**Midterm Exam: BIOS 845**

**Survival Analysis**

**Palash Sharma**

**Date: 04/01/2019**

**Ques No 1(a):**

The Kaplan-Meier estimates for both of these treatment groups were shown below using SAS procedure and the corresponding survival curve shown in the same graph. It seems that group CPVM has higher survival probability than the group BCG which implies that group CPVM tend to have more longevity than group BCG.

Also, corresponding log rank test and a Wilcoxon test were performed to assess the survival time for these two treatment groups.

**Log rank Test:**

Chi- square test statistics using log rank test is 3.4060 with 1 d.f. and the corresponding value is

**Decision Rule**:

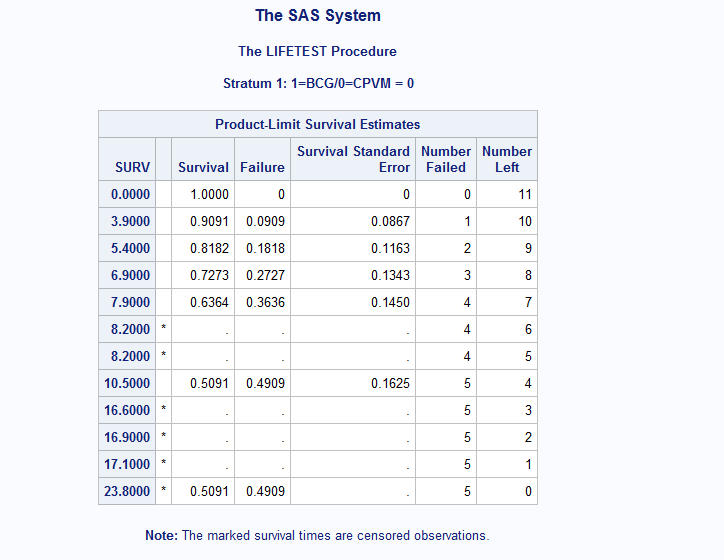
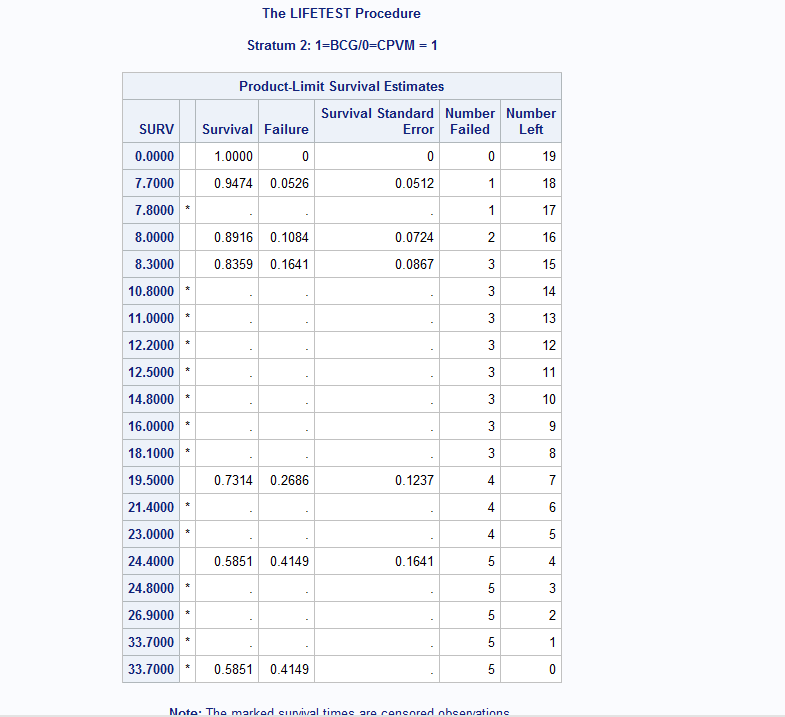
Since value greater that 0.05, we fail to reject the null hypothesis in favor of alternative hypothesis. Our conclusion is, we don’t have enough evidence to support the claim that survival curve from these two-treatment group is statistically different.

