COMPARISON BETWEEN WEIBULL AND COX PROPORTIONAL HAZARDS MODELS

by

ANGELA MARIA CRUMER

B.S., Southeast Missouri State University, 2008

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2011

Approved by:

Major Professor
Dr. James Higgins

Abstract

The time for an event to take place in an individual is called a survival time. Examples include the time that an individual survives after being diagnosed with a terminal illness or the time that an electronic component functions before failing. A popular parametric model for this type of data is the Weibull model, which is a flexible model that allows for the inclusion of covariates of the survival times. If distributional assumptions are not met or cannot be verified, researchers may turn to the semi-parametric Cox proportional hazards model. This model also allows for the inclusion of covariates of survival times but with less restrictive assumptions. This report compares estimates of the slope of the covariate in the proportional hazards model using the parametric Weibull model and the semi-parametric Cox proportional hazards model to estimate the slope. Properties of these models are discussed in Chapter 1. Numerical examples and a comparison of the mean square errors of the estimates of the slope of the covariate for various sample sizes and for uncensored and censored data are discussed in Chapter 2. When the shape parameter is known, the Weibull model far out performs the Cox proportional hazards model, but when the shape parameter is unknown, the Cox proportional hazards model and the Weibull model give comparable results.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I consider it an honor to have worked with Dr. James Higgins. His never ending patience has been the foundation for much learning and development through this research process, and I am ever grateful for this opportunity.

It is with great pleasure that I acknowledge Zhining Ou, who has helped with the computer programming for this research, and my committee members.

I would also like to thank my parents who gave me a great start in life to serve as the springboard to my achievements.

I owe my deepest gratitude to my closest friend in life and academics, Dr. Imad Khamis, who has dedicated much of his own time to help me study and research effectively. His advice and moral support has pushed me to achieve levels of success that without him would have been unattainable.

Dedication

This report is dedicated to my son, Ryan M. Soehlke, whose patience and advanced maturity has made this research and writing possible.

Chapter 1 - Weibull and Cox Proportional Hazard Models

This chapter will introduce parametric and semi-parametric models for survival data analysis.

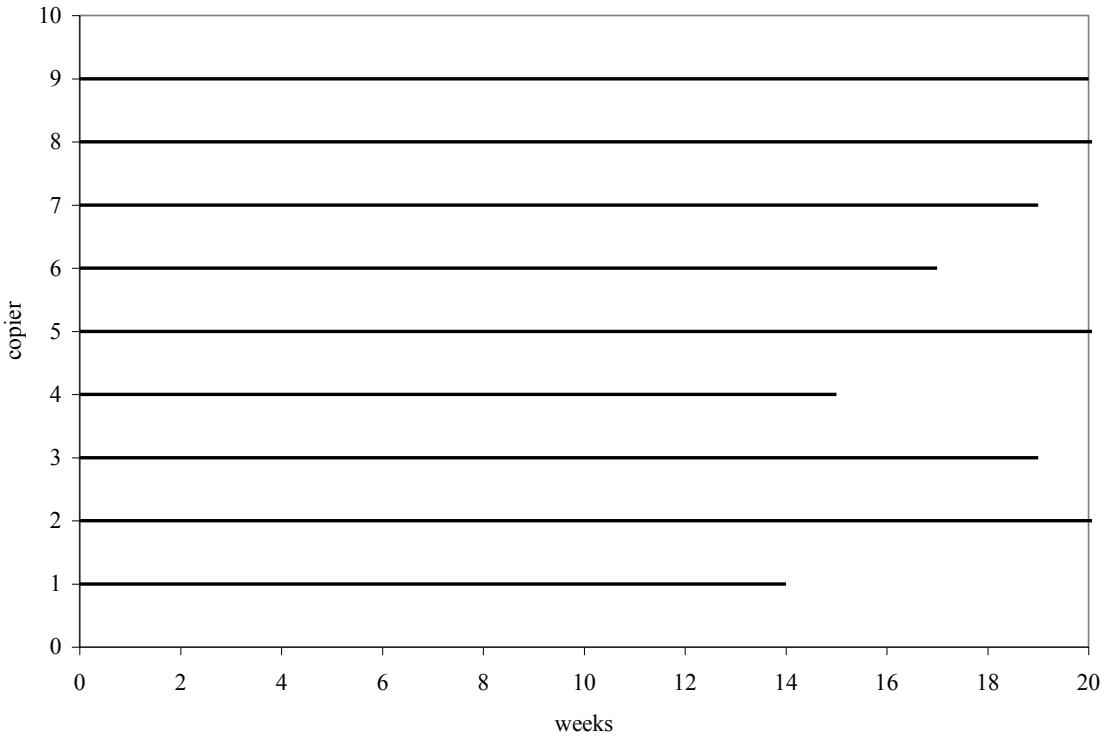
Survival Time and Censoring

To begin a discussion on survival or life data analysis, one must first conceptualize the basic definitions and explanations of survival time and censoring. Survival time is the time it takes for a certain event to take place in a given individual. This could include the survival time for terminally ill patients, amount of time of a treatment before an illness is cured, and length of remission. Survival data analysis has typically been used to study the survival times of human patients with an illness or disease and the survival times of animals in an experiment, but with the rapid influx of technological knowledge, it has many applications in the industrial and business world as well. Some examples of life data analysis in non-medical fields include the amount of time before a malfunction of a mechanical component, lifetime of the battery in a laptop, and the employment time of employees for a certain company.

