Survival Analysis Take-home Midterm

☞ Data introduction:

Kardaun (1983) reports data on 90 males diagnosed with cancer of the larynx during the period 1970V1978 at a Dutch hospital. Times recorded are the intervals (in years) between first treatment and either death or the end of the study (January 1, 1983). Also recorded are the patients age at the time of diagnosis, the year of diagnosis, and the stage of the patients cancer. The four stages of disease in the study were based on the T.N.M. (primary tumor (T), nodal involvement (N) and distant metastasis (M) grading) classification used by the American Joint Committee for Cancer Staging (1972). The four groups are Stage I, $T1N0M0$ with 33 patients; Stage II, $T2N0M0$ with 17 patients; Stage III, $T3N0M0$ and TxN1M0, with 27 patients; $x = 1, 2,$ or 3; and Stage IV, all other TNM combinations except TIS with 13 patients. The stages are ordered from least serious to most serious.

☞ I. Parametric model

1. `Fit a Weibull model to the data including only one variable, disease stage. Find the MLEs of` λ `and` α `and their standard errors. Plot the survival functions for the patients of all four stages in one figure.`

(i) ANOVA Table for $\hat{\mu}, \hat{\sigma}$, and $\hat{\gamma}_i$, i=1,2,3

		Parameter	Standard		
Variables	df	Estimates	Error	Chi-Square	p-Value
Intercept($\hat{\mu}$)	1	2.3691	0.2396		
Scale($\hat{\sigma}$)	1	0.8846	0.1082		
Stage II($\hat{\gamma}_1$)	1	-0.0868	0.4049	-0.214	8.30e-01
Stage III($\hat{\gamma}_2$)	1	-0.5566	0.3186	-1.747	8.06e-02
Stage IV($\hat{\gamma}_3$)	1	-1.5786	0.3632	-4.346	1.38e-05

The fitted model is

$$Y = \log X = \hat{\mu} + \hat{\gamma}^T Z + \hat{\sigma} W = 2.3691 - 0.0868 Z_1 - 0.5566 Z_2 - 1.5786 Z_3 + 0.8846 W$$

where $\gamma = (\gamma_1, \gamma_2, \gamma_3)$, $Z = (Z_1, Z_2, Z_3)$ and

$$\begin{aligned} Z_1 &= 1 \text{ if the patient is in stage II, 0 otherwise,} \\ Z_2 &= 1 \text{ if the patient is in stage III, 0 otherwise,} \\ Z_3 &= 1 \text{ if the patient is in stage IV, 0 otherwise.} \end{aligned}$$

The covariance matrix of $\hat{\mu}$, $\hat{\gamma}$ and $\log \hat{\sigma}$ is

$$\hat{\Sigma} = \begin{bmatrix} 0.0574 & -0.0519 & -0.0563 & -0.0589 & 0.0088 \\ -0.0519 & 0.1640 & 0.0520 & 0.0519 & 0.0004 \\ -0.0563 & 0.0520 & 0.1015 & 0.0575 & -0.0070 \\ -0.0589 & 0.0519 & 0.0575 & 0.1319 & -0.0114 \\ 0.0088 & 0.0004 & -0.0070 & -0.0114 & 0.0150 \end{bmatrix}.$$

Let $f_1(x, y, z, w, v) = (x, y, z, w, e^w)$ then $\dot{f}_1(x, y, z, w, v) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & e^w \end{bmatrix}$, that is,

$$\dot{f}_1(\hat{\mu}, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3, \log \hat{\sigma}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \hat{\sigma} \end{bmatrix}.$$

So the covariance matrix of $\hat{\mu}, \hat{\gamma}$ and $\hat{\sigma}$ is

$$\begin{aligned} \hat{\Sigma}_1 &= \dot{f}_1(\hat{\mu}, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3, \log \hat{\sigma}) \hat{\Sigma} \dot{f}_1(\hat{\mu}, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3, \log \hat{\sigma})^t \\ &= \begin{bmatrix} 0.0574 & -0.0519 & -0.0563 & -0.0589 & 0.0078 \\ -0.0519 & 0.1640 & 0.0520 & 0.0519 & 0.0004 \\ -0.0563 & 0.0520 & 0.1015 & 0.0575 & -0.0062 \\ -0.0589 & 0.0519 & 0.0575 & 0.1319 & -0.0100 \\ 0.0078 & 0.0004 & -0.0062 & -0.0100 & 0.0117 \end{bmatrix} \end{aligned}$$

So the ANOVA Table for $\hat{\lambda}, \hat{\beta}_i$, $i = 1, 2, 3$, and $\hat{\alpha}$

Variables	df	Parameter Estimates	Standard Error	Chi-Square	p-Value
Intercept($\hat{\lambda}$)	1	0.0687	0.0245		
Scale($\hat{\alpha}$)	1	1.1305	0.1383		
Stage II($\hat{\beta}_1$)	1	0.0981	0.4580	0.0459	0.8304
Stage III($\hat{\beta}_2$)	1	0.6293	0.3544	3.1517	0.0758
Stage IV($\hat{\beta}_3$)	1	1.7845	0.4128	18.6853	1.5416e-05

where $\lambda = e^{-\mu/\sigma}$, $\alpha = 1/\sigma$ and $\beta_i = -\gamma_i/\sigma$, $i = 1, 2, 3$.

Let $f_2\left(x,y,z,w,v\right)=\left(e^{-x/v},-y/v,-z/v,-w/v,1/v\right)$ then

$$\begin{aligned}
\dot{f}_2\left(x,y,z,w,v\right) &= \begin{bmatrix} -\frac{1}{v}e^{-x/v} & 0 & 0 & 0 & \frac{x}{v^2}e^{-x/v} \\ 0 & -\frac{1}{v} & 0 & 0 & \frac{y}{v^2} \\ 0 & 0 & -\frac{1}{v} & 0 & \frac{z}{v^2} \\ 0 & 0 & 0 & -\frac{1}{v} & \frac{w}{v^2} \\ 0 & 0 & 0 & 0 & -\frac{1}{v^2} \end{bmatrix} \\
\Rightarrow \dot{f}_2\left(\hat{\mu},\hat{\gamma}_1,\hat{\gamma}_2,\hat{\gamma}_3,\hat{\sigma}\right) &= \begin{bmatrix} -\frac{1}{\hat{\sigma}}e^{-\hat{\mu}/\hat{\sigma}} & 0 & 0 & 0 & \frac{\hat{\mu}}{\hat{\sigma}^2}e^{-\hat{\mu}/\hat{\sigma}} \\ 0 & -\frac{1}{\hat{\sigma}} & 0 & 0 & \frac{\hat{\gamma}_1}{\hat{\sigma}^2} \\ 0 & 0 & -\frac{1}{\hat{\sigma}} & 0 & \frac{\hat{\gamma}_2}{\hat{\sigma}^2} \\ 0 & 0 & 0 & -\frac{1}{\hat{\sigma}} & \frac{\hat{\gamma}_3}{\hat{\sigma}^2} \\ 0 & 0 & 0 & 0 & -\frac{1}{\hat{\sigma}^2} \end{bmatrix}
\end{aligned}$$

So the covariance matrix of $\hat{\lambda},\ \hat{\alpha}$ and $\hat{\beta}_i,\ i=1,2,3$ is

$$\hat{\Sigma}_2=\begin{bmatrix} 0.0006 & -0.0048 & -0.0048 & -0.0065 & -0.0023 \\ -0.0048 & 0.2098 & 0.0669 & 0.0685 & 0.0022 \\ -0.0048 & 0.0669 & 0.1256 & 0.0680 & 0.0017 \\ -0.0065 & 0.0685 & 0.0680 & 0.1704 & 0.0157 \\ -0.0023 & 0.0022 & 0.0017 & 0.0157 & 0.0191 \end{bmatrix}.$$

(ii) The plot of the survival functions for the patients of all four stages are shown in figure 1. The survival probabilities for the patients of stage I and II are similarly and the highest survival probability is the patients of stage I. This means the patients of stage I have the longer survival time than the remainder. On the other hand, the patients of stage IV have the smallest survival probability. And the survival curves of stage III and IV are different clearly form those of stage I and stage II. And all survival probabilities of the patients of the four stages are going decreased by time. That is a normal situation in the fact. Finally, we tend to believe that there is a discrepancy for the patients of stage IV and the remainder, and the stage III is.