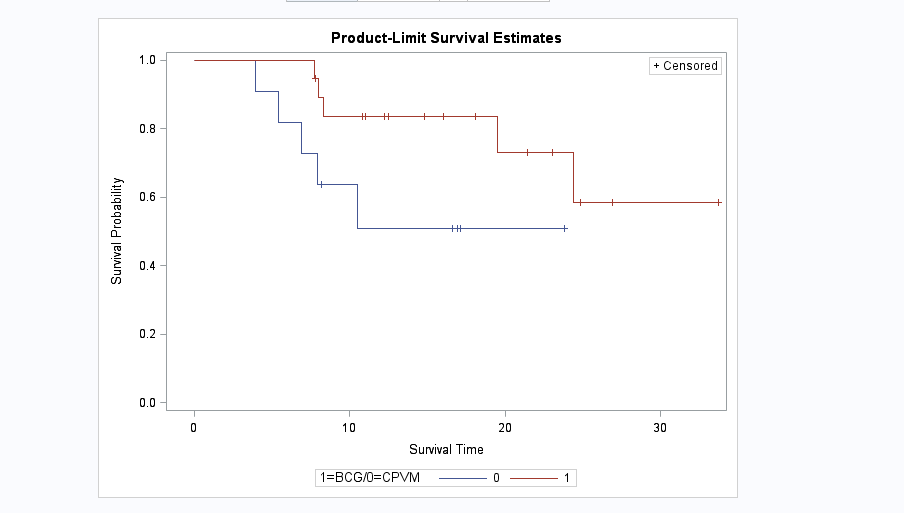
**Wilcoxon Test:**

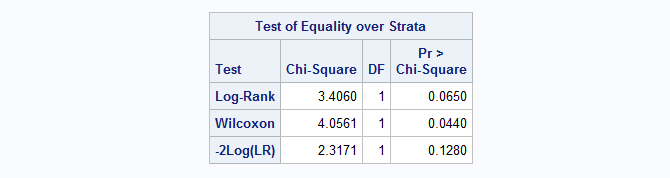
Chi- square test statistics using Wilcoxon test is 4.0561 with 1 d.f. and the corresponding value is

**Decision Rule**:

Since value less that 0.05, we reject the null hypothesis in favor of alternative hypothesis. Our conclusion is that survival curve from these two-treatment group is statistically different using Wilcoxon test.







**Ques No 1(b):**

In this analysis we are comparing time to remission among of these two treatment groups. The Kaplan-Meier estimates for both treatment groups were shown below using SAS procedure and the corresponding curve shown in the same graph. It seems that group CPVM has higher survival probability than the group BCG which implies that group CPVM tend to have more longevity than group BCG.

Also, corresponding log rank test and a Wilcoxon test were performed to assess the remission time for these two treatment groups.

**Log rank Test:**

Chi- square test statistics using log rank test with d.f. and the corresponding value is