The analysis of these survival times is best done when all the survival times are known. However there are many instances when this is not the case. Observations in this category are said to be censored data. A terminally ill patient may live to end of the study, or a mechanical component may not malfunction during the times it is being observed. In these cases, the survival times of the observations are not known, but it is known to be at least as long as the time of the study. This is called Type I censoring when all censored data have the same length (Lee, 1992). For example, suppose a company has 9 copy machines in their building. All 9 copiers are observed for 6 months

and the time until a repair is needed is recorded. As Figure 1 shows, copiers 1, 2, 3, 4, 6, 7, and 9 needed repairs after 14, 21, 19, 15, 17, 19, and 20 weeks, respectively while copiers 5 and 8 did not need repairs during this time of observation. Hence copiers 5 and 8 are the *censored* data and the remaining copiers are the *uncensored* data. The survival data would be 14, 21, 19, 15, 24+, 17, 19, 24+, 20, where the plus sign symbolizes a censored observation.

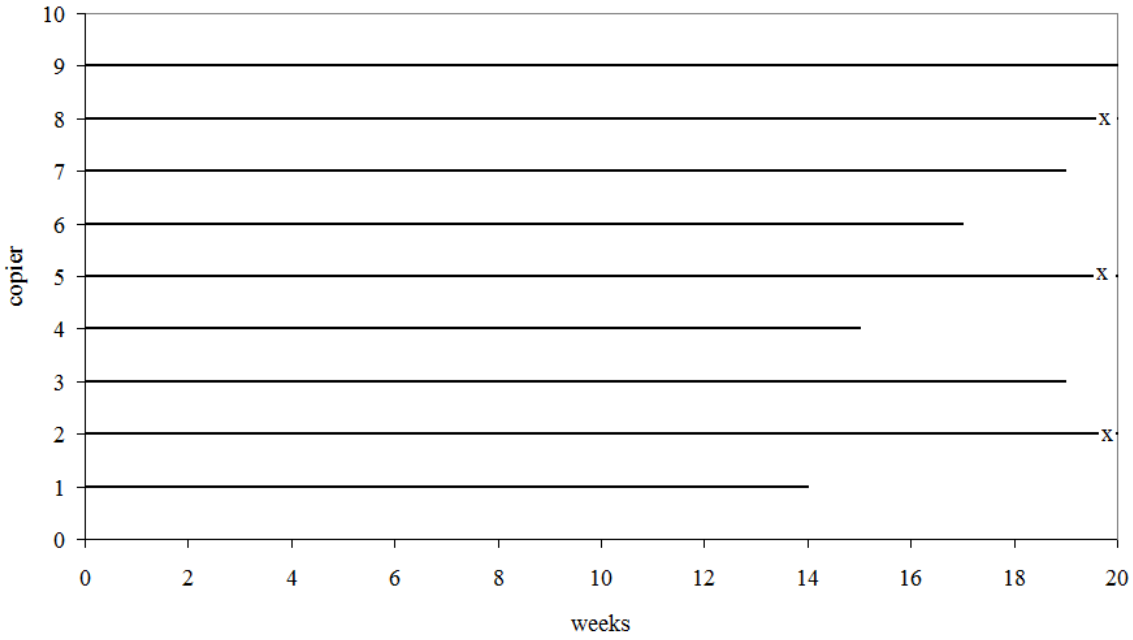
Figure 1 Example of Type I Right Censored Data



Type II censoring is a type of censoring in which all individuals begin at the same time and the study is terminated once a specified number of failures is reached. The remaining observations are then censored to the point at which the longest uncensored observation failed (Lawless, 2003). Using the data from the previous example, consider the situation where the company is interested in the failure times until 6 failures is

reached. The uncensored observations would be copiers 1, 3, 4, 6, 7, and 9 and the survival data would then be 14, 20+, 19, 15, 20+, 17, 19, 20+, 20, shown in Figure 2.

Figure 2 Example of Type II Right Censoring



Type I and II censoring are both types of *right* censoring. Another, less common, type of censoring is *left* censoring in which all individuals do not begin simultaneously. A fourth type of censoring is *random* censoring, also called *non-informative*. This is when a subject has a censoring time that is statistically independent of its failure time, so the observed value is the minimum of the censoring and failure times. Examples of random censoring are common in medical studies where patients may leave the study for reasons unrelated to the treatment such as a family emergency or a change of residence. Those with failure times greater than their censoring time are right-censored.

Survival and Hazard Functions

Two important functions for describing survival data are the survival function and the hazard function. The survival function is the probability that an observation survives longer than t, that is

$$S(t) = P\ (\ T > t\ ).$$

In terms of the cumulative distribution function $F(t)$, the survival function can be written as

$$\begin{aligned} S(t) &= 1 - P\ (\text{ an individual fails before time t }) \\ &= 1 - F(t). \end{aligned}$$

From this, it is easy to see that $S(t)$ is nonincreasing and has the following properties

$$S(t) = 1 \quad \text{for } t = 0$$

$$S(t) \to 0 \quad \text{as } t \to \infty.$$

The survival rate can be depicted using a survival curve, in which a steep curve would indicate a low survival rate and a gradual curve would represent a high survival rate (Lee 1992).

The hazard function is the rate of death/failure at an instant t, given that the individual survives up to time t. It measures how likely an observation is to fail as a function of the age of the observation. This function is also called the instantaneous failure rate or the force of mortality (Nelson, 1982). It is defined as

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

where $f(t)$ is the probability density function of T.