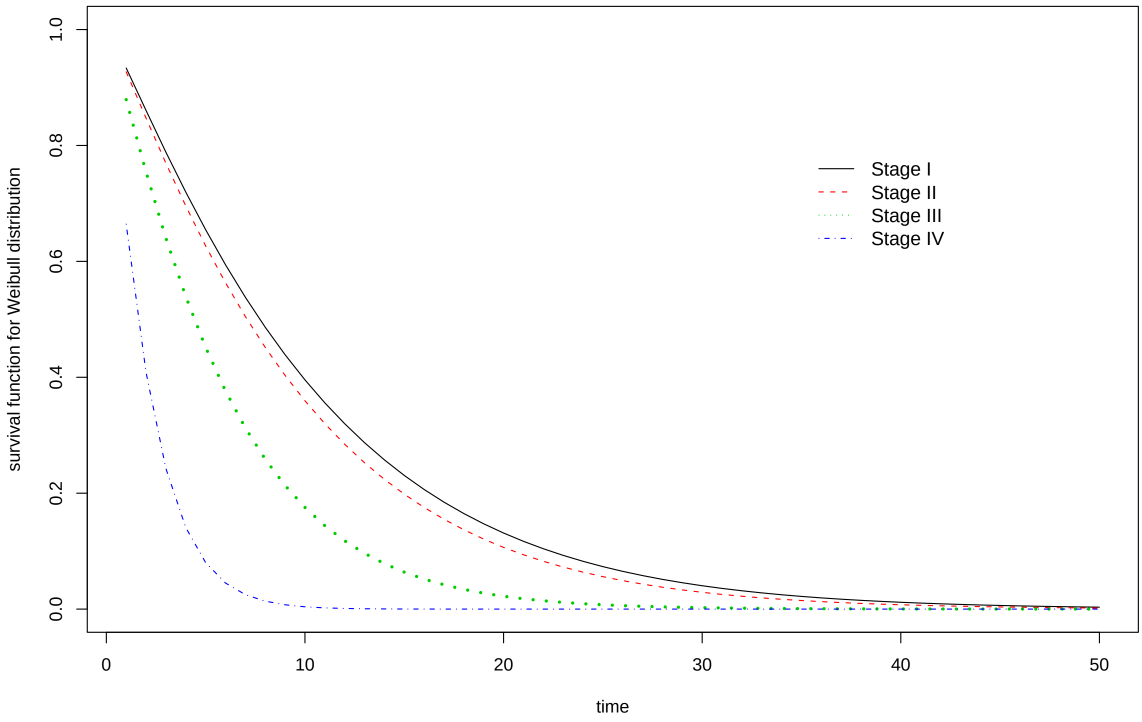

Figure 1: The survival functions for the patients of all four stage.

2. `Find the MLEs of the median survival for disease stage 1 and 4 patients. Use the delta method to find the corresponding standard errors.`

Since the model is Weibull distribution, the survival function is

$$\begin{aligned} S(x|\mathbf{Z}) &= \exp\left\{-x^{1/\sigma}e^{-\frac{\gamma^t\mathbf{Z}+\mu}{\sigma}}\right\} = \exp\left\{-\lambda x^{\alpha}e^{\beta^t\mathbf{Z}}\right\} \\ 0.5 &= S(x_{0.5}|\mathbf{Z}) = \exp\left\{-\lambda x_{0.5}^{\alpha}e^{\beta^t\mathbf{Z}}\right\} \\ \Rightarrow x_{0.5} &= \left(\frac{\log 2}{\lambda}e^{-\beta^t\mathbf{Z}}\right)^{1/\alpha}, \end{aligned}$$

where $\lambda = \exp\{-\mu/\sigma\}$, $\alpha = 1/\sigma$ and $\beta_i = -\gamma_i/\sigma,\ i = 1, 2, 3$.

Let $f_3(x,y,z,w,v) = \left(\frac{\log 2}{x}\, e^{yZ_1+zZ_2+wZ_3}\right)^{1/v}$ then $\dot{f}_3(x,y,z,w,v) = (g_1, g_2, g_3, g_4, g_5)$, where

$$\begin{aligned}
g_1 &= -\left(\frac{e^{-(yZ_1+zZ_2+wZ_3)}\log 2}{x}\right)^{1/v-1} \frac{e^{-(yZ_1+zZ_2+wZ_3)}\log 2}{x^2 v} \\
g_2 &= -\left(\frac{e^{-(yZ_1+zZ_2+wZ_3)}\log 2}{x}\right)^{1/v-1} \frac{Z_1 e^{-(yZ_1+zZ_2+wZ_3)}\log 2}{xv} \\
g_3 &= -\left(\frac{e^{-(yZ_1+zZ_2+wZ_3)}\log 2}{x}\right)^{1/v-1} \frac{Z_2 e^{-(yZ_1+zZ_2+wZ_3)}\log 2}{xv} \\
g_4 &= -\left(\frac{e^{-(yZ_1+zZ_2+wZ_3)}\log 2}{x}\right)^{1/v-1} \frac{Z_3 e^{-(yZ_1+zZ_2+wZ_3)}\log 2}{xv} \\
g_5 &= -\left(\frac{e^{-(yZ_1+zZ_2+wZ_3)}\log 2}{x}\right)^{1/v} \log\left(\frac{e^{-(yZ_1+zZ_2+wZ_3)}\log 2}{x}\right)\frac{1}{v^2}
\end{aligned}$$

So the covariance of $\hat{x}_{0.5}$ is

$$var(\hat{x}_{0.5}) = \dot{f}_3\left(\hat{\lambda}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\alpha}\right)\hat{\Sigma}_2 \dot{f}_3\left(\hat{\lambda}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\alpha}\right)^t.$$

(i) For stage I, $\mathbf{Z} = (0,0,0)$ and then the MLE of median survival and its variance are

$$\begin{aligned}
\hat{x}^I_{0.5} &= \left(\frac{\log 2}{\hat{\lambda}}\right)^{1/\hat{\alpha}} = \left(\frac{\log 2}{0.0687}\right)^{1/1.1305} = 7.7286 \\
var(\hat{x}_{0.5}) &= \left(-\left(\frac{\log 2}{\hat{\lambda}}\right)^{1/\hat{\alpha}-1}\frac{\log 2}{\hat{\alpha}\hat{\lambda}^2}, 0, 0, 0, -\left(\frac{\log 2}{\hat{\lambda}}\right)^{1/\hat{\alpha}} \log\left(\frac{\log 2}{\hat{\lambda}}\right)\frac{1}{\hat{\alpha}^2}\right)\hat{\Sigma}_2 \\
&\quad \left(-\left(\frac{\log 2}{\hat{\lambda}}\right)^{1/\hat{\alpha}-1}\frac{\log 2}{\hat{\alpha}\hat{\lambda}^2}, 0, 0, 0, -\left(\frac{\log 2}{\hat{\lambda}}\right)^{1/\hat{\alpha}} \log\left(\frac{\log 2}{\hat{\lambda}}\right)\frac{1}{\hat{\alpha}^2}\right)^t \\
&= 3.1796,
\end{aligned}$$

where $\hat{\Sigma}_2, \hat{\lambda}, \hat{\beta}_i,\ i = 1,2,3$ and $\hat{\alpha}$ are defined in problem 1. Thus, the standard error of MLE of the median survival for disease stage I is se$(\hat{x}^I_{0.5}) = 1.7832$.