**Decision Rule**:

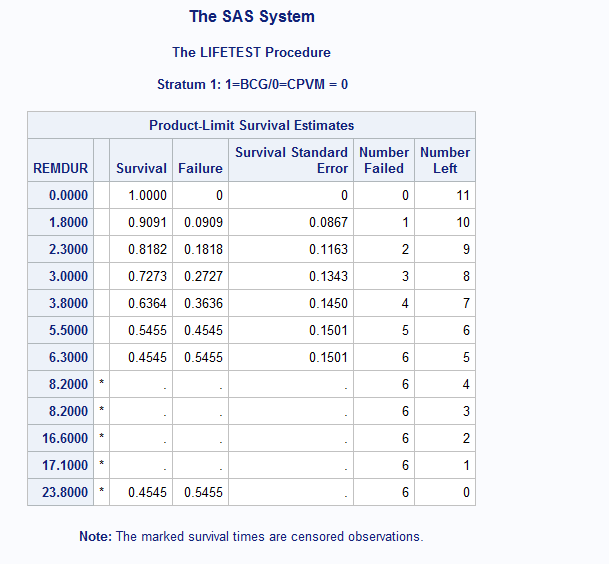
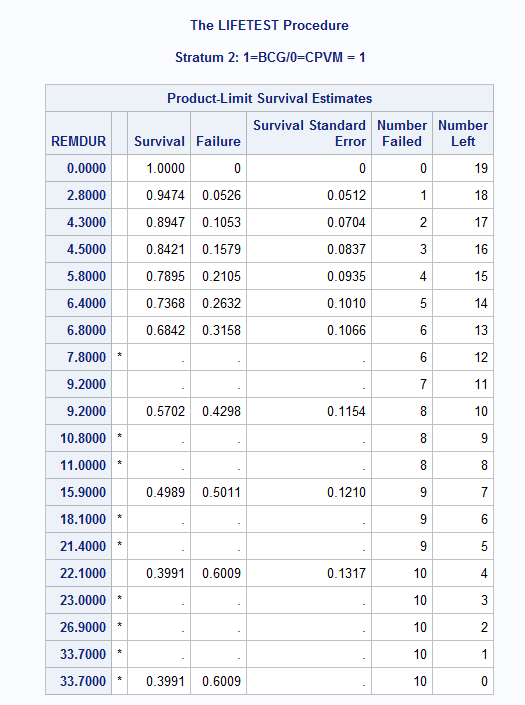
Since value greater that 0.05, we fail to reject the null hypothesis in favor of alternative hypothesis. Our conclusion is, we don’t have enough evidence to support the claim that time to remission curve from these two-treatment groups are statistically different.

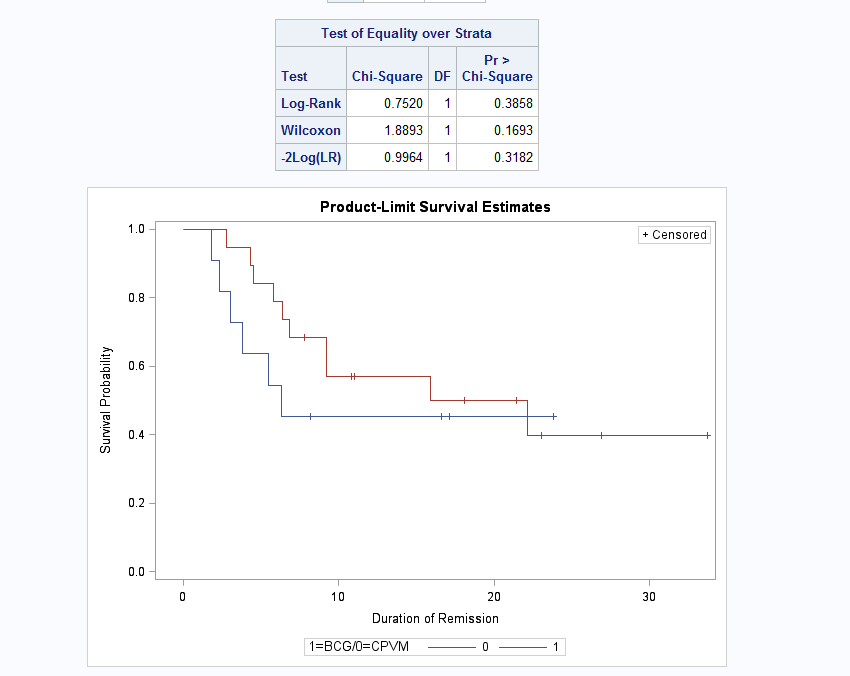
**Wilcoxon Test:**

Chi- square test statistics using Wilcoxon test is with 1 d.f. and the corresponding value is

**Decision Rule**:

Since value greater that 0.05, we fail to reject the null hypothesis in favor of alternative hypothesis. Our conclusion is, we don’t have enough evidence to support the claim that time to remission curve from these two-treatment groups are statistically different.