Hence, in terms of the survival function,

$$h(x) = -\frac{d}{dx}\log S(x).$$

Thus,

$$\log S(x) = -\int_0^t h(x)dx\,,$$

and since $S(0) = 1$,

$$S(t) = \exp\left\{\int_0^t h(x)dx\right\}$$

Therefore the pdf of the distribution can be found from the hazard and survival functions,

$$f(t) = h(t)\exp\left\{\int_0^t h(x)dx\right\}.$$

Exponential and Weibull Distributions

Statisticians chose the exponential distribution to model life data because the statistical methods for it were fairly simple (Lawless, 2003). The exponential density function is

$$f(t) = \lambda \exp\{-\lambda t\},\ for\ \lambda > 0\ and\ t > 0.$$

It has a constant hazard function

$$h(t) = \lambda$$

and its survival function is

$$S(t) = \exp\{-\lambda t\}.$$

Thus, a large λ implies a high risk and a short survival. Conversely, a small λ indicates a low risk and a long survival. This distribution has the memoryless property meaning that how long an individual has survived does not affect its future survival (Lee, 1992). It is

used with ordered data, that is, the first individual to fail is the weakest, the second to fail is the second weakest, and so on (Epstein and Sobel, 1953).

The exponential distribution is limited in applicability because it has only one parameter, the scale parameter λ. By adding a shape parameter the distribution becomes more flexible and can fit more kinds of data. The generalization of the exponential distribution to include the shape parameter is the Weibull distribution. The cumulative distribution function of the Weibull distribution is

$$F(t) = 1 - \exp\left\{-\theta t^{\gamma}\right\}, \quad t > 0$$

where θ is the shape parameter and γ is the scale parameter, and the probability density function of the Weibull distribution is

$$f(t) = \gamma \theta t^{\gamma-1} \exp\left\{-\theta t^{\gamma}\right\}, \quad t > 0.$$

The survival function and hazard function of the Weibull distribution are

$$S(t) = \exp\left\{-\theta t^{\gamma}\right\},$$

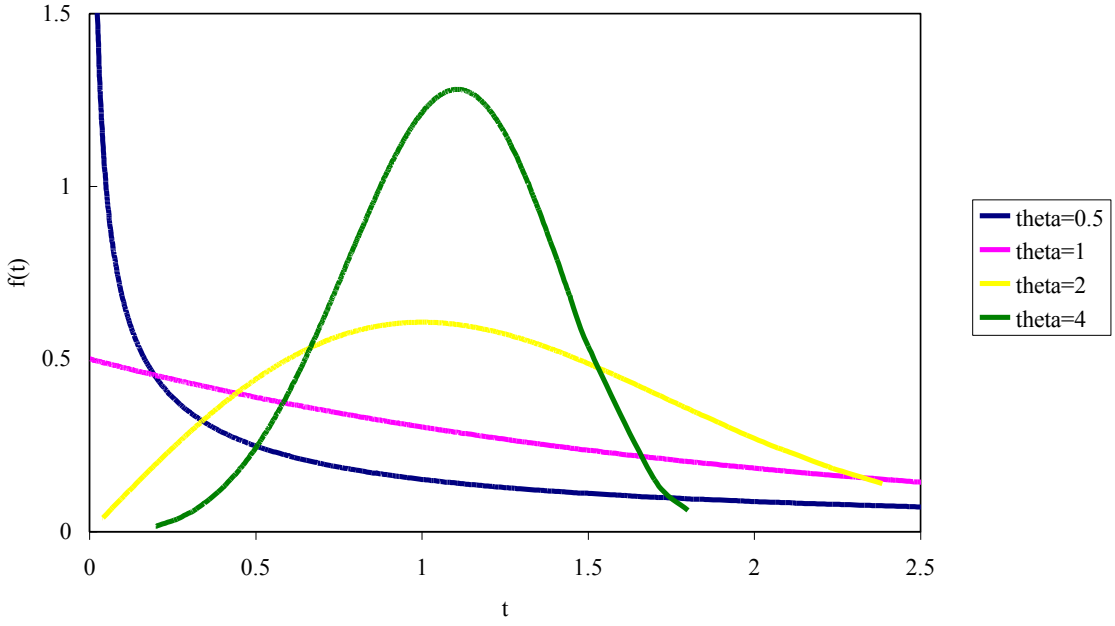
and,

$$h(t) = \gamma \theta t^{\gamma-1}$$

respectively.

It is easy to see just how flexible the Weibull distribution can be. When γ=1, the Weibull distribution becomes the exponential distribution with θ = λ and the hazard rate remains constant as time increases, and when γ=2 it is the Rayleigh distribution. For $3 \leq \gamma \leq 4$, it is close to the normal distribution and when γ is large, say $\gamma \geq 10$ it is close to the smallest extreme value distribution (Nelson, 1982). When $\gamma > 1$ the hazard rate increases as time increases, and for $\gamma < 1$ the hazard rate decreases. Figure 3 shows the Weibull distribution for different values of γ.

Figure 3 Weibull Distribution for Different Shape Parameters



Because of the Weibull distribution's flexibility, it is used for many applications including product life and strength/reliability testing. It models the rate of failure as time increases (Nelson, 1982). It can be shown that the mean and standard deviation are

$$E(T) = \left(\frac{1}{\theta}\right)^{\frac{1}{\gamma}} \Gamma\left(1+\frac{1}{\gamma}\right)$$

and

$$sd = \left(\frac{1}{\theta}\right)^{\frac{1}{\gamma}} \left[\Gamma\left(1+\frac{2}{\gamma}\right) - \Gamma^2\left(1+\frac{1}{\gamma}\right)\right]^{\frac{1}{2}}$$

see appendices for derivations.