(ii) For stage IV, $\mathbf{Z} = (0,0,1)$ and then the MLE of median survival and its variance are

$$\begin{aligned}
\hat{x}^{IV}_{0.5} &= (K)^{1/\hat{\alpha}} = 1.5942 \\
var(\hat{x}^{IV}_{0.5}) &= \left(-K^{1/\hat{\alpha}-1}\frac{K}{\hat{\alpha}\hat{\lambda}}, 0, 0, -K^{1/\hat{\alpha}-1}\frac{K}{\hat{\alpha}}, -K^{1/\hat{\alpha}}\log K\frac{1}{\hat{\alpha}^2}\right)\hat{\Sigma}_2 \\
&\quad \left(-K^{1/\hat{\alpha}-1}\frac{K}{\hat{\alpha}\hat{\lambda}}, 0, 0, -K^{1/\hat{\alpha}-1}\frac{K}{\hat{\alpha}}, -K^{1/\hat{\alpha}}\log K\frac{1}{\hat{\alpha}^2}\right)^t \\
&= 0.1907
\end{aligned}$$

where $K = \frac{\log 2}{\hat{\lambda}} e^{-\hat{\beta}_3} = 1.6942$. Thus, the standard error of MLE of the median survival for disease stage IV is $\text{se}(\hat{x}_{0.5}^{IV}) = 0.4366$.

3. `Test the hypothesis that the death rates the same for the patients of all four stages.`

For Weibull distribution,

$$\begin{aligned}
S(x|Z) &= P\left(W > \frac{\log x - \gamma^t Z - \mu}{\sigma}|Z\right) \\
&= P\left(e^W > \exp\left\{\frac{\log x - \gamma^t Z - \mu}{\sigma}\right\}\right) \\
&= \exp\left\{-x^{1/\sigma} e^{-\frac{\gamma^t Z - \mu}{\sigma}}\right\} \\
-f(x|Z) &= \frac{d}{dx} S(x|Z) = -\frac{1}{\sigma} x^{1/\sigma - 1} \exp\left\{-\frac{\gamma^t Z + \mu}{\sigma}\right\} S(x|Z) \\
\Rightarrow f(x|Z) &= \frac{1}{\sigma} x^{1/\sigma - 1} \exp\left\{-\frac{\gamma^t Z + \mu}{\sigma}\right\} S(x|Z) \\
h(x|Z) &= \frac{f(x|Z)}{S(x|Z)} = \frac{1}{\sigma} x^{1/\sigma - 1} \exp\left\{-\frac{\gamma^t Z + \mu}{\sigma}\right\}
\end{aligned}$$

and the fitted model is

$$Y = \log X = \hat{\mu} + \hat{\gamma}_1 Z_1 + \hat{\gamma}_2 Z_2 + \hat{\gamma}_3 Z_3 + \hat{\sigma} W = 2.3691 - 0.0868 Z_1 - 0.5566 Z_2 - 1.5786 Z_3 + 0.885 W.$$

Testing the death rates are the same for the patients of all four stages that we can take the hypotheses $H_0 : h(t|\mathbf{Z}_0) = h(t|\mathbf{Z}_1) = h(t|\mathbf{Z}_2) = h(t|\mathbf{Z}_3)$ vs. H_1 :At least one of $h(t|\mathbf{Z}_i)$ is different, $i = 0,\ 1,\ 2,\ 3$, where $\mathbf{Z} = (Z_1,\ Z_2,\ Z_3)$ and

$$\begin{aligned}
\mathbf{Z}_0 &= (Z_1,\ Z_2,\ Z_3) = (0,\ 0,\ 0) \\
\mathbf{Z}_1 &= (Z_1,\ Z_2,\ Z_3) = (1,\ 0,\ 0) \\
\mathbf{Z}_2 &= (Z_1,\ Z_2,\ Z_3) = (0,\ 1,\ 0) \\
\mathbf{Z}_3 &= (Z_1,\ Z_2,\ Z_3) = (0,\ 0,\ 1)\,.
\end{aligned}$$

The testing hypotheses in the above is the same as testing $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$ vs. $H_1 : \gamma_i \neq 0,\ i = 1,\ 2,\ 3$.

Define the contrast matrix $\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Thus, the testing hypotheses becomes $H_0 : \mathbf{C}\gamma = 0$ vs. $H_1 : \mathbf{C}\gamma \neq 0$ with $\gamma = (\gamma_1,\ \gamma_2,\ \gamma_3)^t$. So the variance of $\mathbf{C}\gamma$ is

$$\begin{aligned} var(\mathbf{C}\gamma) &= \mathbf{C}var(\gamma)\mathbf{C}^t = var(\gamma) \\ &= \begin{bmatrix} 0.1640 & 0.0520 & 0.0519 \\ 0.0520 & 0.1015 & 0.0575 \\ 0.0519 & 0.0575 & 0.1319 \end{bmatrix} \end{aligned}$$

and hence, the test statistic is

$$X_W^2 = (\mathbf{C}\hat{\gamma})\left[var(\mathbf{C}\hat{\gamma})\right]^{-1}(\mathbf{C}\hat{\gamma})^t = 20.8814$$

with $\mathcal{X}_3^2$ distribution and its corresponding p-value is 0.0001. Since the p-value < 0.05, we reject H_0 at significant level 0.05, that is, we know that the death rates are different for the patients of all four stages. This result is the same as (ii) in problem 1 which obtained by the survival functions for the patients of all four stages in figure 1.

4. `Fit all possible parametric models (exponential, Weibull, log normal, and log logistic) including only one variable, disease stage. Use all the graphic techniques you learn in the class to determine which model is appropriate to the data.`

For all possible parameter models (exponential, Weibull, log normal, and log logistic), the fitted models are

$$\begin{aligned} Y &= \log X = 2.4476 - 0.0827Z_1 - 0.6164Z_2 - 1.6758Z_3 + W \\ Y &= \log X = 2.3691 - 0.0868Z_1 - 0.5566Z_2 - 1.5786Z_3 + 0.885W \\ Y &= \log X = 2.188 - 0.228Z_1 - 0.889Z_2 - 1.906Z_3 + 1.27W \\ Y &= \log X = 2.107 - 0.115Z_1 - 0.784Z_2 - 1.780Z_3 + 0.714W, \end{aligned}$$

respectively. And the hazard plots for all possible parameter models are as follows

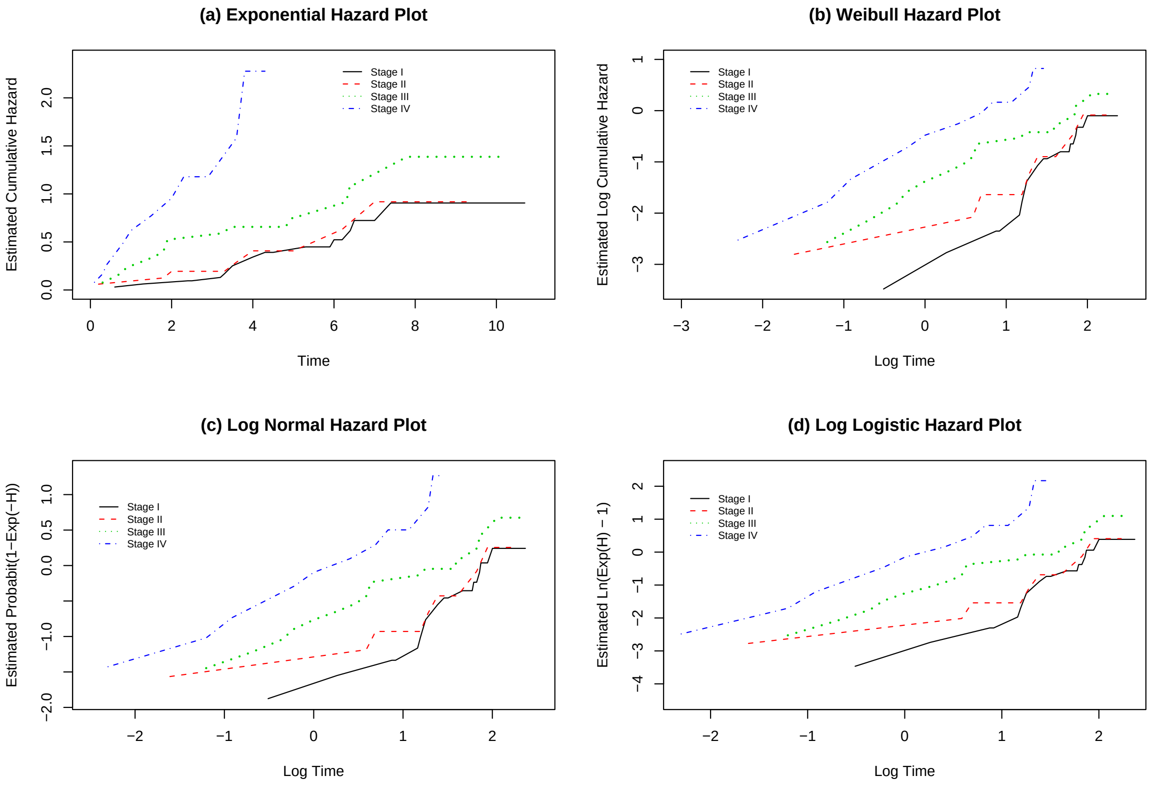

Figure 2: The hazard plot for all possible parameter models.

