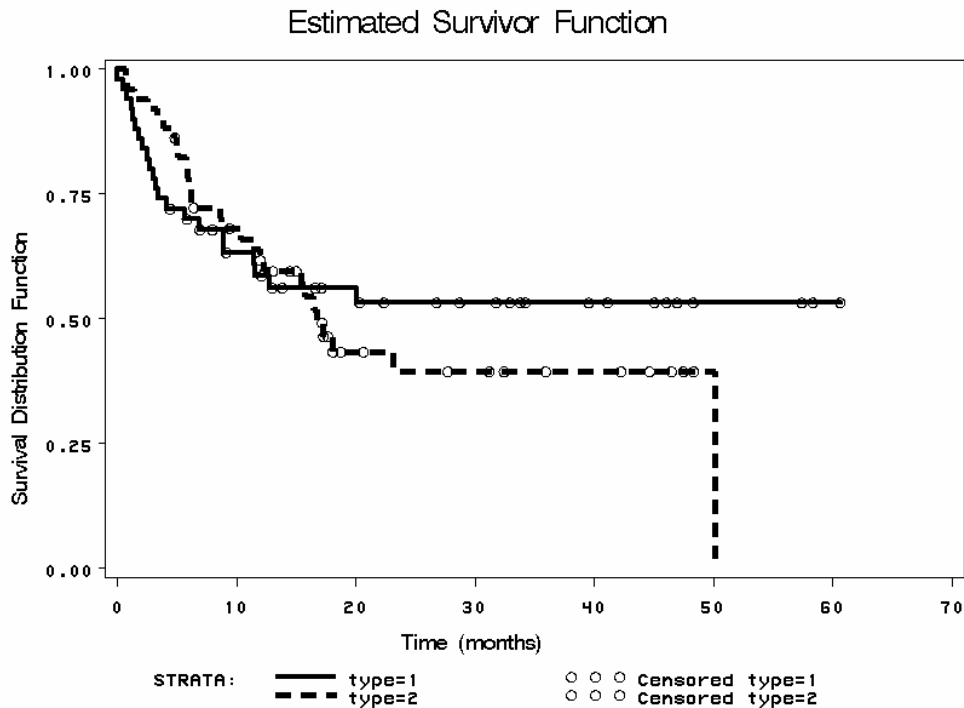


1. Since  $S(t)=1-F(t)$  and  $h(t) = \frac{f(t)}{S(t)} = \frac{\partial F(t)/\partial t}{S(t)} = \frac{-\partial S(t)/\partial t}{S(t)} = \frac{-\partial \log(S(t))}{\partial t}$ . It follows that

$$\log(S(t)) = \int_0^t -h(u)\partial u = \int_0^t -h_0(u)g(x,\beta)\partial u = g(x,\beta)\int_0^t -h_0(u)\partial u = g(x,\beta)\log(S_0(t))$$

Consequently,  $S(t) = [S_0(t)]^{g(x,\beta)}$ .

2. Analysis of leukemia-free survival data (in months) from a sample of 101 patients with advanced acute myelogenous leukemia. Fifty-one patients received an autologous (auto) bone marrow transplant in which, after high doses of chemotherapy, their own bone marrow was reinfused to replace their destroyed immune system. Fifty patients received an allogeneic (allo) bone marrow transplant using marrow from an HLA (Histocompatibility Leukocyte Antigen) matched sibling. An event time is recorded if either the patient dies or the patient exhibits signs of leukemia, whichever occurs first.
- A. Plot of the Kaplan-Meier estimators of survivor functions for allogeneic and autologous transplant patients.



- B Since there are no failures between 12.796 and 20.066 months, use the estimate of the survivor function at 12.796 month to estimate the probability of leukemia free survival for at least 18 months with the allo transplant.

$$\hat{S}(18) = \hat{S}(12.796) = 0.5617 \text{ with standard error } s_{\hat{S}(18)} = 0.0726$$

- (ii) Because of the high degree of censoring, we cannot obtain an approximate 95% confidence interval for the median leukemia free survival time for allogeneic transplant patients. Since the estimated survivor function is larger than 0.5 at the last observed failure time (20.066 months), the estimated median must be beyond 20.066 months but the Kaplan-Meier estimator provides no information about the survival time distribution beyond 20.066, other than an estimated probability of 0.5321 of surviving beyond 20.066 months. There is no information about how far beyond 20.066 to place the median. This is a limitation of non-parametric estimation procedures.