Proportional Hazards Model

Sometimes it is helpful to add an explanatory variable, or covariate, to describe the effects of factors which may influence the time-to-failure of an individual. These variables can be continuous such as levels of radiation until remission or indicator variables such as gender. In the Weibull model, covariates can be used to explain some of the variability in θ, the scale parameter, or γ, the shape parameter.

Most often it is useful to consider only the scale parameter as a function of the covariates, that is,

$$\theta = \exp\{\beta_0 + \beta_1 x\} \quad \text{or} \quad \log(\theta) = \beta_0 + \beta_1 x\,.$$

In the engineering context where survival times are of electrical or mechanical components, these two regression are sometimes referred to as the accelerated failure regression models (Lawless, 2003).

Recall the hazard function with no covariates is

$$h(t) = \theta\gamma t^{\gamma-1}.$$

Thus hazard function with covariates is

$$h(t \mid x) = \gamma \exp\{\beta_0 + \beta_1 x\} t^{\gamma-1}$$

The ratio of the hazard functions of two different values of x is given by

$$\frac{h(t \mid x_1)}{h(t \mid x_2)} = \frac{\gamma \exp\{(\beta_0 + \beta_1 x_1)\} t^{\gamma-1}}{\gamma \exp\{(\beta_0 + \beta_1 x_2)\} t^{\gamma-1}} = \exp\{\beta_1 (x_1 - x_2)\}$$

This ratio is a constant proportion that depends only on the covariate and not on time. Thus it is called a *proportional hazards model*. This simple regression model can easily be extended to a multiple regression model by letting

$$\theta = \exp\{\beta_0 + \boldsymbol{\beta x}\},$$

where $\boldsymbol{\beta}$ is a 1 x p vector of coefficients and **x** is a p x 1 vector of explanatory variables.

Cox Proportional Hazards Model

Cox (1972) introduced a model for survival time that allows for covariates but does not impose a parametric form for the distribution of survival times. Specifically he assumed that the survival distribution satisfies the condition

$$h(t \mid x) = h_0(t)\exp\{\beta x\},\ t > 0$$

where x is a covariate, but he made no assumption about the form of $h_0(t)$ which is called the baseline hazard function because it is the value of the hazard function when x = 0. When using a covariate of the form

$$\theta = \exp\{\beta_0 + \beta_1 x\}$$

β_0 is incorporated into the baseline hazard function $h_0(t)$. When x is changed, the conditional hazard functions change proportionally with one another. Hazard functions for any pair of different covariate values *i* and *j* can be compared using a hazard ratio:

$$HR = \frac{h_0(t)\exp\{\beta x_i\}}{h_0(t)\exp\{\beta x_j\}} = \exp\{\beta(x_i - x_j)\} \quad for\ i \neq j$$

Hence, the hazard ratio is a constant proportion and the Cox's model is, indeed, a proportional hazards model. This model is used when the covariates have a multiplicative effect on the hazard function and can be extended for multiple regression situations by allowing

$$h(t|x) = h_0(t)\exp\{\boldsymbol{\beta x}\}.$$

where $\boldsymbol{\beta}$ and **x** to be vectors. It is mostly used in biostatistics.

Chapter 2 - Numerical Examples of Weibull and Cox Models

This chapter will illustrate the use of the Weibull and Cox proportional hazards models with real data sets and will use simulation to compare estimates of the proportional hazards slope parameter.

SAS implementation

The parameters in the Weibull model may be estimated in SAS with the LIFEREG procedure which uses the maximum likelihood estimates. The parameters in the Cox proportional hazards model may be estimated with the PHREG procedure which uses a form of a partial likelihood function proposed by Breslow (1974) as the default option. When calculating parameter estimates, it is important to understand that LIFEREG and PHREG use different parameterizations. The coefficients that are estimated by the two procedures are not the same, but they are related. PROC PHREG uses the model

$$h(t) = h_o(t)\exp\{\beta x\}$$

where *h(t)* is the hazard function and $h_o(t)$ is the baseline hazard function. PROC LIFEREG uses the model

$$T = T^* e^{\delta_0+\delta_1 x}$$

where T is the survival time and T^* is a random variable that has the Weibull survival function

$$S^*(t) = \exp\{-t^\gamma\}$$

In terms of the survival function, the parameterization of the Weibull model for T is

LIFEREG: $S(t) = S^*(t e^{-\delta_0+\delta_1 x}) = e^{-(te^{-\delta_0+\delta_1 x})^\gamma} = (e^{-t^\gamma e^{-\gamma\delta_0}})^{e^{-\gamma\delta_1 x}}$

On the other hand, the parameterization for PHREG gives the following form of the survival function,

PHREG: $S(t) = (e^{-t^{\gamma} e^{-\gamma\delta_0}})^{e^{\beta x}}$

It follows that the relationship between the parameterizations of the Weibull model for these LIFEREG and PHREG is

$$-\gamma\delta_1 = \beta .$$

If $\hat{\delta}_1$ and $\hat{\gamma}$ are estimates of the slope and shape parameters from LIFEREG and $\hat{\beta}$ is the estimate of the slope from PHREG, it follows that $-\hat{\delta}_1\hat{\gamma}$ and $\hat{\beta}$ are estimates of the same parameter which we call "PH-slope". This chapter shows numerical examples of estimates of PH-slope using real data and compares the mean square errors of estimates of this parameter when estimated by the maximum likelihood method for the Weibull model and the Breslow method for the semi-parametric Cox proportional hazards model. Computations are done using LIFEREG and PHREG.