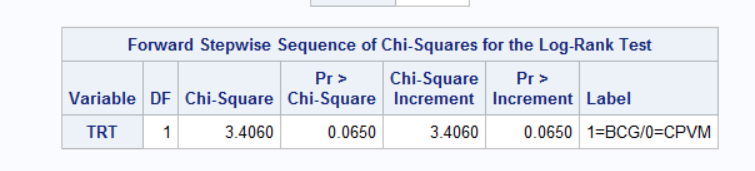


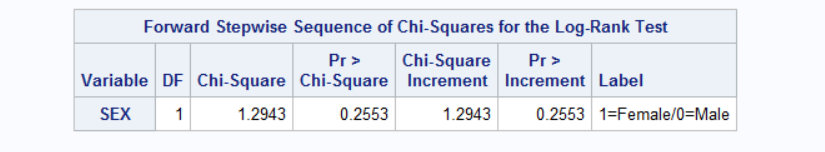


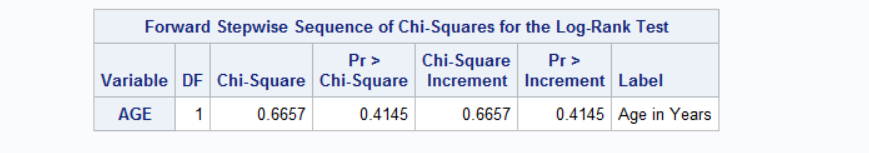
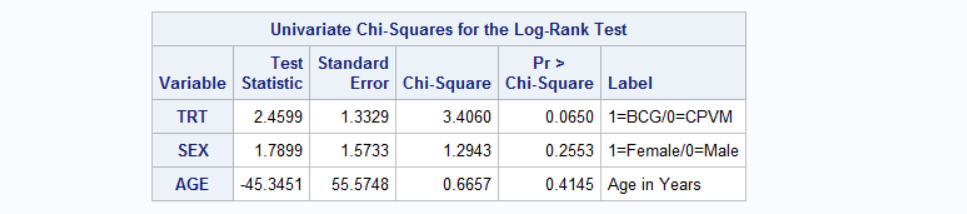
Ques No 1(d):

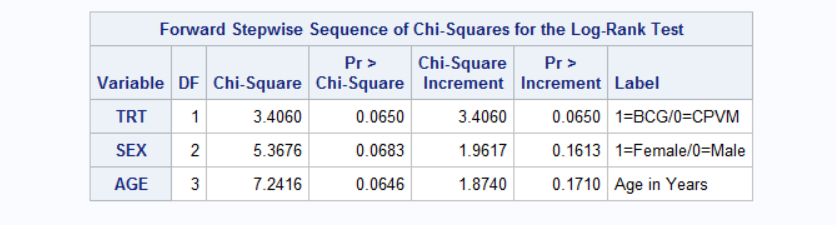
**Ques No 1 (c):**

At the beginning of the forward selection procedure, I tested the effect of each individual covariate treatment, sex and age group. Based on the significance of the variable, I was adjusted other covariates. None of the covariates are statistically significant based on their P value. Finally, I tested the treatment variable adjusted for sex and age group. I did not find any statistically significant covariates that contribute to the time to survival.





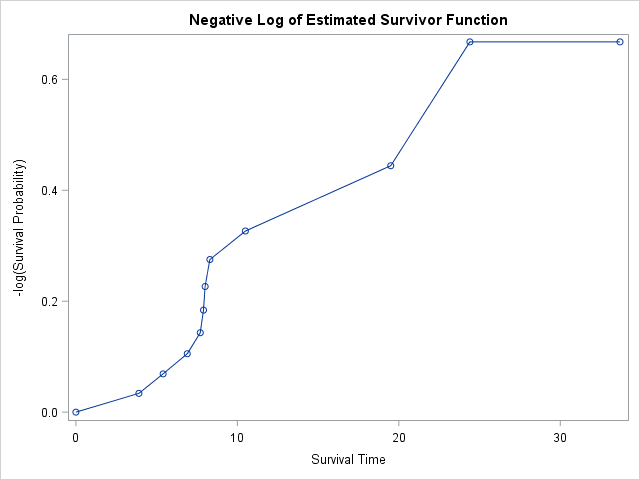
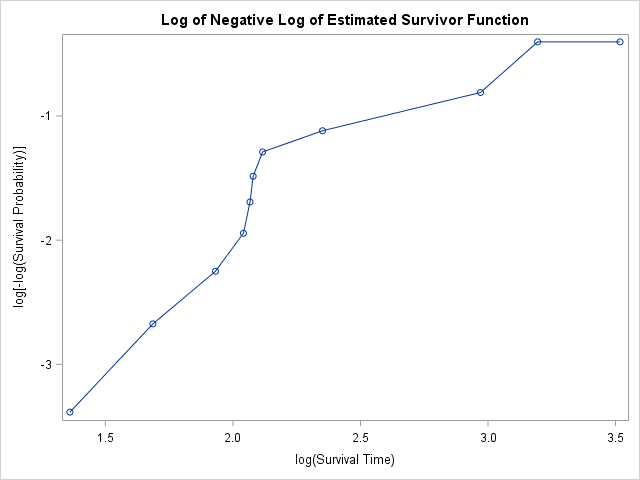