- B (i) Since there are no failures between 17.237 and 18.092 months, use the estimate of the survivor function at 17.237 month to estimate the probability of leukemia free survival for at least 18 months with the auto transplant.

$$\hat{S}(18) = \hat{S}(17.237) = 0.4643 \text{ with standard error } s_{\hat{S}(18)} = 0.0754$$

- (ii) The median leukemia-free survival time is estimated as the smallest time for which estimated survivor function is less than 0.5. From the following table of survivor function estimates

time	Survival	Failure	Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	51
0.6580	0.9804	0.0196	0.0194	1	50
0.8220	0.9608	0.0392	0.0272	2	49
1.4140	0.9412	0.0588	0.0329	3	48
2.5000	0.9216	0.0784	0.0376	4	47
3.3220	0.9020	0.0980	0.0416	5	46
3.8160	0.8824	0.1176	0.0451	6	45
4.7370	0.8627	0.1373	0.0482	7	44
4.8360*	.	.	.	7	43
4.9340	0.8427	0.1573	0.0511	8	42
5.0330	0.8226	0.1774	0.0537	9	41
5.7570	0.8026	0.1974	0.0560	10	40
5.8550	0.7825	0.2175	0.0581	11	39
5.9870	0.7624	0.2376	0.0599	12	38
6.1510	0.7424	0.2576	0.0616	13	37
6.2170	0.7223	0.2777	0.0631	14	36
6.4470*	.	.	.	14	35
8.6510	0.7017	0.2983	0.0646	15	34
8.7170	0.6810	0.3190	0.0659	16	33

9.4410*	.	.	.	16	32
10.3290	0.6597	0.3403	0.0672	17	31
11.4800	0.6385	0.3615	0.0683	18	30
12.0070	0.6172	0.3828	0.0693	19	29
12.0070*	.	.	.	19	28
12.2370	0.5951	0.4049	0.0702	20	27
12.4010*	.	.	.	20	26
13.0590*	.	.	.	20	25
14.4740*	.	.	.	20	24
15.0000*	.	.	.	20	23
15.4610	0.5693	0.4307	0.0718	21	22
15.7570	0.5434	0.4566	0.0730	22	21
16.4800	0.5175	0.4825	0.0740	23	20
16.7110	0.4916	0.5084	0.0747	24	19
17.2040*	.	.	.	24	18
17.2370	0.4643	0.5357	0.0754	25	17
17.3030*	.	.	.	25	16
17.6640*	.	.	.	25	15
18.0920	0.4334	0.5666	0.0764	26	14
18.0920*	.	.	.	26	13
18.7500*	.	.	.	26	12
20.6250*	.	.	.	26	11
23.1580	0.3940	0.6060	0.0790	27	10
27.7300*	.	.	.	27	9
31.1840*	.	.	.	27	8
32.4340*	.	.	.	27	7
35.9210*	.	.	.	27	6
42.2370*	.	.	.	27	5
44.6380*	.	.	.	27	4
46.4800*	.	.	.	27	3
47.4670*	.	.	.	27	2
48.3220*	.	.	.	27	1
50.0860	0	1.0000	0	28	0

the smallest time for which the estimated survivor function is less than 0.5 is 16.711 months. Thus the median leukemia-free survival time for auto transplant patients is estimated at 16.711 months. An approximate 95% confidence interval for the median leukemia-free survival time produced by PROC LIFETEST in SAS for allogeneic transplant patients is ( 12.007, 50.086) months.

- C. Since the data were not sufficient to estimate the median leukemia-free survival time for the allogeneic transplant patients, the data are also not sufficient to construct a 95% confidence interval for the difference in the median leukemia-free survival times for allogeneic and auto transplant patients.
- D. Results of the log-rank and Wilcoxon tests the null hypothesis that the survivor curves are the same for allogeneic and auto transplant patients are

Test	Chi-Square	DF	Chi-Square
Log-Rank	0.3816	1	0.5368
Wilcoxon	0.0969	1	0.7556

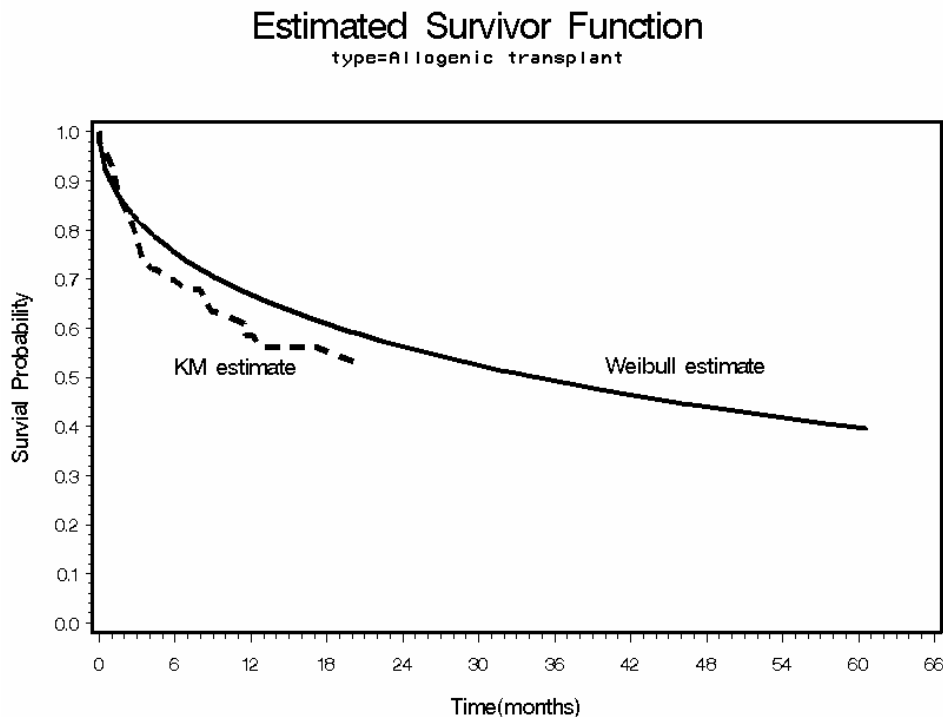
Neither test indicates that the estimated survivor curves are sufficiently different to reject the homogeneity null hypothesis at the .05 level of significance.

