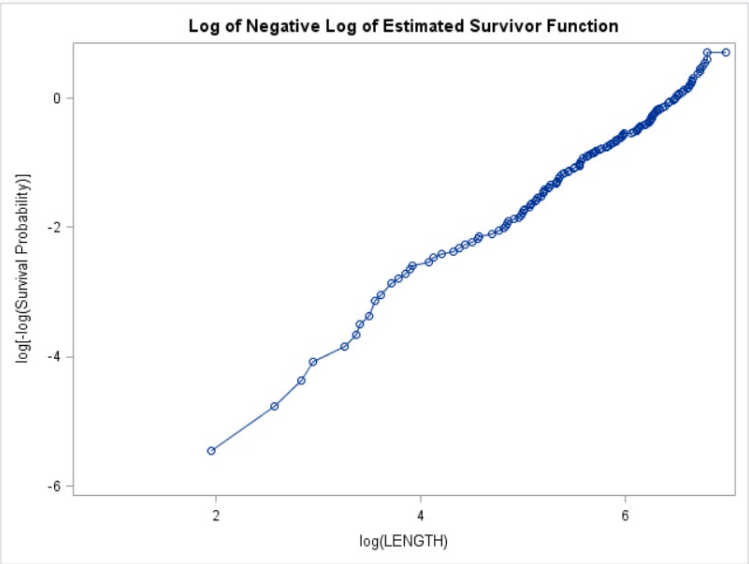
different since previously the effect of clinic and prison is adjusted for fixed numeric value of dosage, while right now the effect of clinic and prison is adjusted for a given range of dosage values.

For part (F):

We may be able to use parametric model instead of Cox model here for our data. Our sample size is large, in fact if we graph the negative log survive versus time and log-log survival versus log time with proc lifetest, we got the following:

```
proc lifetest data=Addicts plots= (ls lls);
    time length*status(0);
run;
```





Both graph displays a very good linear pattern, which makes it possible for us to consider parametric mode like exponential model or weibull model. The parametric model will give us more information on the nature of the data than the cox model.

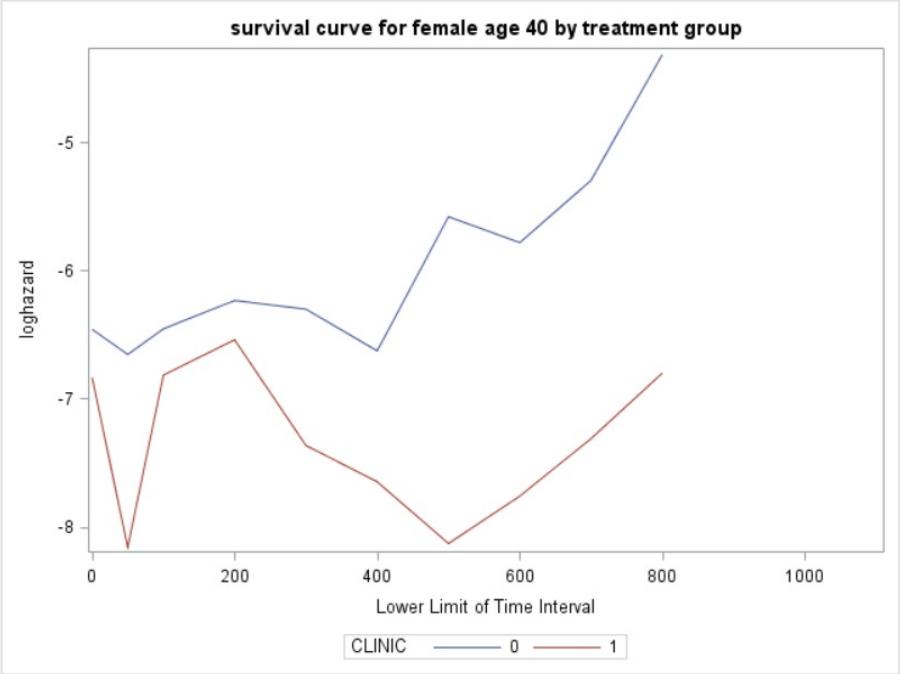
For part (G):

The only way we can get an estimate for hazard function directly is through the lifetable method in proc lifetest. Then we create another variable for log(hazard) and use proc sgplot to plot the log(hazard) versus length variable grouped by clinic and prison separately.

The code for clinic is:

```
proc lifetest data=Addicts method = life outsurv = results
         plots = (H) intervals = 50 100 200 300 400 500 600 700 800 900 1000 1100;
    time length*status(0);
    strata clinic;
run;

data loghazard;
    set results;
    loghazard = log(hazard);
run;

proc sgplot data=loghazard;
    series x = length y = loghazard/group = clinic;
run;
```

and the plot is:

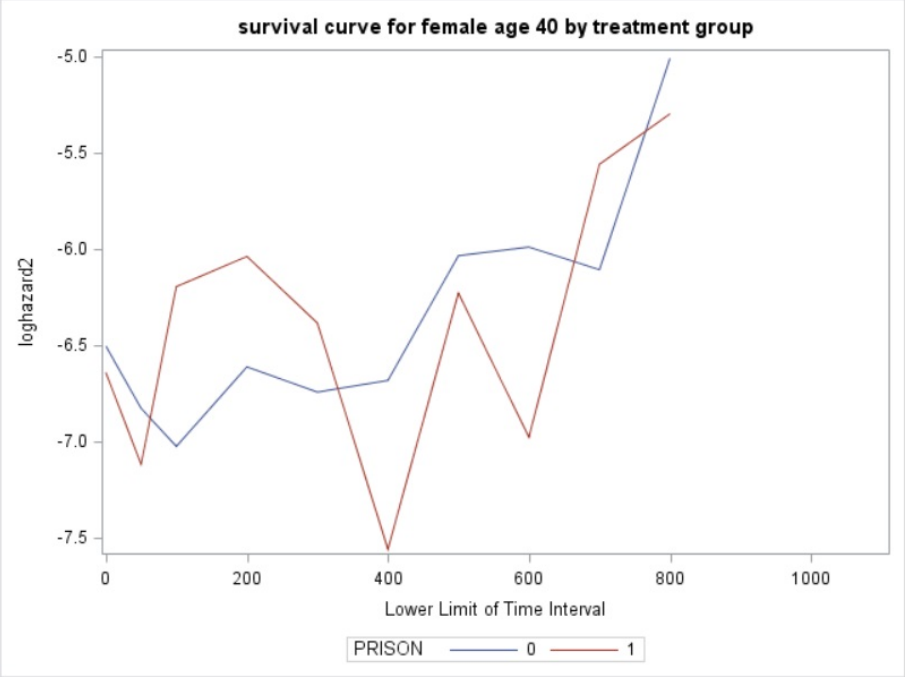


The curves are approximately parallel, supporting the assumption of proportionl hazard with the presence of clinic as a covariate.

The code for checking prisoin is similar:

```
proc lifetest data=Addicts method = life outsurv = results2
        plots = (H) intervals = 50 100 200 300 400 500 600 700 800 900 1000 1100;
    time length*status(0);
    strata prison;
run;

data loghazard2;
    set results2;
    loghazard2 = log(hazard);
run;

proc sgplot data=loghazard2;
    series x = length y = loghazard2/group = prison;
run;
```

and the plot is:



However this time the curves are not parallel, so the PH assumption may not be valid for prison.