For all plots in figure 2, both death risks for the patients between stage I and stage II are closed to each other. Furthermore, the patients of stage IV have the highest death risk no mater for any possible parameter models. Also, the hazard plot for the patients of all four stages in figure 2(a) is similar to be a linear before the survival time 7.5 months and its tail is horizontal for stage I-III. However, we can still claim this plot to be linear. Except plot in figure 2(a). the rest plots are approximate to be a linear. Therefore, the hazard plots, all of which should be linear, suggest that any of the models would be reasonable. Thus, we try to plot the Cox-Snell residuals plot to determine which model is suitable.

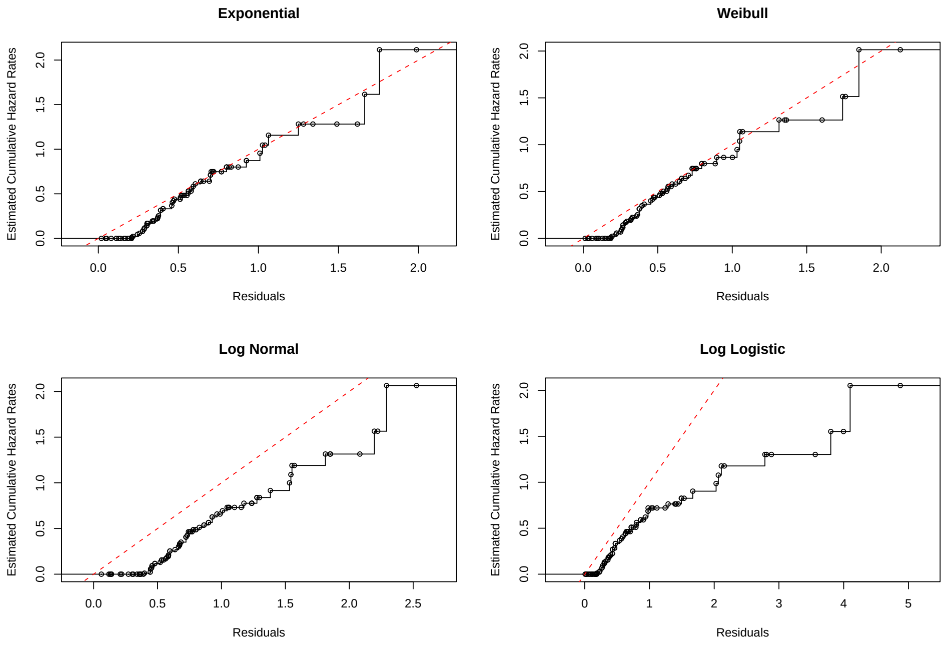

Figure 3: The hazard plot for all possible parameter models.

The Cox-Snell residual, r_j, is defined by $r_j = \hat{H}(T_j|\mathbf{Z}_j)$, where $\hat{H}$ is the fitted model. If the model fits the data then the r_j's should have a certain distribution, so that a hazard plot of r_j versus the Nelson-Aalen estimator of the cumulative hazard of the r_j's should be a straight line with slope 1 where the Cox-Snell residuals for the four models considered are

$$\begin{aligned} \text{Exponential} \qquad & r_i = \hat{\lambda} t_j \exp\left\{\hat{\beta}^t Z_i\right\} \\ \text{Weibull} \qquad & r_i = \hat{\lambda} \exp\left\{\hat{\beta}^t Z_i\right\} t_j^{\hat{\alpha}} \\ \text{Log normal} \qquad & r_i = \log\left\{1 - \Phi\left(\frac{\log T_j - \hat{\mu} - \hat{\gamma}^t Z_i}{\hat{\sigma}}\right)\right\} \\ \text{Log logistic} \qquad & r_i = \log\left\{\frac{1}{1 + \hat{\lambda} \exp\left\{\hat{\beta}^t Z_i\right\} t_j^{\hat{\alpha}}}\right\}. \end{aligned}$$

Thus we see that both exponential and weibull distributions are closed to a straight line in the figure 3. Although the rest of models (log normal and log logistic) are also closed to a straight line, both of their slopes are not equal to 1 (the red line in the figure 4 means a straight line with slope 1). So we attend to believe that exponential or weibull model is appropriate to the data. But using exponential distribution to fit the model is much better than weibull distribution because we only estimate an unknown parameter by using exponential distribution.

5. `Repeat question 4 but including all variables of the data. Use the Cox-Snell residuals plot to determine a suitable model for the data.`

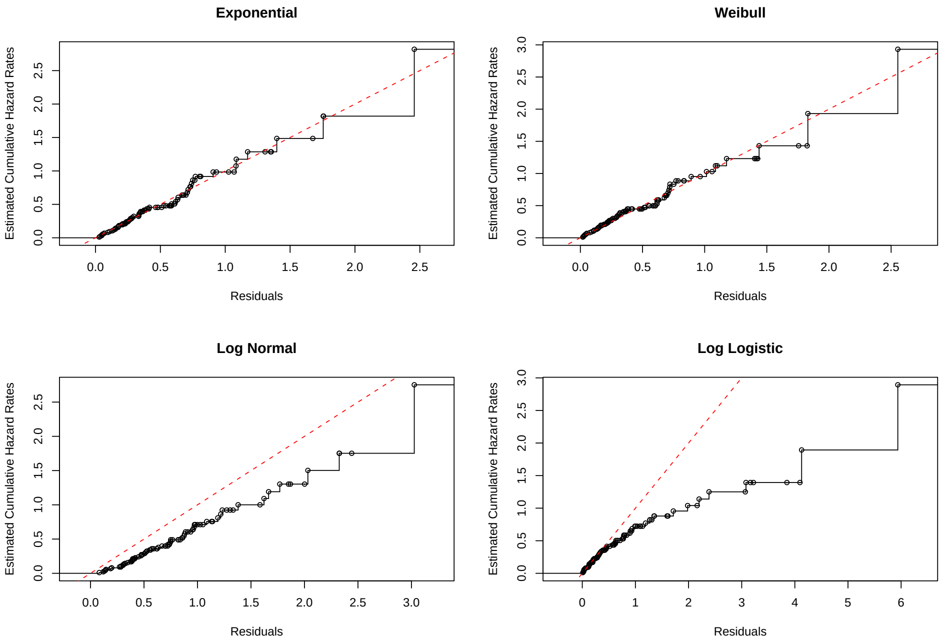

Figure 4: The hazard plot for all possible parameter models.

Figure 4 is the Cox-Snell residuals plot with including all variables of the data and the red line means a straight line with slope 1. We can obtain the same result as problem 4 form the above figure. That is, both exponential and weibull models are closed to a straight line with slope 1. Although the log normal and log logistic models are closed to a straight line, the slopes of those straight lines are not equal to 1. Furthermore, we have to estimate the number of unknown parameters for the exponential model are less than the weibull model needed. Hence, we tend to believe the exponential model is a suitable model for the data.

6. `Perform statistical method to find a best model from question 6.`

First, we compare the fit of the exponential, Weibull, log normal, and generalized gamma models for the data on the laryngeal cancer. We fit the log linear model

$$Y = \log X = \mu + \sum_{k=1}^{5} \gamma_k Z_k + \sigma W,$$

where we have five covariates:

Z_1 : 1if Stage II cancer, 0 otherwise,

Z_2 : 1if Stage III cancer, 0 otherwise,

Z_3 : 1if Stage IV cancer, 0 otherwise, and

Z_4 : Patient's age at diagnosis,

Z_5 : the year of diagnosis.