Example 1 Survival Data with a Continuous Covariate using the Weibull Model and Proportional Hazard Model

Survival data from 30 patients with AML is given in Table 1 (Lee, 1992). Age is a continuous covariate and Censor indicates censoring where Censor = 1 is a censored observation. Table 2 shows the SAS code and results when analyzed using PROC LIFEREG in SAS, and Table 3 shows the SAS code and results when analyzed using PROC PHREG in SAS. The estimate of the slope and scale parameters in LIFEREG are -.0261 and .8345, respectively. Using the relationship above, the estimate of PH-slope

from LIFEREG is -(1.1983)×(-.0261) = .0313. This compares to .0266, with standard error 0.01384, which is the estimate of PH-slope from the PHREG procedure.

Table 1 Survival Times vs. Age of 30 AML Patients

Survival Time	Censor	Age	Survival Time	Censor	Age
18	0	35	8	0	72
9	0	42	2	0	60
28	1	33	26	1	56
31	0	20	10	0	61
39	1	22	4	0	59
19	1	45	3	0	69
45	1	37	4	0	70
6	0	19	18	0	54
8	0	44	8	0	74
15	0	26	3	0	53
23	0	48	14	0	66
28	1	32	3	0	64
7	0	21	13	0	54
12	0	51	13	0	60
9	0	65	35	1	68

Table 2 SAS Results LIFEREG

Parameter	DF	Estimate	Std Err	95% CI Limits		Chi-Square	Pr<ChiSq
Intercept	1	4.2454	0.6001	3.0693	5.4215	50.05	<.0001
Age	1	-0.0261	0.0112	-0.048	-0.0043	5.49	0.0192
Scale	1	0.8345	0.1395	0.6013	1.1582		
Shape	1	1.1983	0.2004	0.8634	1.663		

```
proc lifereg;
model surv *censor(1)= age /dist=weibull;
```

Table 3 SAS Results PHREG

Covariate	DF	Estimates	Std Error	Chi-Suare	Pr>ChiSq	Hazard Ratio
age	1	0.02655	0.01384	3.6818	0.055	1.027

```
proc phreg;
model surv *censor(1) = age;
```

Example 2 Survival Data with Continuous Covariate and transformed survival time using the Weibull Model and Proportional Hazard Model

For illustrative purposes, Table 4 shows the results for LIFEREG when analyzing the square root of survival times from Example 1. If the original data are Weibull, the transformed data will also be Weibull but with different slope and scale parameters. The estimate of the PH-slope using PHREG will not change because the estimate using the Breslow partial likelihood depends only on the order of the observations and the pattern of censoring, not the actual survival times.

Table 4 SAS Results LIFEREG

Parameter	DF	Estimate	Std Err	95% CI Limits		Chi-Square	Pr<ChiSq
Intercept	1	2.1227	0.3	1.5347	2.7108	50.05	<.0001
Age	1	-0.0131	0.0056	-0.024	-0.0021	5.49	0.0192
Scale	1	0.4173	0.0698	0.3007	0.5791		
Shape	1	2.3966	0.4007	1.7269	3.326		

The estimate of PH-slope using the results of LIFEREG is - (2.3966)×(-.0131) = .0314 which except for rounding is the same as obtained by the analysis of the original survival data, showing that the semi-parametric model does not depend on the shape parameter.

Example 3 Survival Data with Categorical Covariates using the Weibull Model and Proportional Hazard Model

Using the same survival times, Table 5 shows age now recorded as a discrete variable; age = 1 if patient ≥ 50 years, 0 if patent < 50 years, and clot = 1 if cellularity of marrow clot section is 100%, 0 otherwise. Table 6 shows the results from SAS using LIFEREG and Table 7 shows the results from SAS using PHREG.

Table 5 Survival Times vs. Age and Clot of 30 AML Patients							
Survival Time	Censor	Age	Clot	Survival Time	Censor	Age	Clot
18	0	0	0	8	0	1	0
9	0	0	1	2	0	1	1
28	1	0	0	26	1	1	0
31	0	0	1	10	0	1	1
39	1	0	1	4	0	1	0
19	1	0	1	3	0	1	0
45	1	0	1	4	0	1	0
6	0	0	1	18	0	1	1
8	0	0	1	8	0	1	1
15	0	0	1	3	0	1	1
23	0	0	0	14	0	1	1
28	1	0	0	3	0	1	0
7	0	0	1	13	0	1	1
12	0	1	0	13	0	1	1
9	0	1	0	35	1	1	0

Table 6 SAS Results LIFEREG							
Parameter	DF	Estimate	Std Err	95% CI Limits		Chi-Square	Pr<Chi-Sq
Intercept	1	2.3514	0.2677	1.8267	2.876	77.16	<.0001
age	1	-1.0191	0.366	0.3018	1.7364	7.75	0.0054
clot	1	-0.3838	0.3517	-0.3056	1.0732	1.19	0.2752
scale	1	0.8034	0.1363	0.5762	1.1202		
shape	1	1.2448	0.2111	0.8927	1.7356		

```
proc lifereg;
model surv *censor (1)= age clot /dist=weibull
```

Table 7 SAS Results PHREG						
Covariate	DF	Estimates	Std Error	Chi-Suare	Pr>ChiSq	Hazard Ratio
age	1	1.01317	0.4574	4.9065	0.0268	2.754
clot	1	0.35025	0.43917	0.636	0.4252	1.419

```
proc phreg;
model surv *censor(1) = age clot;
```

If we multiply the estimates of the coefficients of age and clot by negative the shape parameter, we obtain the LIFEREG estimates of the proportion hazards parameters, that is, - (1.2448)×(-1.0191) = 1.269 for age and - (1.2448)×(-.3838) = .4778 for clot. This compares to the estimates of 1.0132 and .3503, with standard errors 0.4574 and 0.43917, for the respective estimates of these parameters from PHREG.