E. Since allogeneic transplant patients tend to have more complications early in the recovery process, the primary interest is the comparison of leukemia free survival rates among long-term survivors. Using Fleming and Harrington weights with  $p=0$  and  $q=1$  results in weights  $W(t_i) = 1 - \hat{S}(t_{i-1})$ ,  $i=1,2,\dots,r$ , that give more relative weight to deviations in the survivor curves at later failure times than either the log-rank test, which gives the same weight to differences at any failure time, or the Wilcoxon test, which gives more weight to differences at early failure times because its weights are proportional to the number of at risk subjects at each failure time. The Fleming and Harrington weights with  $p=0$  and  $q=1$  produce a chi-square test statistic with value 4.20 and  $p\text{-value}=0.0404$ . This test result is sufficient to reject the null hypothesis of homogeneous survivor functions at the 0.05 level of significance. It appears that allogeneic transplants provide higher probabilities for longer term survival.

3. A. A Weibull model with survivor function  $S(t) = \exp(-\theta t^\alpha)$  was fit the leukemia free survival times for the allogeneic transplant patients in problem 1. The maximum likelihood estimates of the parameters are  $\hat{\theta} = 0.112$  and  $\hat{\alpha} = 0.515$ , with standard errors 0.0405 and 0.0974, respectively. The estimate of the covariance matrix for the large sample normal distribution for these estimates is

$$V = \begin{bmatrix} 0.0016382 & -0.003184 \\ -0.003184 & 0.0094951 \end{bmatrix}$$

The following plot shows the estimated Weibull and Kaplan-Meier survivor curves.



This plot suggests that the Weibull model may not be able to completely account for an initial sharp decrease in the hazard that is reflected in the Kaplan-Meier estimate of the survivor curve.

(i) The maximum likelihood estimate of the probability of leukemia free survival for at least 18 months is  $\hat{S}(18) = \exp(-\hat{\theta}(18)^{\hat{\alpha}}) = \exp(-(.112)(18)^{.515}) = 0.609$ .

Using the delta method the large sample variance of  $\hat{S}(t)$  is

$\text{var}(\hat{S}(t)) = D(t) V D(t)^T$  where  $V$  is the large sample covariance matrix for

$(\hat{\theta}, \hat{\alpha})$  and  $D(t) = \begin{pmatrix} \frac{\partial S(t)}{\partial \theta} & \frac{\partial S(t)}{\partial \alpha} \end{pmatrix} = \begin{pmatrix} -t^\alpha S(t) & -\theta \log(t) t^\alpha S(t) \end{pmatrix}$ . Evaluating

everything at  $(\hat{\theta}, \hat{\alpha}) = (0.112, 0.515)$  yields an estimated variance of 0.00416 and a standard error of 0.0645.

(ii) The maximum likelihood estimate of the median leukemia-free survival time

is  $\hat{M}_{\text{allo}} = \left[ \frac{\log(2)}{\hat{\theta}} \right]^{1/\hat{\alpha}} = \left[ \frac{\log(2)}{0.112} \right]^{1/0.515} = 34.38$ . The large sample variance is

obtained from the delta method as  $\text{var}(\hat{M}_{\text{allo}}) = D(t) V D(t)^T$  where  $V$  is the large sample covariance matrix for  $(\hat{\theta}, \hat{\alpha})$  and

$D(t) = \begin{pmatrix} \frac{\partial M_{\text{allo}}}{\partial \theta} & \frac{\partial M_{\text{allo}}}{\partial \alpha} \end{pmatrix} = \begin{pmatrix} -\frac{M_{\text{allo}}}{\alpha \theta} & -\frac{M_{\text{allo}}}{\alpha^2} \log\left(\frac{\log(2)}{\theta}\right) \end{pmatrix}$ . Evaluating everything

at  $(\hat{\theta}, \hat{\alpha}) = (0.112, 0.515)$  yields an estimated variance of 215.40 and a standard error of 14.68. Then an approximate 95% confidence interval is

$\hat{M}_{\text{allo}} \pm (1.96)(14.68) \Rightarrow (5.61, 63.15)$ .