We use SAS PORC LIFEREG to fit the exponential, Weibull, log normal, and generalized gamma models and their output in the table as follows:

Table: *Parameteric Models for the Laryngeal Cancer Study*

	Exp.		Weibull		Log Normal		Log Logistic		G. Gamma	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
μ	1.18	5.37	2.11	4.92	0.52	5.52	2.08	5.47	1.39	5.42
γ_1	-0.17	0.46	-0.16	0.41	-0.25	0.46	-0.14	0.43	-0.19	0.45
γ_2	-0.65	0.36	-0.59	0.32	-0.93	0.37	-0.82	0.36	-0.79	0.41
γ_3	-1.70	0.42	-1.59	0.40	-1.94	0.47	-1.80	0.46	-1.80	0.49
γ_4	-0.02	0.01	-0.02	0.01	-0.02	0.01	-0.01	0.01	-0.02	0.01
γ_5	0.03	0.07	0.09	0.07	0.04	0.07	0.01	0.07	0.03	0.07
σ	1.00	0	0.89	0.11	1.27	0.14	0.72	0.09	1.13	0.26
θ									0.42	0.60
Log L	-108.38		-107.98		-107.85		-108.17		-107.60	
AIC	228.76		229.97		299.71		230.35		231.21	

In the above table, we see that all models fit equally well. The exponential model has the smallest AIC and, in the sense, is the best fitting model. For this model,

$$Y = \log X = 1.1782 - 0.1722Z_1 - 0.6536Z_2 - 1.7015Z_3 - 0.0191Z_4 + 0.0346Z_5 + W.$$

The negative values of the coefficients of $Z_1 - Z_4$ in the log linear model suggest that the individuals with stage II-IV cancer have shorter lifetimes than individuals with stage I disease. Also, grey-haired patients at diagnosis have shorter lifetimes than young patients. A second model of interest is the generalized gamma distibution. For this model, $Y = \log X$ follows the linear model

$$Y = \log X = \mu + \gamma^t \mathbf{Z} + \sigma W$$

with W having the following probability density function

$$f(w) = \frac{|\theta| \left[\exp\left(\theta w\right)/\theta^2\right]^{1/\theta^2} \exp\left[-\exp\left(\theta w\right)/\theta^2\right]}{\Gamma(1/\theta^2)}, \qquad -\infty < w < \infty.$$

When θ equals 1, this model reduces to the Weibull regression model, and when θ is 0, the model reduces to the log normal distribution. When $\theta = 1$ and $\sigma = 1$ in the linear model, then it reduces to the exponential regression model. Wald or likelihood ratio tests of the hypotheses that $\theta = 1$ or $\theta = 0$ provid a means of checking the assumption of a Weibull or log normal regression model, respectively. Thus, we can test $\theta = 1$ and $\sigma = 1$, that is, $H_0 : \theta = 1, \ \sigma = 1$ vs. H_1 : At least one is false. Then the Wald test statistic is

$$X_W^2 = \begin{bmatrix} \hat{\theta}_{GG} - 1 \\ \hat{\sigma}_{GG} - 1 \end{bmatrix}^t \hat{\Sigma}_{GG} \begin{bmatrix} \hat{\theta}_{GG} - 1 \\ \hat{\sigma}_{GG} - 1 \end{bmatrix} = 0.0489$$

and its p-value is 0.9758>0.05 where θ_{GG} and σ_{GG} are the fitted values for the generalized gamma model

$$\hat{\Sigma}_{GG} = \begin{bmatrix} 0.067314 & -0.137374 \\ -0.137374 & 0.357744 \end{bmatrix}$$

is the covariance matrix of $\hat{\theta}$ and $\hat{\sigma}$ for the generalized gamma model. That means we do not reject H_0 at significant level 0.05, that is, we attend to believe that $\theta = 1$ and $\sigma = 1$ and Hence, we prefer the exponential model to be a best model for the data on laryngeal cancer.

7. `Perform model selection for your best model from question 6.`

The objective of AIC model selection is to estimate the information loss when the probability distribution f associated with the true (generating) model is approximated by probability distribution g, associated with the model that is to be evaluated. Akaike (1973; Bozdogan, 1987) has shown that choosing the model with the lowest expected information loss is asymptotically equivalent to choosing a model $M_i, i = 1, \ 2, \cdots, \ K$ that has the lowest AIC value. The AIC is defined as

$$\text{AIC} = -2 * \log\left(\text{Likelihood}\right) + 2\left(p + k\right),$$

where the $p+k$ is free parameters in such a way as to maximize the probability that the candidate model has generated the observed data.

Despite the widespread use of the AIC, some believe that it is too liberal and tends to select overly complex models. 1995). It has been pointed out that the AIC neglects the sampling variability of the estimated parameters. When the likelihood values for these parameters are not highly concentrated around their maximum value, this can lead to overly optimistic assessments. Furthermore, the AIC is not consistent. That is, as the number of observations n grows very large, the probability that the AIC recovers a true low-dimensional model does not approach unity. A popular alternative model selection criterion is the Bayesian information criterion or BIC. The BIC for model i is defined as

$$\text{BIC} = -2 * \log(\text{Likelihood}) + \log n/2$$

where n is the number of observations that enter into the likelihood calculation. In contrast to the AIC, the BIC is consistent as $n \to \infty$ and does take parameter uncertainty into account.

A comparison of BIC and AIC shows that the BIC penalty term is larger than the AIC penalty term when $n > e^2$. Although the equations of AIC and BIC look very similar, they originate from quite different frameworks. The BIC assumes that the true generation model is in the set of candidate models,and it measures the degree of belief that a certain model is the true data-generating model. The AIC does not assume that any of the candidate models is necessarily true.And for the forward step, the BIC contains a few model parameters than AIC does. Thus, we follow the BIC to perform model selection for the best model from problem 6.

First step: *The BIC for the model selection.*

Model	BIC
$Y = \log X = \mu + W$	237.6908
$Y = \log X = \mu + \gamma_1 Z_1 + W$	236.5680
$Y = \log X = \mu + \gamma_2 Z_2 + W$	236.8570
$Y = \log X = \mu + \gamma_3 Z_3 + W$	224.4834
$Y = \log X = \mu + \gamma_4 Z_4 + W$	234.8472
$Y = \log X = \mu + \gamma_5 Z_5 + W$	237.6202

We want to choose the smallest value of BIC. So, from the above table, the smallest BIC occurs in the model $Y_1 = \mu + \gamma_3 Z_3 + W$. Again, we consider the other covariate into the model Y_1 and compute the BIC values.

Second step: *The BIC for the model selection.*

Model	BIC
$Y = \log X = \mu + \gamma_3 Z_3 + W$	224.4834
$Y = \log X = \mu + \gamma_1 Z_1 + \gamma_3 Z_3 + W$	224.2495
$Y = \log X = \mu + \gamma_2 Z_2 + \gamma_3 Z_3 + W$	221.2630
$Y = \log X = \mu + \gamma_3 Z_3 + \gamma_4 Z_4 + W$	222.7064
$Y = \log X = \mu + \gamma_3 Z_3 + \gamma_5 Z_5 + W$	224.1159

From the table of the second step, the model $Y_2 = \mu + \gamma_2 Z_2 + \gamma_3 Z_3 + W$ is better model than the model Y_1.

Third step: *The BIC for the model selection.*

Model	BIC
$Y = \log X = \mu + \gamma_2 Z_2 + \gamma_3 Z_3 + W$	221.2630
$Y = \log X = \mu + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + W$	221.0687
$Y = \log X = \mu + \gamma_2 Z_2 + \gamma_3 Z_3 + \gamma_4 Z_4 + W$	219.3505
$Y = \log X = \mu + \gamma_2 Z_2 + \gamma_3 Z_3 + \gamma_5 Z_5 + W$	220.9023

From the table of the third step, the model $Y_3 = \mu + \gamma_2 Z_2 + \gamma_3 Z_3 + \gamma_4 Z_4 + W$ is better than the model Y_2. By the same argument, we put the covariate into the model Y_3 and compute the BICs.