Simulation

Is there an advantage to using a parametric form of the survival distribution instead of the semi-parametric Cox proportional hazards model in estimating the effect of a covariate of survival time when the parametric form of the model is known? To investigate this question, a simulation study was done to compare the mean square errors of the Weibull maximum likelihood estimate and the Cox proportional hazards model estimate of β = PH-slope when data come from a Weibull model.

The data were simulated from a Weibull model with survival function

$$S(t) = (e^{-t^2})^{e^x}$$

That is, the model is Weibull with $\beta = 1$ for the slope of the covariate x, shape parameter $\gamma = 2$, and baseline survival function $S_0(t) = e^{-t^2}$. The values of the covariate are x = 1, 2, 3, 4, and 5. The total sample sizes are 15, 30, and 90 with 3, 6, or 18 observations for each value of x. The data were simulated using the fact that the random variable U = F(T) has a uniform distribution where T is a Weibull random variable with cumulative distribution function F(t). For this study, a value of T was obtained as

$$T = ((-\log(1-U))e^{-x})^{1/2}$$

The uniform random variable was generated using the SAS random number generator. Data were simulated without censoring and with twenty percent random censoring. With

random censoring a uniform variable U* was generated independently of U and an observation was denoted as censored if U* ≤ .20.

The maximum likelihood estimate of PH-slope using the parametric Weibull model was obtained from LIFEREG as $\hat{\beta} = -\hat{\gamma}\hat{\delta}_1$ where $\hat{\gamma}$ is the estimate of the shape parameter and $\hat{\delta}_1$ is the estimate of the slope of the Weibull model as parameterized in LIFEREG. Since the shape parameter is known to be 2, an estimate of PH-slope that takes advantage of this fact, $-2\hat{\delta}_1$, was also obtained. The estimate of PH-slope from the Cox proportional hazards model was computed using PHREG.

One-thousand replications of each sample size were run and the mean square estimated as

$$\sum_{i=1}^{1000} \frac{(\hat{\beta}_i - \beta)^2}{1000}$$

where β = 1. The standard error of the mean square error was computed as the standard deviation of the squared deviations $(\hat{\beta}_i - \beta)^2$, i = 1, 2,…1000, divided by the square root of 1000. The distributions of $\hat{\beta}$ from the maximum likelihood estimates of the Weibull parameters and from the Cox proportional hazards model do not depend on the value of the shape parameter γ. Thus, the mean square errors apply to all Weibull shape parameters.

Results for Complete Samples

Table 8 has the means square errors for the complete sample case. Here it can be seen that when the shape parameter is unknown, the estimates for the Cox proportion hazards model and the maximum likelihood estimates of the Weibull model perform

similarly, but when the shape parameter is known, the estimate $-2\hat{\delta}_1$ far out-performs the Cox proportional hazards model. From this, it can be concluded that when the distributional assumptions are not known, or are not met, the Cox proportional hazards model should be used.

Table 8 MSEs and Standard Errors for Complete samples			
	N=15	N=30	N=90
Weibull mle shape known	0.0381(.0016)	0.0180 (.0008)	0.0059 (.0002)
Weibull MLE shape unknown	0.1616 (.0158)	0.0527 (.0029)	0.0140 (.0007)
Cox PH model estimate	0.1494 (.0098)	0.0547 (.0031)	0.0143 (.0009)

Results for Censored Samples

Results for the censored sample case are shown in Table 9. The patterns are similar to the complete sample case. The MSEs are similar for the maximum likelihood estimates and the proportional hazards model estimates when the shape parameter is unknown, but much smaller for the maximum likelihood estimates of the Weibull model when the shape parameter is known. As would be expected, the MSEs for censored data are larger than uncensored data, but not appreciably so, except in one notable case. The small sample case, N=15, the Cox PH model occasionally produces unusual estimates, sometimes very large, in both the uncensored and censored cases yielding inconsistent MSE calculations. This problem is exacerbated in the presence of censored data, but is not present in either case for larger sample sizes. Although the Cox model is generally comparable to the Weibull model, perhaps it is not for small sample sizes.

Table 9 MSEs and standard errors of censored data			
	N=15	N=30	N=90
Weibull MLE shape known	0.0484 (.0024)	0.0247 (.0043)	0.0067 (.0003)
Weibull MLE shape unknown	0.1876 (.013)	0.0709 (.0039)	0.0194 (.0011)
Cox PH model estimate	0.1385 (.012)	0.0756 (.0043)	0.0183 (.0009)

Conclusions

In conclusion, the Weibull model is the best option for analyzing lifetime data if the distributional assumptions can be met and the shape parameter is known. The mean square errors are smallest in this case. However, when the shape parameter is unknown, the Cox proportional hazards model is a good alternative. It requires fewer assumptions than the parametric Weibull model and provides comparable mean square errors of the estimates of PH-slope. There may be a concern for smaller samples with the Cox proportional hazards model depending on the particular data set being analyzed.