**Ques No 1(d):**

Log survival plot is not linear and going upward direction indicating that hazard increased with time. Thus, we can say that exponential model might not be appropriate for this dataset.

Similarly log-log survival plot is also going upward indicating that hazard is increasing with time. Since The curve did not look like linear with time, we can infer that Weibull model might not be a good ft for this dataset.



**Ques No 2:**

**Ques No 2(a):**

**Statistical Method:**

Under the parametric framework, we did not know the exact distribution of Melanoma data set. We need to fit various distributions to assess the assumption of distribution for this dataset. Hence, we examined with common parametric distribution such as exponential, Weibull, log normal, Logistic, Log-logistic, Gompertz, Normal, generalized gamma distribution etc. To find the best fitted model, we cross valided our fitted distribution with various fit statistics criterion. We also examine our models with visual inspection of log survival and log-log survival plot. We evaluated the AIC, AICc, BIC values and selected the distribution that produce the smallest value. Furthermore, visual inspection was performed to assess the fitted criteria for exponential, Weibull distribution. For the nested distribution, we compared their log likelihood value to find the best model. All the statistical procedure was performed with SAS version 9.4 and with the procedure PROC LIFEREG and PROC NLMIXED.

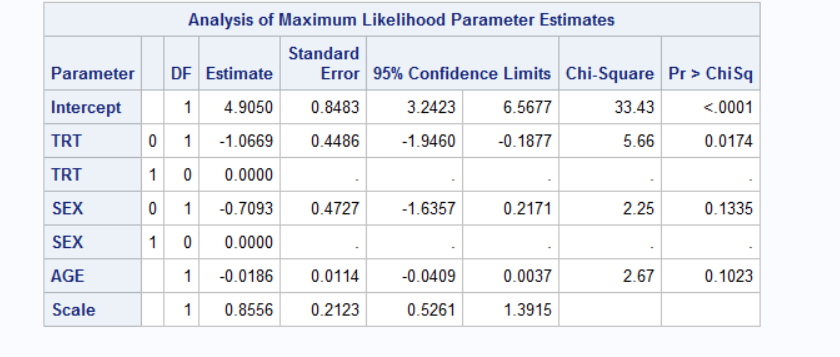
**Results:**

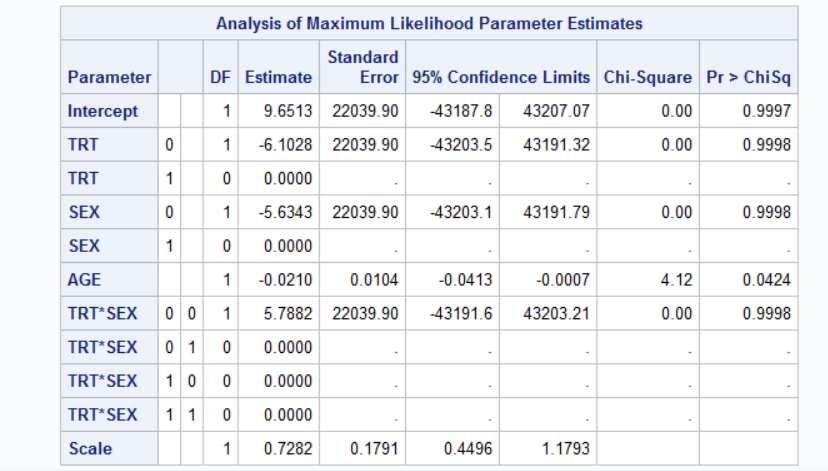
In the previous question, we plotted the log survival and log-log survival plot which indicated that exponential and Weibull distribution is not a goof fit for the melanoma dataset. Moreover, AIC, AICc and BIC values also indicated that those distribution might not be a good fit for this dataset. Similarly, based on the fit statistics values, we can also exclude the logistic, normal and Gompertz distributions.