Fourth step: *The BIC for the model selection.*

Model	BIC
$Y = \log X = \mu + \gamma_2 Z_2 + \gamma_3 Z_3 + \gamma_4 Z_4 + W$	219.3505
$Y = \log X = \mu + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + \gamma_4 Z_4 + W$	219.2522
$Y = \log X = \mu + \gamma_2 Z_2 + \gamma_3 Z_3 + \gamma_4 Z_4 + \gamma_5 Z_5 + W$	219.1502

Since all the values of BIC in the table of the fourth step are similar, we will not add any covaraite into the model Y_3. Hence, we preform the model $Y_3 = \mu + \gamma_2 Z_2 + \gamma_3 Z_3 + \gamma_4 Z_4 + W$ is the best model for the data on the laryngeal cancer.

8. `Interpret your final model. Write your answer as much as you can.`

The best fitted model in problem 7 is

$$Y = 3.6812 - 0.6038 Z_2 - 1.5925 Z_3 - 0.0193 Z_4.$$

The negative values of the coefficients of Z_2 and Z_3 in the log linear model suggest that individuals with stages III and IV cancer have shorter lifetimes than individuals without stage III or stage IV disease. Also, the negative values of the coefficient of Z_4 (age) in the log linear model means that there has shorter lifetimes with the advanced age than with the young age and the lifetimes decreases 0.0193 years when patient's age at diagnosis increases one month. On the other hand, the hazard function of the exponential model is

$$h(t|\mathbf{Z}^*) = \exp\{-(\gamma^* \mathbf{Z}^* + \mu)\}$$

where $\hat{\gamma}^* = (-0.6038, -1.5925, -0.0193)^t$ and $\mathbf{Z}^* = (Z_2, Z_3, Z_4)^t$.SO, we can compare the relative risk of death between the patients at stage III cancer and at stage IV cancer. Then,the point estimator of the relative risk of death for the stage III cancer as compared to a stage IV cancer is

$$\begin{aligned} \frac{\hat{h}(t|Z_2=1,\ Z_3=0,\ Z_4=c)}{\hat{h}(t|Z_2=0,\ Z_3=1,\ Z_4=c)} &= \frac{\exp\{-\hat{\gamma}_2^*-c\hat{\gamma}_4^*\}}{\exp\{-\hat{\gamma}_3^*-c\hat{\gamma}_4^*\}} \\ &= \exp\{\hat{\gamma}_3^*-\hat{\gamma}_2^*\}=\exp\{-1.5925+0.6038\} \\ &= 0.3721 \end{aligned}$$

where the patient's age at disease is fixed at a constant c. This means the death risk for the patients of stage III cancer is 0.3721 times for the patients of stage IV cancer under the same age at disease, that is, there has the higher death risk in stage IV cancer than in stage III cancer. Next, we also can obtain the 95% confidence interval of the relative risk of death by using the delta method. Thus, the covariance matrix of $\hat{\gamma}_2^*$ and $\hat{\gamma}_3^*$ is

$$\hat{\Sigma}^* = \begin{bmatrix} 0.1044 & 0.0449 \\ 0.0449 & 0.1387 \end{bmatrix}.$$

Then the point estimate of $\gamma_3^*-\gamma_2^*$ is equal to $\hat{\gamma}_3^*-\hat{\gamma}_2^*=-0.9887$ and its relative variance of $\hat{\gamma}_3^*-\hat{\gamma}_2^*$ is

$$var\left(\hat{\gamma}_3^*-\hat{\gamma}_2^*\right)=var(\hat{\gamma}_2^*)-2*cov(\hat{\gamma}_2^*,\hat{\gamma}_3^*)+var(\hat{\gamma}_3^*)=0.1533.$$

So the 95% confidence interval for $\gamma_3^*-\gamma_2^*$ is

$$\hat{\gamma}_3^*-\hat{\gamma}_2^*\pm 1.96\sqrt{var\left(\hat{\gamma}_3^*-\hat{\gamma}_2^*\right)}=-0.9887\pm 1.96*0.3916=\left(-1.7561,-0.2212\right).$$

Hence, the 95% confidence interval for the relative risk of death for the stage III cancer as compared to a stage IV cancer is

$$\left(\exp\left\{\hat{\gamma}_3^*-\hat{\gamma}_2^*-1.96\sqrt{var\left(\hat{\gamma}_3^*-\hat{\gamma}_2^*\right)}\right\},\hat{\gamma}_3^*-\hat{\gamma}_2^*+1.96\sqrt{var\left(\hat{\gamma}_3^*-\hat{\gamma}_2^*\right)}\right)=\left(0.1727,0.8015\right).$$

This means we have 95% confidence level to see that the death risk for the patients of stage III cancer is between 0.1727 and 0.8015 times for the patients of stage IV cancer.

From the all above results, we know that if the lifetimes of an individual's advanced age at disease will be shorter than other young person under the same stage level. And the patients of stage IV cancer will have shorter lifetime than stage III cancer if the patient's age at disease is fixed, that is, the patients of stage IV cancer has higher death risk than stage III cancer and their relative risk of death is about 0.3217.

☞ II. Nonparametric model

9. Obtain the Kaplan-Meier estimates of survival functions for the patients of four stages and perform a test to find if there are differences in survival among the four stages. Make a brief conclusion for both graphs and the test.

(i) The plot of the Kaplan-Meier estimates of survival functions for the patients of four stages shows that the survival probabilities for the patients between stage I and II are similar. Furthermore, their survival probabilities are identical after four months. On the other hand, the stage III and stage IV have smaller survival probability than stage I and II. And there is a discrepancy survival probabilities for the patients between stage (III, IV) and stage (I, II). Also, the plot in the below shows that the patients of stage IV have the smallest survival probability.

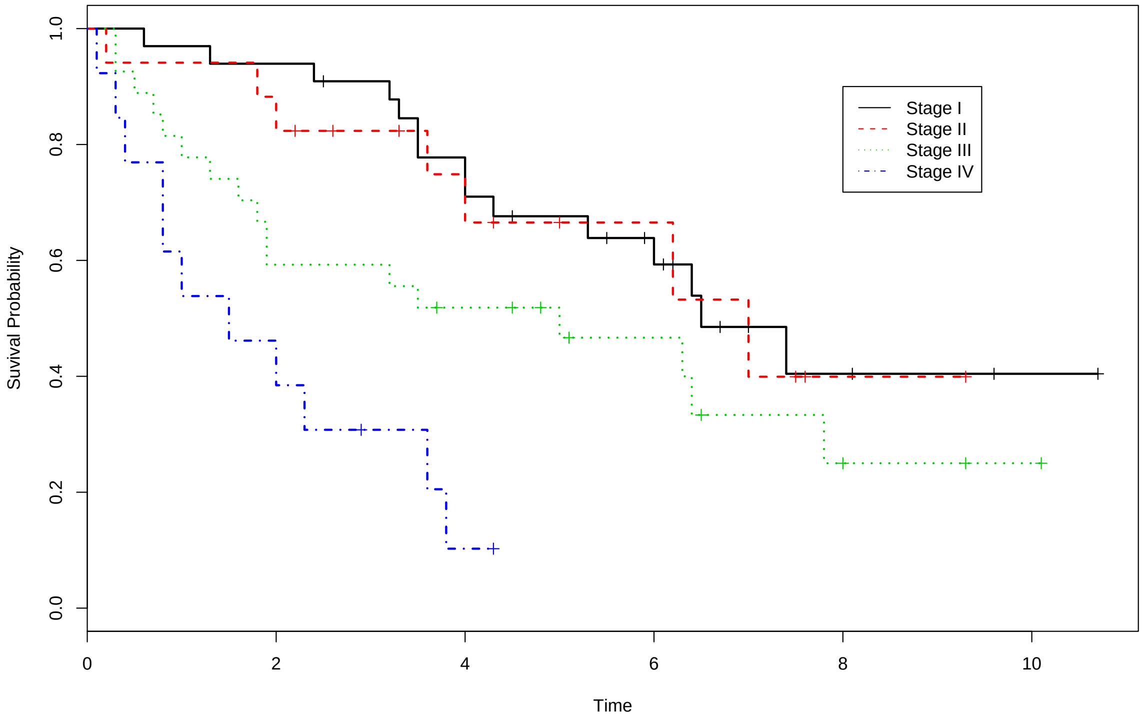

Figure 5: The survival functions for the patients of all four stage using Kaplan-Meier estimation.