On pages 166-167, Collet derives an approximate confidence interval using the large sample normal approximation to the logarithm of the estimated median

$$\log(\hat{M}_{\text{allo}}) = \frac{1}{\hat{\alpha}} \left[ \log(\log(2)) - \log(\hat{\theta}) \right] = \frac{1}{.515} \left[ \log(\log(2)) - \log(.112) \right] = 3.538.$$

Using the delta method, the asymptotic standard error of the logarithm of the estimated median is  $s_{\log(\hat{M}_{\text{allo}})} = \sqrt{D(t) V D(t)^T}$  where  $V$  is the large sample covariance matrix for  $(\hat{\theta}, \hat{\alpha})$  and

$$D(t) = \begin{pmatrix} \frac{\partial \log(M_{\text{allo}})}{\partial \theta} & \frac{\partial \log(M_{\text{allo}})}{\partial \alpha} \end{pmatrix} = \begin{pmatrix} -\frac{1}{\alpha \theta} & -\frac{1}{\alpha^2} \log\left(\frac{\log(2)}{\theta}\right) \end{pmatrix}$$

is evaluated at  $(\hat{\theta}, \hat{\alpha}) = (0.112, 0.515)$ . Then  $s_{\log(\hat{M}_{\text{allo}})} = 0.42665$  and an approximate 95% confidence interval for  $\log(M_{\text{allo}})$  is

$$\log(\hat{M}_{\text{allo}}) \pm (1.96)s_{\log(\hat{M}_{\text{allo}})} \Rightarrow 3.538 \pm (1.96)(0.42665) \Rightarrow (2.7018, 4.3743).$$

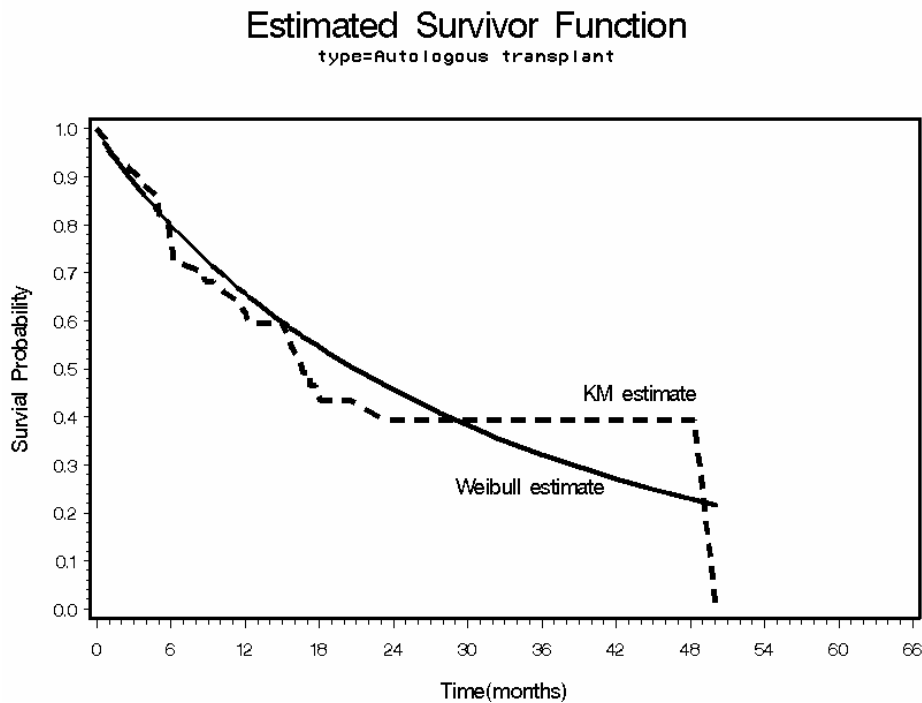
Then, an approximate 95% confidence interval for the median leukemia-free survival time for patients treated with the allogeneic transplant is

$$(\exp(2.7018), \exp(4.3743)) \Rightarrow (14.91, 79.38).$$

3. B. A Weibull model with survivor function  $S(t) = \exp(-\theta t^\alpha)$  was fit the leukemia free survival times for the auto transplant patients in problem 1. The maximum likelihood estimates of the parameters are  $\hat{\theta} = 0.0444$  and  $\hat{\alpha} = 0.9045$ , with standard errors 0.02123 and 0.14327, respectively. The estimate of the covariance matrix for the large sample normal distribution for these estimates is

$$V = \begin{bmatrix} 0.0004507 & -0.002794 \\ -0.002794 & 0.0205269 \end{bmatrix}$$

The following plot shows the estimated Weibull and Kaplan-Meier survivor curves.