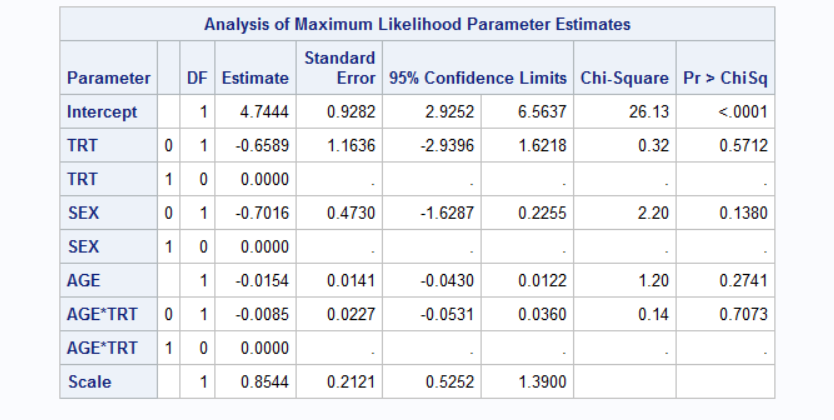
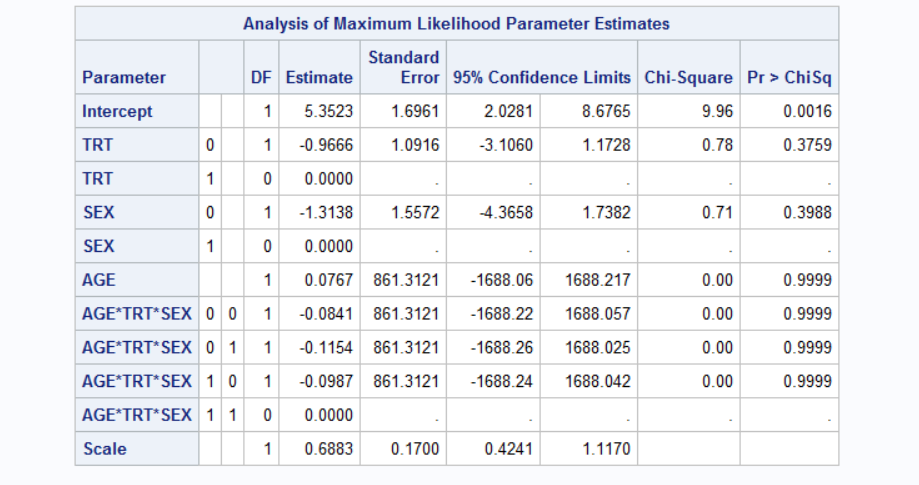
We can observe that log normal and generalized gamma distribution can be a good candidate for this dataset. Since log normal distribution is nested within the generalized gamma distribution, we can perform a Likelihood ratio test to compare the “best” model. Under null hypothesis, we assumed that simpler log normal model be a good fit for this dataset and under alternative hypothesis, generalized distribution be good fit for the dataset. Lognormal distribution has likelihood value is -21.015 and generalized gamma has log likelihood value is -19.39. Under null, the chi square test statistics is 3.24 with 1 df and the associated P value is .0718> 0.05 . Thus, we fail to reject the null hypothesis in favor of alternative hypothesis and concluded that log normal distribution is a good fit for our dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Distribution | -2 lL | AIC | AICc | BIC |
| Exponential | 46.360 | 54.360 | 55.960 | 59.965 |
| Weibull | 43.474 | 53.474 | 55.974 | 60.480 |
| Log normal | **42.030** | **52.030** | **54.530** | **59.036** |
| Logistic | 93.842 | 103.842 | 106.342 | 110.848 |
| Log logistic | 43.144 | 53.144 | 55.644 | 60.150 |
| Gompertz | 143.5 | 153.5 | 156 | 160.5 |
| Normal | 92.624 | 102.624 | 105.124 | 109.630 |
| Generalized Gamma | 38.790 | 50.790 | 54.442 | 59.197 |