References

Breslow N. (1974). Covariance analysis of censored survival data. *Biometrics*, *30,* 89-99.

Cox, D. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society, 34*, 187–220.

Epstein, B. and Sobel, M. (1953). Life Testing. *Journal of the American Statistical Association, 48(263),* 486-502.

Lawless, J. (2003). Statistical Models and Methods for Lifetime Data. Hoboken: John Wiley & Sons, Inc.

Lee, E. (1992). Statistical Methods for Survival Data Analysis. New York: John Wiley & Sons, Inc.

Nelson, W. (1982). Applied Life Data Analysis. New York: John Wiley & Sons, Inc.

Appendix 1

Probability Density and Likelihood Functions of Weibull Distribution

The probability density function of the Weibull random variable T is

$$f(t) = \frac{d}{dt}\left(1 - \exp\left\{-\theta t^{\gamma}\right\}\right) = \left(-\exp\left\{-\theta t^{\gamma}\right\}\right)\left(-\theta\gamma t^{\gamma-1}\right) = \theta\gamma \exp\left\{-\theta t^{\gamma}\right\} t^{\gamma-1}$$

In the case of the complete sample with observed survival times $t_1, t_2, ..., t_n$, the log likelihood function is

$$\log L(\theta, \gamma) = n\log(\theta) + n\log(\gamma) + (\gamma - 1)\sum \log t_i - \sum \theta t_i^{\gamma}$$

The partial derivative of the log likelihood function with respect to θ depends on the survival times through the quantities t_i^{γ}. The random variable $W = T^{\gamma}$ has an exponential distribution

$$f_W(w) = \theta \exp\{-\theta w\},\ w > 0.$$

Thus the distribution of the maximum likelihood estimate of θ depends only on the distribution of *W*, which is independent of γ. If $\log(\theta) = \beta_0 + \beta_1 x$, then the distributions of the maximum likelihood estimates of β_0 and β_1 will not depend on γ.

In the case of the sample with censored data, the likelihood function is

$$\prod_{t_i} f(t_i)^{1-c}\left(1 - F(t_i)\right)^{c},$$

where $c = 0$ if the time is not censored, and $c = 1$ if the time is censored. From the form of $f(t)$ and $F(t)$ for the Weibull distribution, it can be seen that the distribution of the maximum likelihood estimate of θ does not depend on γ, as in the case of the complete sample.

Mean and Variance of Weibull Distribution

The following steps show the derivations of the mean and variance of the Weibull random variable T.

$$E(T) = \int_0^\infty t(\theta\gamma)\exp\{-\theta t^\gamma\}t^{\gamma-1}dt$$

$$Let\ u = \theta t^\gamma \Rightarrow t = \left(\frac{u}{\theta}\right)^{\frac{1}{\gamma}}.$$

$$Then,\ du = \theta\gamma t^{\gamma-1}dt.$$

$$E(T) = \int_0^\infty \left(\frac{u}{\theta}\right)^{1/\gamma} \exp\{-u\}du = \left(\frac{1}{\theta}\right)^{1/\gamma}\int_0^\infty u^{1/\gamma}\exp\{-u\}du$$

$$= \left(\frac{1}{\theta}\right)^{1/\gamma}\int_0^\infty u^{\left(1+1/\gamma\right)-1}\exp\{-u\}du = \left(\frac{1}{\theta}\right)^{1/\gamma}\Gamma\left(1+\frac{1}{\gamma}\right)$$

$$E(T^2) = \int_0^\infty t^2(\theta\gamma)\exp\{-\theta t^\gamma\}t^{\gamma-1}dt$$

$$Let\ u = \theta t^\gamma \Rightarrow \left(\frac{u}{\theta}\right)^{1/\gamma}.$$

$$Then,\ du = \theta\gamma t^{\gamma-1}dt.$$

$$E\left(T^2\right) = \int_0^\infty \left(\frac{u}{\theta}\right)^{2/\gamma} \exp\{-u\}du = \left(\frac{1}{\theta}\right)^{2/\gamma}\int_0^\infty u^{\left(1+2/\gamma\right)-1}\exp\{-u\}du$$

$$= \left(\frac{1}{\theta}\right)^{2/\gamma}\Gamma\left(1+\frac{2}{\gamma}\right)$$

$$Var(T) = \left(\frac{1}{\theta}\right)^{\frac{2}{\gamma}}\Gamma\left(1+\frac{2}{\gamma}\right) - \left[\left(\frac{1}{\theta}\right)^{\frac{1}{\gamma}}\Gamma\left(1+\frac{1}{\gamma}\right)\right]^2$$

$$= \left(\frac{1}{\theta}\right)^{\frac{2}{\gamma}}\Gamma\left(1+\frac{2}{\gamma}\right) - \left(\frac{1}{\theta}\right)^{\frac{2}{\gamma}}\Gamma^2\left(1+\frac{1}{\gamma}\right) = \left(\frac{1}{\theta}\right)^{\frac{2}{\gamma}}\left[\Gamma\left(1+\frac{2}{\gamma}\right) - \Gamma^2\left(1+\frac{1}{\gamma}\right)\right]$$

Appendix 2

Log Transformation of Weibull Distribution and Cumulative Distribution Function for Extreme Value Distribution

Let T have a Weibull distribution with cdf

$$F_T(t) = 1 - \exp\{-\theta t^{\gamma}\}.$$

Let $Y = \log(T)$. The cdf of Y is

$$F_Y(y) = P(Y \le y) = P(\log(T) \le y) = P(T \le \exp\{y\}) = 1 - \exp\{-\theta \exp\{y\gamma\}\}.$$