This plot suggests that the Weibull model may be appropriate for these data. The estimate of the Weibull survivor function matches the Kaplan-Meier quite well. The mismatch with the Kaplan-Meier estimate after 24 months should not be considered too seriously because it is based on the reaction of the Kaplan-Meier estimate to a single failure around 50 months.

- (i) The maximum likelihood estimate of the probability of leukemia free survival for at least 18 months is  $\hat{S}(18) = \exp(-\hat{\theta}(18)^{\hat{\alpha}}) = \exp(-(.0444)(18)^{.9045}) = 0.545$ .

Using the delta method the large sample variance of  $\hat{S}(t)$  is  $\text{var}(\hat{S}(t)) = D(t) V D(t)^T$  where  $V$  is the large sample covariance matrix for  $(\hat{\theta}, \hat{\alpha})$  and  $D(t) = \begin{pmatrix} \frac{\partial S(t)}{\partial \theta} & \frac{\partial S(t)}{\partial \alpha} \end{pmatrix} = \begin{pmatrix} -t^\alpha S(t) & -\theta \log(t) t^\alpha S(t) \end{pmatrix}$ . Evaluating everything at  $(\hat{\theta}, \hat{\alpha}) = (0.0444, 0.9045)$  yields an estimated variance of 0.0039744 and a standard error  $s_{\hat{M}_{\text{auto}}} = 0.06304$ .

(ii) The maximum likelihood estimate of the median leukemia-free survival time is  $\hat{M}_{\text{auto}} = \left[ \frac{\log(2)}{\hat{\theta}} \right]^{1/\hat{\alpha}} = \left[ \frac{\log(2)}{0.0444} \right]^{1/0.9045} = 20.83$ . The large sample variance is obtained from the delta method as  $\text{var}(\hat{M}_{\text{auto}}) = D(t) V D(t)^T$  where  $V$  is the large sample covariance matrix for  $(\hat{\theta}, \hat{\alpha})$  and

$D(t) = \begin{pmatrix} \frac{\partial M_{\text{auto}}}{\partial \theta} & \frac{\partial M_{\text{auto}}}{\partial \alpha} \end{pmatrix} = \begin{pmatrix} -\frac{M_{\text{auto}}}{\alpha \theta} & -\frac{M_{\text{auto}}}{\alpha^2} \log\left(\frac{\log(2)}{\alpha}\right) \end{pmatrix}$ . Evaluating everything at  $(\hat{\theta}, \hat{\alpha}) = (0.0444, 0.9045)$  yields an estimated variance of 19.0235 and a standard error  $s_{\hat{M}_{\text{auto}}} = 4.362$ . Then an approximate 95% confidence interval is  $\hat{M}_{\text{auto}} \pm (1.96)(4.362) \Rightarrow (12.28, 29.38)$ .

Alternatively, using the large sample normal approximation for the distribution of the estimated median,  $s_{\log(\hat{M}_{\text{auto}})} = 0.209$  and an approximate 95% confidence interval for  $\log(M_{\text{auto}})$  is

$$\log(\hat{M}_{\text{auto}}) \pm (1.96) s_{\log(\hat{M}_{\text{auto}})} \Rightarrow \log(20.8709) \pm (1.96)(0.209) \Rightarrow (2.629, 3.448)$$

Then, an approximate 95% confidence interval for the median leukemia-free survival time for patients treated with the allogeneic transplant is

$$(\exp(2.629), \exp(3.448)) \Rightarrow (13.86, 31.44).$$

- C. Assuming that the allogeneic transplant patients respond independently of the auto transplant patients,  $\hat{M}_{\text{auto}}$  is independent of  $\hat{M}_{\text{allo}}$  and an approximate 95% confidence interval for the difference in the medium leukemia-free survival times for the allo and auto transplant patients is

$$\begin{aligned} \hat{M}_{\text{allo}} - \hat{M}_{\text{auto}} \pm (1.96) \sqrt{s_{\hat{M}_{\text{allo}}}^2 + s_{\hat{M}_{\text{auto}}}^2} &\Rightarrow (34.38 - 20.83) \pm (1.96) \sqrt{(14.677)^2 + (4.362)^2} \\ &\Rightarrow (13.55) \pm (1.96) (15.31) \Rightarrow (-16.46, 43.56) \end{aligned}$$

Since this confidence interval does not contain zero, the data do not provide enough evidence to reject the null hypothesis of equal median leukemia-free survival times

for allogeneic and autologous transplant patients. Because of the relatively small sample size and the high level of censoring, the median survival times are not very well estimated from these data.