After Analyzing the AFT lognormal model, we investigated the parametric maximum likelihood estimation and the associated P value. Our analysis is suggested that treatment has significant effect on survival (P value = 0.0174). But Sex and age covariates did not have any statistically significant effect on survival time. We also investigate whether treatment and sex or treatment and age had any interaction effect on survival. We did not observe any significant interaction from our analysis fitted with log normal distribution. May be need more interpretation

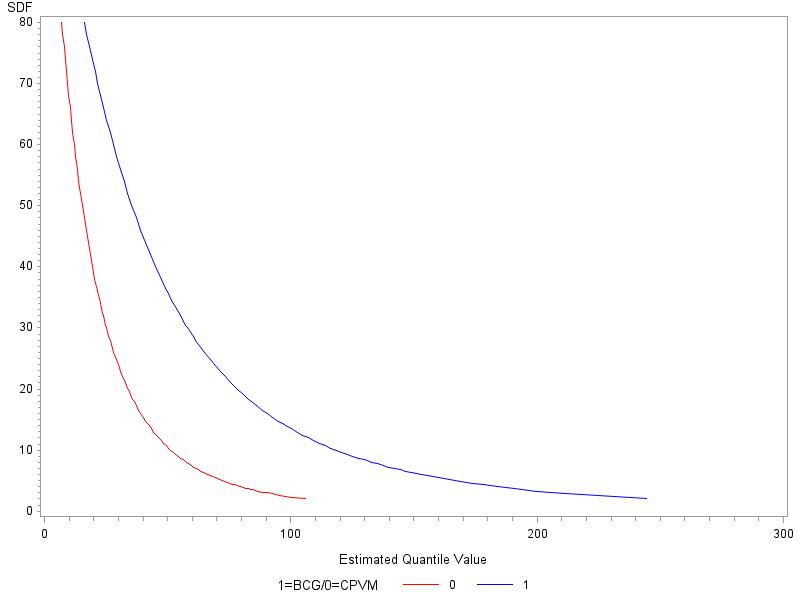






**Ques No 2(b):**

Since Age and Sex variable was not significant in our dataset, we did not include those variables in our final model. Below are the estimated survival curves for females 40 years of age receiving two treatment on the same plot.



**Ques No 4(a):**

To obtain parameter estimates for a parametric survival model using the likelihood function, we usually use the full information. In this Addiction study dataset, we have 150 events and 88 censored observations. Thus, all 238 observations were used to estimate parameters using MLE.

On the other hand, Cox regression model only used partial likelihood information. Thus, only 150 events were used to construct the likelihood, hence it is called partial likelihood estimation.

**Ques No 4(b):**

Cox proportional hazard model can be written as:

where, x1 = Clinic

x2 = prison and x3 = dose

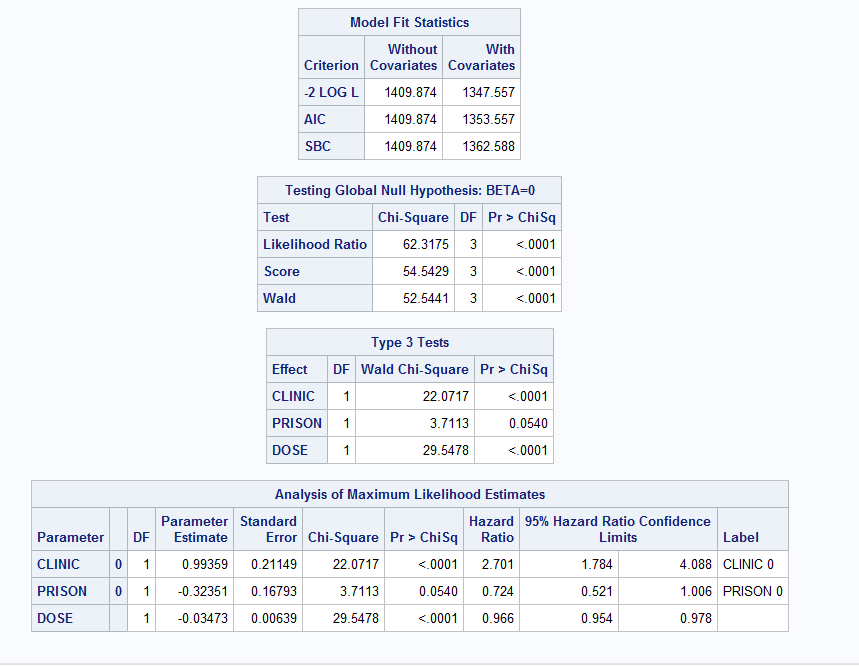
Also, represent the baseline hazard function.