(ii) So we can test the differences in survival among the four stages by the log rank test. That is, testing the hypotheses $H_0 : S_I(t_0) = S_{II}(t_0) = S_{III}(t_0) = S_{IV}(t_0)$ vs. H_1 : at least one of the $S_j(t_0)$ is different. First, we take the weight $W_{k_i} = n_i$ (Gehan-Wilcoxon) then the test statistic is

$$X_{GW}^2 = 22.8 \sim \mathcal{X}_3^2 \text{ and its p-value is } 4.53e - 05 < 0.05.$$

This means we reject H_0 at significant level 0.05. We see that at least one of $S_j(t_0)$ is different at significant level 0.05. So there are differences in survival among the four stages. Second, we try to take the weight $W_{k_i} = \hat{S}(t_i)$ (Peto) then the test statistic is

$$X^2_{Peto} = 23.1 \sim \mathcal{X}^2_3 \text{ and its p-value is } 3.85e - 05 < 0.05.$$

This has the same result, rejects H_0 at significant level 0.05, as taking the weight $W_{k_i} = n_i$. Hence, we know that there are differences in survival among the four stages.

10. `Find Nair's` $c_\alpha(a_L, A_U)$ `and Hall-Wellners'` $k_\alpha(a_L, a_U)$ `for overall survival time period.`

Define

$$a_L = \frac{n\sigma_s^2(t_L)}{1 + n\sigma_s^2(t_L)}$$

and

$$a_U = \frac{n\sigma_s^2(t_U)}{1 + n\sigma_s^2(t_U)},$$

where $\sigma_s^2(t) = \sum_{t_i \le t} = \frac{d_i}{Y_i(Y_i - d_i)}$ and Y_i means the number of individuals who are at risk at time t_i, d_i means the number of events (sometimes simply referred to as deaths) at time t_i. Also we compute a_L and a_U and use the interpolation to obtain $c_\alpha(a_L, a_U)$ and $k_\alpha(a_L, a_U)$. So the outcomes are as follows:

Stage	Time (t_L, t_U)	(a_L, a_U)	Nair's $c_\alpha(a_L, a_U)$	Hall-Wellner's $k_\alpha(a_L, a_U)$
I	(0.6, 7.4)	(0.03,0.72)	3.0595	1.3501
II	(0.2, 7.0)	(0.06,0.74)	3.0059	1.3525
III	(0.3,7.8)	(0.07,0.83)	3.0377	1.3576
IV	(0.1,3.8)	(0.08,0.92)	3.0892	1.3581

11. `Use nonparametric method to obtain the point estimates and confidence intervals for the mean and median survival time of the four stages patients. Make a brief interpretation on the median survival time.`

(i) The estimated mean restricted to the interval $[0, \tau]$, with τ either the longest observed time or preassigned by the investigator, is given by

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt.$$

The variance of this estimator is

$$\hat{V}(\hat{\mu}_{\tau}) = \sum_{i=1}^{D} \left[\int_{0}^{\tau} \hat{S}(t)dt \right]^2 \frac{d_i}{Y_i(Y_i - d_i)}$$

where the definitions of Y_i and d_i are the same as problem 10. A $100(1-\alpha)\%$ confidence interval for the mean is expressed by

$$\hat{\mu}_{\tau} \pm Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\mu}_{\tau})}.$$

Here, we use R program to obtain the point estimates and confidence intervals for the mean of the four stages and the codes are I-10 in the Appendix.

Disease Stage	Mean	variance	Standard Error	95% confidence interval
I	7.0833	0.4309	0.6565	(5.7966,8.3699)
II	6.8668	1.0354	1.0176	(1.8724,8.8612)
III	5.1272	0.6481	0.8050	(3.5494,6.7051)
IV	2.5641	0.7830	0.8849	(0.8298,4.2984)

(ii) Define $x_p = \inf\{t : S(t) \leq 1-p\}$, that is, x_p is the smallest time at which the survival function is less than or equal to $1-p$. To estimate x_p, we find the smallest time $\hat{x}_p$ for which the Product-Limit estimator is less than or equal to $1-p$. That is, $\hat{x}_p = \inf\left\{t : \hat{S}(t) \leq 1-p\right\}$. So a $100(1-\alpha)\%$ confidence interval for x_p, based on the linear confidence interval, is the set of all time points t which satisfy the following condition:

$$-Z_{1-\alpha/2} \leq \frac{\hat{S}(t)-(1-p)}{\hat{V}^{1/2}\left[\hat{S}(t)\right]} \leq Z_{1-\alpha/2}.$$

Thus, we can obtain the point estimates and confidence interval for the median survival time of the four stages patients as follows

Disease Stage	$\hat{x}_{0.5}$	95% confidence interval
I	6.5	$(5.3, \infty)$
II	7.0	$(4.0, \infty)$
III	5.0	(1.8,6.4)
IV	1.5	(0.8,2.3)

Again, the above results are computed by R program and the codes and complete outputs are I-11 in the Appendix.

☞ Appendix

I-1

```
##Fit a Weibull model to the data including only one variable, disease variable
fit1 = survreg(Surv(futime, fustat) ~ factor(stage), dist = "weibull")
time = 1 : 50
z1 = c(1, 0, 0, 0) ; z2 = c(1, 1, 0, 0)
z3 = c(1, 0, 1, 0) ; z4 = c(1, 0, 0, 1)
##To compute the survival functions of all four stages
gamma.z1 = sum(z1 * fit1 $ coef) ; gamma.z2 = sum(z2 * fit1 $ coef)
gamma.z3 = sum(z3 * fit1 $ coef) ; gamma.z4 = sum(z4 * fit1 $ coef)
arg1 = time * exp(-gamma.z1) ; arg2 = time * exp(-gamma.z2)
arg3 = time * exp(-gamma.z3) ; arg4 = time * exp(-gamma.z4)
alpha.1 = (fit1 $ scale) ^ (-1)
surv.1 = exp(-arg1 ^ alpha.1) ; surv.2 = exp(-arg2 ^ alpha.1)
surv.3 = exp(-arg3 ^ alpha.1) ; surv.4 = exp(-arg4 ^ alpha.1)
##According the above result to plot survival functions of four stages in one figure
##The lines of stage III and IV are similar to lines(...)
plot(time, surv.1, xlab = "time", ylab = "survival function
     for Weibull distribution", type = "l", ylim = c(0, 1))
lines(time, surv.2, lty = 2, col = 2)
```

I-2

```
##To compute MLEs of the median survival for stage I and IV
alpha.1 = 1 / fit1 $ scale ; lambda.1 = exp(-fit1 $ coef[1] / fit1 $ scale)
beta.1 = as.matrix(-fit1 $ coef[-1] / fit1 $ scale)
z.1 = c(0, 0, 0) ; z.4 = c(0, 0, 1)
median.1 = (log(2) * exp(-z.1 %*% beta.1) / lambda.1) ^ (1 / alpha.1)
median.4 = (log(2) * exp(-z.4 %*% beta.1) / lambda.1) ^ (1 / alpha.1)
```

```
##Use the programs deriv(....) and attr(eval(),"gradient") to apply the delta method
```