- D. A proportional hazards Weibull model was fit to these data using the following code for the LIFEREG procedure in SAS. This model incorporates the type of transplant as a covariate coded as type=0 for allogeneic transplant patients and type=1 for autologous transplant patients. The formulas for the survivor functions are

$$S(t) = \begin{cases} \exp(-[\theta + \beta(0)]t^\alpha) & \text{for allogeneic transplant patients} \\ \exp(-[\theta + \beta(1)]t^\alpha) & \text{for autologous transplant patients} \end{cases}$$

This differs from fitting the two separate models in parts A and B, because the  $\alpha$  parameter is the same for both the allogeneic and autologous transplant patients. The  $\alpha$  parameter controls the increase or decrease of the hazard function across time and it must be the same for both Weibull survivor functions to preserve the proportional hazards property. The null hypothesis that the survivor curves are the same for allogeneic and autologous transplant patients can be tested by testing  $H_0 : \beta = 0$  versus  $H_a : \beta \neq 0$ . The relevant output from the LIFEREG procedure in SAS is

**The LIFEREG Procedure**

**Analysis of Parameter Estimates**

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.3389	0.6966	2.9736	5.7042	38.80	<.0001
type	1	-0.3754	0.4195	-1.1976	0.4467	0.80	0.3708
Scale	1	1.4704	0.1809	1.1554	1.8713		
Weibull Shape	1	0.6801	0.0837	0.5344	0.8655		

**Estimated Covariance Matrix**

	Intercept	type	Scale
Intercept	0.485269	-0.276384	0.022122
type	-0.276384	0.175949	-0.003875
Scale	0.022122	-0.003875	0.032720

SAS parameterizes the model in a different way and the values of the parameters in our model are obtained as

$$\hat{\alpha} = \frac{1}{\text{scale}} \quad \hat{\theta} = \exp\left(\frac{-\text{intercept}}{\text{scale}}\right) \quad \hat{\beta} = \exp\left(\frac{-(\text{coefficient for type})}{\text{scale}}\right).$$

Nevertheless, testing

$$H_0 : \text{coefficient for type} = 0 \quad \text{versus} \quad H_a : \text{coefficient for type} \neq 0$$

is equivalent to testing  $H_0 : \beta = 0$  versus  $H_a : \beta \neq 0$ . This test can be done as a Wald

test by comparing  $\left(\frac{0.3754}{0.4195}\right)^2 = 0.80$  to the percentiles of a central chi-square

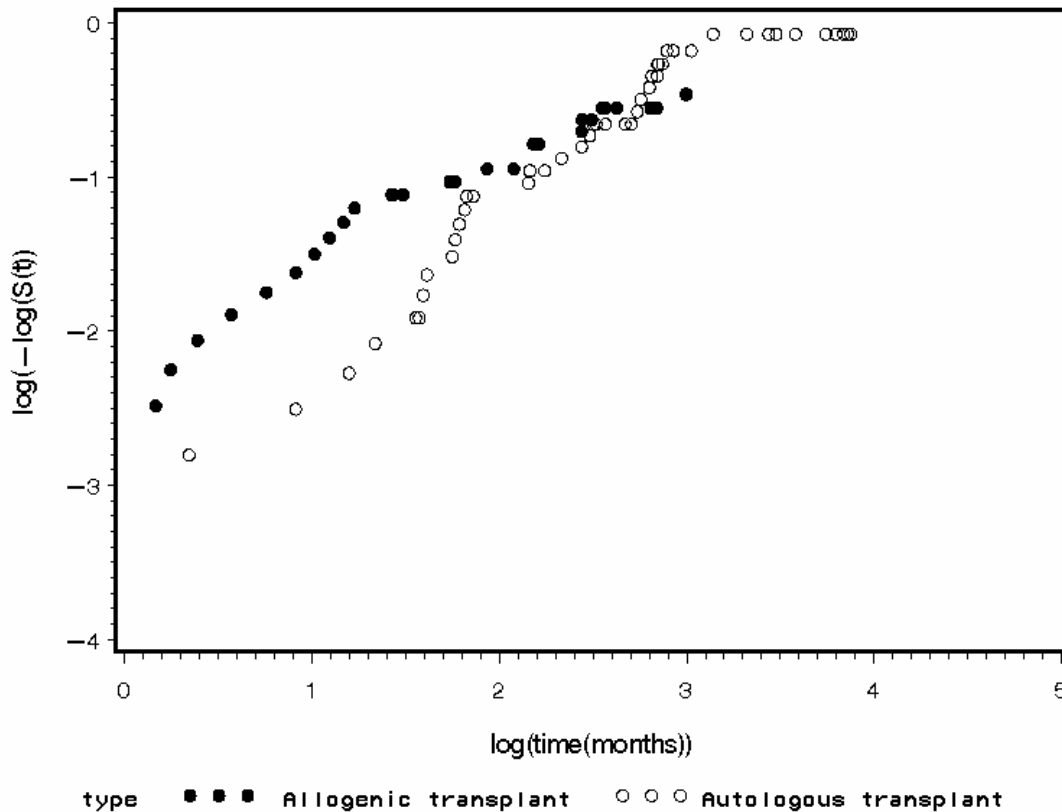
distribution to get p-value of 0.3708. This test does not provide enough evidence to reject the hypothesis of homogeneous leukemia-free survival time distributions for allergenic and autologous transplant patients.