Using Partial likelihood estimation Cox PH model for this dataset is:

Since, Prison variable is not significant, I did not include it in my model building process and the baseline hazard is when clinic = 1.

**Ques No 4(c):**

We found a statistically significant different in the amount of time spent in the two clinics adjusting for prison and dose variable. Our best estimate of the hazard ratio for amount of time spent in clinic 1 vs. clinic 2 is: HR = 2.701., w/ 95% C.I. (1.784,4.088).



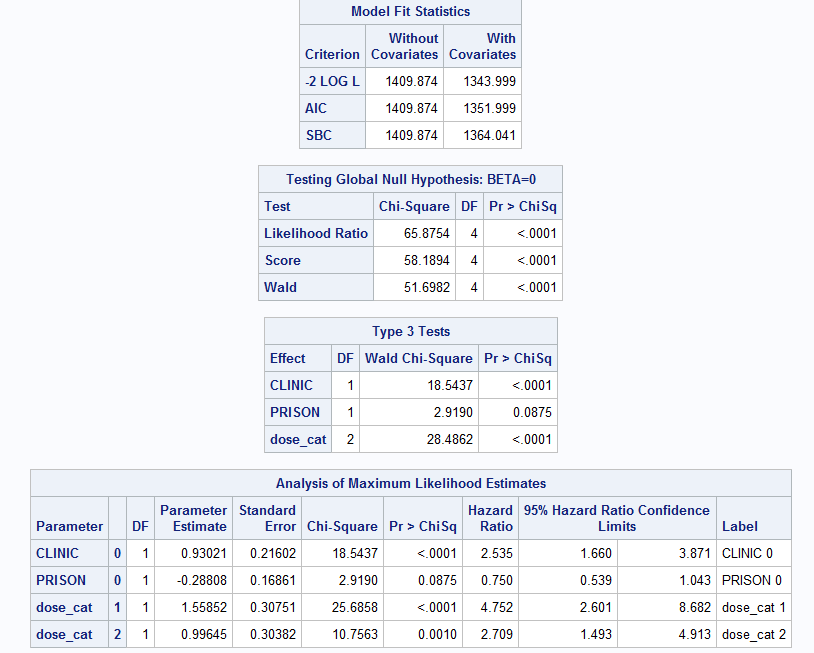
**Ques No 4 (d):**

When we use dose as a quantitative variable, the estimate is

Which implies that the adjusted hazard decreases by approximately 3.4% for each unit increase of dose at the time spent in clinic.

On the contrary, when we consider dose as a ordinal variable, the interpretation is changed. Considering dose >=80 as a reference group, we can say that, estimated value for the dose <60 group is exp(-1.558) = 4.75. which implies that, at any given point in time, the adjusted hazard for those receiving dose 1 is approximately 5 times the adjusted hazard for those who receiving dose 3.

Similarly, estimated value for the dose [60-80) group is exp(-.996) = 2.707. which implies that, at any given point in time, the adjusted hazard for those receiving dose 12 is approximately 3 times the adjusted hazard for those who receiving dose 3.



**Ques No 4 (e):**

The resulting plot are shown for the addiction study dataset. The martingale residuals plots showed an increasing pattern and skewed to the right.

Also, deviance residual plot suggested a distinguishable pattern indicated that Cox model might not be appropriate for this dataset.