I-4

```
##Change dist="" to weibull, log normal and log logistic can fit different model
fit.exp = survreg(Surv(futime, fustat) ~ factor(stage), dist = "exponential")
##Computing the Kaplan-Meier estimator for each stage i.e. change "stage ==..."
fit.I = survfit(Surv(futime, fustat) ~ 1, data = larynx[stage == 1, ])
##plot the hazard rate for all stages
##the below program is to plot exponential and the rest omitted
plot(time.I, H1, type = "l", main = "(a) Exponential Hazard Plot", xlab = "Time",
     ylab = "Estimated Cumulative Hazard", xlim = c(0, 11), ylim = c(0, 2.4))
lines(time.II, H2, lty = 2, col = 2)
lines(time.III, H3, lty = 3, col = 3, lwd = 2)
lines(time.IV, H4, lty = 4, col = 4)
```

I-5

```
##plot the Cox-snell residuals plot
##This is to fit exponential model and the rest are similar
fit.all.exp = survreg(Surv(futime, fustat) ~ factor(stage) + age + year
                , dist = "exponential")
summary(fit.all.exp)
ri.exp = exp(log(futime) - (fit.all.exp $ linear.predictors) / fit.all.exp $ scale)
fh.surv.exp = survfit(Surv(ri.exp, fustat) ~ 1, type = "fleming-harrington") $ surv
cum.hz.exp = -log(fh.surv.exp)
s.exp = stepfun(sort(ri.exp), c(0, cum.hz.exp))
```

I-6 SAS Program

```
##Find the fitted model for all possible parametric models
##Only change DIST option to weibull lnormal llogistic
```

```
PROC LIFETIME;
    MODEL FUTIME*FUSTAT(0)=Z1 Z2 Z3 AGE YEAR/DIST=EXPONENTIAL;
##Fit a model with generalized gamma
##COVB option is use to call the covariance matirx of all parameters
PROC LIFETIME;
    MODEL FUTIME*FUSTAT(0)=Z1 Z2 Z3 AGE YEAR/DIST=GAMMA COVB;
```

I-9

```
##test to find if there are differences in survival among the four
##stages with different weight.rho=0 means wieght =ni and rho=1 means
##wieight=S(ti)
test0 = survdiff(Surv(futime, fustat) ~ factor(stage), rho = 0)
test1 = survdiff(Surv(futime, fustat) ~ factor(stage), rho = 1)
```

I-10

```
##computing the aL and aU for all four stages
km = function(data, stage, i)
{
    km = summary(survfit(Surv(futime, fustat) ~ stage, data = data[stage == i, ]))
    time = km $ time
    y = km $ n.risk
    d = km $ n.event
    surv = km $ surv
##each time of d / (y * (y-d))
    sigma.s = cumsum(d / (y * (y - d)))
    temp = d / (y * (y - d))
    v.s = surv ^ 2 * sigma.s
    tL = time[1] ; tU = rev(time)[1]
##y[1] means the sample size for each stage
    tempL = y[1] * sigma.s[1]
    tempU = y[1] * rev(sigma.s)[1]

    aL = tempL / (1 + tempL) ; aU = tempU / (1 + tempU)

    output1 = cbind(time, d, y, surv, temp, sigma.s, v.s)

    output2 = c(y[1], tL, tU, aL, aU)

    return(list(prod.lim = output1, est = round(output2, 2)))

}

km1 = km(larynx, stage, 1)

km2 = km(larynx, stage, 2)

km3 = km(larynx, stage, 3)

km4 = km(larynx, stage, 4)
```

I-11 Tables of a 95% confidence interval for the median of all four stages

Table : stage I

t_i	$\hat{S}(t_i)$	$\sqrt{\hat{V}\left[\hat{S}(t_i)\right]}$	$\frac{\hat{S}(t_i)-(1-p)}{\hat{V}^{1/2}\left[\hat{S}(t_i)\right]}$
0.6	0.9697	0.0298	15.7403
1.3	0.9394	0.0415	10.5786
2.4	0.9091	0.0500	8.1747
3.2	0.8777	0.0573	6.5922
3.3	0.8452	0.0637	5.4165
3.5	0.7776	0.0744	3.7293
4.0	0.7010	0.0819	2.5641
4.3	0.6762	0.0847	2.0804
5.3	0.6386	0.0879	1.5766
6.0	0.5930	0.0927	1.0030
6.4	0.5391	0.0987	0.3960
6.5	0.4852	0.1025	-0.1445
7.4	0.4043	0.1129	-0.8474

Table : stage II

t_i	$\hat{S}(t_i)$	$\sqrt{\hat{V}\left[\hat{S}(t_i)\right]}$	$\frac{\hat{S}(t_i)-(1-p)}{\hat{V}^{1/2}\left[\hat{S}(t_i)\right]}$
0.2	0.9412	0.0571	7.7308
1.8	0.8824	0.0781	4.8930
2.0	0.8235	0.0925	3.4991
3.6	0.7487	0.1103	2.2549
4.0	0.6655	0.1255	1.3182
6.2	0.5324	0.1557	0.2079
7.0	0.3993	0.1641	-0.6137

Table : stage III

t_i	$\hat{S}(t_i)$	$\sqrt{\hat{V}\left[\hat{S}(t_i)\right]}$	$\frac{\hat{S}(t_i)-(1-p)}{\hat{V}^{1/2}\left[\hat{S}(t_i)\right]}$
0.3	0.9259	0.0504	8.4507
0.5	0.8889	0.0605	6.4299
0.7	0.8519	0.0684	5.1465
0.8	0.8148	0.0745	4.2112
1.0	0.7778	0.0800	3.4718
1.3	0.7407	0.0843	2.8545
1.6	0.7037	0.0879	2.3180
1.8	0.6667	0.0907	1.8371
1.9	0.5926	0.0946	0.9792
3.2	0.5556	0.0956	0.5809
3.5	0.5185	0.0962	0.1926
5.0	0.4667	0.0995	-0.3349
6.3	0.4000	0.1053	-0.9496
6.4	0.3333	0.1068	-1.5606
7.8	0.2500	0.1078	-2.3188

Table : stage IV

t_i	$\hat{S}(t_i)$	$\sqrt{\hat{V}\left[\hat{S}(t_i)\right]}$	$\frac{\hat{S}(t_i)-(1-p)}{\hat{V}^{1/2}\left[\hat{S}(t_i)\right]}$
0.1	0.9231	0.0739	5.7246
0.3	0.8462	0.1001	3.4592
0.4	0.7692	0.1169	2.3040
0.8	0.6154	0.1349	0.8551
1.0	0.5385	0.1383	0.2782
1.5	0.4615	0.1383	-0.2782
2.0	0.3846	0.1349	-0.8551
2.3	0.3077	0.1280	-1.5023
3.6	0.2051	0.1196	-2.4662
3.8	0.1026	0.0940	-4.2286

```
##compute the point estimates and confidence interval for the
```

```
##mean and median survival time of the four stages
nonpa.surv = function(value, data, j){
datai = data[stage == j, ]
mtime = max(futime)
##reverse the time and the smallest time is 0
time.death = rev(c(0, value[, 1], mtime))
##compute the integrable value between every time interval
mu = 0
for(i in 1 : (length(time.death) - 1)){
    mu[i] = (time.death[i] - time.death[i + 1]) * rev(c(1, value[, 4]))[i]
}
##point estimate of the mean survival time
mu.est = sum(mu)
sigma = value[, 5]
##the variance of the mean survival time
var.est = sum(rev(cumsum(mu[-length(mu)]) ^ 2) * sigma)
ci.lower = mu.est - 1.96 * sqrt(var.est)
ci.upper = mu.est + 1.96 * sqrt(var.est)
return(list(mu = mu.est, var = var.est, ci.lower = ci.lower, ci.upper = ci.upper))
}
##Compute the median confidence interval
##The output is in the above four tables
ci1 = cbind(km1[, 1], km1[, 4], sqrt(km1[, 6]), (km1[, 4] - 0.5) / sqrt(km1[, 6]))
ci2 = cbind(km2[, 1], km2[, 4], sqrt(km2[, 6]), (km2[, 4] - 0.5) / sqrt(km2[, 6]))
ci3 = cbind(km3[, 1], km3[, 4], sqrt(km3[, 6]), (km3[, 4] - 0.5) / sqrt(km3[, 6]))
ci4 = cbind(km4[, 1], km4[, 4], sqrt(km4[, 6]), (km4[, 4] - 0.5) / sqrt(km4[, 6]))
```