E. Potential advantages of using Weibull models over Kaplan-Meier estimation are that (1) percentiles beyond the last observed failure time can be estimated (extrapolation is possible because you have a formula for the survivor function), (2) estimates of survival probabilities and percentiles are obtained with smaller standard errors (specifying the form of the survivor function introduces some additional information that reduces variability in the estimation of the survivor function). A potential advantage of using a nonparametric approach like Kaplan-Meier estimation is that it provides consistent estimators for the survivor function and the percentiles of the survival distribution, regardless of the true form of the survivor function.

Consequently, you do not have to check the fit of a family of models (a little less work and an automatic procedure). The price you pay is increased variation in estimates of survival probabilities and percentiles. Using a Weibull model will produce biased estimates of the survivor function and the percentiles of the survival distribution if the Weibull family of models does not contain the true model. If the Weibull family of models provides a reasonable approximation to the true survival time distribution, biases will be small and the reduction in variability may more than compensate for a little bias, especially for small samples. If goodness-of-fit checks indicate that a parametric family of models provides a reasonable approximation to the true model, it is usually better to use the parametric approach than to use a non-parametric approach.

4. Using the Kaplan-Meier estimators of the survivor functions from problem 1, the following is a plot of  $\log(-\log(\hat{S}(t)))$  against  $\log(t)$  for the allogeneic and auto transplant patients.

## Weibull Goodness-of-fit Plot



- (i) For each type of transplant, the plotted points fall nearly along a straight line. (although you have no experience with how much variability about a straight line to expect with these sample sizes). This indicates that Weibull distributions may provide reasonable approximations to the survival time distributions for the two types of transplant. A slope of one would suggest that the exponential distribution may be appropriate. The slope appears to be close to one for the autologous transplant patients but closer to 0.5 for the allogenic transplant patients.
- (ii) The points for the two types of transplant do not appear to fall along parallel lines. As noted above, the nearly linear trend for the allogenic transplant patients appears to have a smaller slope than the linear trend for the autologous transplant patients. This plot suggests that a proportional hazards model may not be appropriate.

Assuming that leukemia-free survival times follow Weibull distributions for both allogenic and autologous transplant patients, one could perform a likelihood ratio test by comparing

$$\begin{aligned}
& -2[\log(\text{likelihood for the Weibull proportional hazards model in part C}) \\
& \quad - \{\log(\text{likelihood for the Weibull model in part A}) \\
& \quad \quad + \log(\text{likelihood for the Weibull model in part B})\}] \\
& = -2[(-143.8442) - (-72.8472 - 68.3583)] = 5.2774
\end{aligned}$$

with the percentiles of a chi-square distribution with one degree of freedom to get a p-value of 0.0216. This is sufficient evidence to reject the proportional hazards hypothesis in this context. Of course, this test could be misleading if the Weibull assumption is inappropriate. For these data, however, the goodness-of fit plots support the Weibull assumption for both groups.

5. A. Let  $\pi$  denote the proportion of the population of smokers who would be successful at permanently quitting smoking with this program. Recidivism times for subjects that are not cured follow a Weibull distribution with survivor function  $S(t) = \exp(-\theta t^\alpha)$ . Then the probability that a randomly selected member of the population of smokers who would participate in the stop smoking program does not return to smoking before time  $t$  is

$$\begin{aligned}
P(\text{smoke free beyond time } t) &= P(\text{smoke free beyond time } t \mid \text{cured})P(\text{cured}) \\
& \quad + P(\text{smoke free beyond time } t \mid \text{recidivist})P(\text{recidivist}) \\
& = (1)(\pi) + (1-\pi)\exp(-\theta t^\alpha) \\
& = \pi + (1-\pi)\exp(-\theta t^\alpha) \\
& = S(t)
\end{aligned}$$

B. The density function is

$$f(t) = \frac{-\partial S(t)}{\partial t} = (1-\pi)\theta\alpha t^{\alpha-1} \exp(-\theta t^\alpha)$$

For data  $(t_i, d_i)$   $i=1,2,\dots,n$ , where  $t_i$  is an observed time and  $d_i$  is a right censoring indicator defined as

$$d_i = \begin{cases} 1 & \text{observed recidivism time} \\ 0 & \text{observed right censored time} \end{cases}$$

the joint likelihood (assuming independent observations) is

$$L(\theta, \alpha, \pi \mid \text{data}) = \prod_{i=1}^n \left[ (1-\pi)\theta\alpha t_i^{\alpha-1} \exp(-\theta t_i^\alpha) \right]^{d_i} \left[ \pi + (1-\pi)\exp(-\theta t_i^\alpha) \right]^{1-d_i}$$

- C. Let  $\mathbf{X}_i$  denote a set of covariates for the  $i$ -th individual in the study. Covariates could be incorporated into the model as

$$S(t) = \pi + (1 - \pi) \exp(-(\theta + \beta^T \mathbf{X}_i) t^\alpha)$$

Then the joint likelihood is

$$L(\theta, \alpha, \pi | \text{data}) = \prod_{i=1}^n \left[ (1 - \pi) (\theta + \beta^T \mathbf{X}_i) \alpha t_i^{\alpha-1} \exp(-(\theta + \beta^T \mathbf{X}_i) t_i^\alpha) \right]^{d_i} \left[ \pi + (1 - \pi) \exp(-(\theta + \beta^T \mathbf{X}_i) t_i^\alpha) \right]^{1-d_i}$$

This model would be difficult to work with, however, because it requires that  $\beta^T \mathbf{X}_i > 0$  for every possible set of covariate values. You could avoid this problem by replacing  $\beta^T \mathbf{X}_i$  with a positive quantity such as  $\exp(\beta^T \mathbf{X}_i)$ .

Alternatively, using a proportional hazards criterion for the recidivists, covariates could be incorporated into the model as

$$S(t) = \pi + (1 - \pi) \exp(-(\theta e^{\beta^T \mathbf{X}_i}) t^\alpha)$$

Then the joint likelihood is

$$L(\theta, \alpha, \pi | \text{data}) = \prod_{i=1}^n \left[ (1 - \pi) (\theta e^{\beta^T \mathbf{X}_i}) \alpha t_i^{\alpha-1} \exp(-(\theta e^{\beta^T \mathbf{X}_i}) t_i^\alpha) \right]^{d_i} \left[ \pi + (1 - \pi) \exp(-(\theta e^{\beta^T \mathbf{X}_i}) t_i^\alpha) \right]^{1-d_i}$$

Another possibility is to expand the time scale as in an accelerated failure time model. This can be done as

$$S(t) = \pi + (1 - \pi) \exp(-\theta (e^{\beta^T \mathbf{X}_i} t)^\alpha)$$

This is equivalent to the previous proportional hazards approach.

One could incorporate covariates into the recidivism rate  $1 - \pi_i$  by using a logit model

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \gamma^T \mathbf{Z}_i \quad \text{or} \quad \pi_i = \frac{\exp(\gamma^T \mathbf{Z}_i)}{1 + \exp(\gamma^T \mathbf{Z}_i)}.$$

The  $\mathbf{Z}_i$  covariates could include some of the  $\mathbf{X}_i$  covariates used in the conditional survival distribution for recidivists, but they would not have to be the same. Allowing  $1 - \pi_i$  to change with changes in covariates would create some numerical problems for maximizing the resulting log likelihood, because it would have some very flat ridges. You may not be able to allow  $1 - \pi_i$  to change with covariates in practice.

D. The proportion of the population of smokers who would be successful at permanently quitting smoking with this program could not be estimated with a nonparametric Kaplan –Meier approach. Since the Kaplan –Meier approach does not assume a parametric form for the survivor function, it cannot extrapolate beyond the last failure time in the data. Some cases that are censored at the last failure time in the data might later fail and there is no way to determine what proportion of cured cases among the cases censored at the last observed failure time. Non-parametric models are not very useful for extrapolation.

6. (a) The number of patients required for the log-rank test is 812, with 406 in each treatment group.
- (b) The number of patients required for the Wilcoxon test is 1132, with 566 in each treatment group. The Wilcoxon test requires more subjects to achieve the same power as the logrank test in this situation because the Wilcoxon test gives relatively more weight to differences in the survivor functions at early time points where they are similar and less weight to differences at later time points where the survivor curves exhibit larger differences.
- (c) By extending the follow-up time from 8 to 12 months (and adding a line at  $t = 16$  months with  $S(t) = .05$  for the placebo group and  $S(t) = .15$  for the CGD treated patients), required sample sizes for the log-rank and Wilcoxon tests are reduced to 656 and 1060, respectively.
- (d) Keeping the follow-up time of 8 months and changing the accrual time to 2 months (and changing the last line of each set of projected survival times from 12 to 10 months with survival probabilities of 0.20 and 0.30 for the placebo and CGD groups, respectively), required sample sizes for the log-rank and Wilcoxon tests are reduced to 848 and 1146, respectively.
- (e) Increasing the follow-up time would be the better option if a reduction from 812 to 656 subjects provided enough savings to more than offset the costs of following the patients for two more months and waiting for two additional months to get the results.