To put the distribution in the form of a location-scale model, let $\theta = \exp\{-\mu\gamma\}$ and $\sigma = \frac{1}{\gamma}$. With this

$$F_Y(y) = 1 - \exp\{-\exp\{-\mu\gamma\}\exp\{y\gamma\}\} = 1 - \exp\left\{-\exp\left\{(y-\mu)/\sigma\right\}\right\}.$$

Probability Density Function for Extreme Value Distribution

$$f(y) = \frac{d}{dy}\left[1 - \exp\left\{-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\}\right] = \left(-\exp\left\{-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\}\right)\left(-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right)\left(\frac{1}{\sigma}\right)$$

$$= \frac{1}{\sigma}\exp\left\{-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\}\exp\left\{\frac{y-\mu}{\sigma}\right\}$$

Mean of Extreme Value Distribution

$$E(Y) = \frac{1}{\sigma}\int_{-\infty}^{\infty} y \exp\left\{\frac{y-\mu}{\sigma}\right\}\exp\left\{-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\}dy$$

$$Let\ u = \exp\left\{\frac{y-\mu}{\sigma}\right\} \Rightarrow y = \sigma \log u + \mu \rightarrow 0 < u < \infty.$$

$$Then\ \ du = \frac{1}{\sigma}\exp\left\{\frac{y-\mu}{\sigma}\right\}dy.$$

$$E(Y) = \int_{0}^{\infty}(\sigma \log u + \mu)\exp\{-u\}du$$

$$= \int_{0}^{\infty}\sigma \log u \exp\{-u\}du + \int_{0}^{\infty}\mu \exp\{-u\}du$$

$$= \sigma\int_{0}^{\infty}\log u \exp\{-u\}du + \mu\int_{0}^{\infty}\exp\{-u\}du$$

$$= -\sigma\left[-\int_{0}^{\infty}\log u \exp\{-u\}du\right] + \mu\int_{0}^{\infty}\exp\{-u\}du$$

$$= -\sigma\gamma - \mu \exp\{-u\}|_{0}^{\infty}$$

where γ is the Euler-Mascheroni constant, 0.57722.

$= \mu$ - 0.57722σ

Standard Deviation of Extreme Value Distribution

$$E\left(Y^2\right)=\int_{-\infty}^{\infty} x^2\left(\frac{1}{\sigma}\right)\exp\left\{\frac{y-\mu}{\sigma}\right\}\exp\left\{-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\}$$

$$Let\ u=\exp\left\{\frac{y-\mu}{\sigma}\right\}\Rightarrow x=\sigma\log u+\mu\rightarrow 0<u<\infty.$$

$$Then\ du=\frac{1}{\sigma}\exp\left\{\frac{y-\mu}{\sigma}\right\}du.$$

$$E\left(Y^2\right)=\int_0^{\infty}\left(\sigma\log u+\mu\right)^2\exp\{-u\}du$$

$$=\int_0^{\infty}\sigma^2\left(\log u\right)^2\exp\{-u\}du+\int_0^{\infty}2\sigma\left(\log u\right)\exp\{-u\}du+\int_0^{\infty}\mu^2\exp\{-u\}du$$

$$=\sigma^2\int_0^{\infty}\left(\log u\right)^2\exp\{-u\}du+2\sigma\gamma+\mu^2$$

$$=\sigma^2 I_2+2\sigma\gamma+\mu^2$$

where I_2 is an Euler-Mascheroni integral and γ is the Euler-Mascheroni constant.

$$=\sigma^2\left(\gamma^2+\frac{1}{6}\pi^2\right)+2\sigma\gamma+\mu$$

$$=\left(\sigma^2\gamma^2+2\sigma\gamma+\mu\right)+\frac{1}{6}\pi^2$$

$$\left(\gamma\sigma+\mu\right)^2+\frac{1}{6}\pi^2\sigma^2$$

So,

$$Var(Y)=\left(.57722\sigma+\mu\right)^2+\frac{1}{6}\pi^2\sigma^2-\left(.57722\sigma+\mu\right)^2$$

$$=\frac{1}{6}\pi^2\sigma^2$$

and,

$$sd=\sqrt{Var(y)}=\frac{1}{\sqrt{6}}\pi\sigma$$

Appendix 3

Estimate of PH-Slope Using Cox Proportional Hazards Model

The Breslow form of the Cox partial likelihood for a single covariate x with values $x_1, x_2, \ldots, x_n$ is

$$L(\beta) = \prod \frac{\exp\{\beta x_i\}}{\sum \exp\{\beta x_i\}}$$

where the product is taken over all uncensored times $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ and the sum for each $t_{(i)}$ is taken over the individuals whose survival time is at least $t_{(i)}$. The estimate of β does not depend on the actual survival times, only the orders of the times and the pattern on censoring.

Appendix 4

Simulation SAS Code for N=15, Uncensored

```
title 'Uncensored Data N=15';
dm "output;clear;log;clear;";
ODS TAGSETS.EXCELXP
file='\\statsrvr\home\asoehlke\My Documents\data15.xls'
STYLE=sasweb
OPTIONS ( Orientation = 'landscape');
%macro one;
%do k = 1 %to 1000 %by 1;

data c;
do x = 1 to 5 by 1;
      do i = 1 to 3 by 1;
      F=rand('uniform');
      t=(-exp(-x)*LOG(1-F))**(1/2);
      output;
      end;
end;
run;


ods trace on;
proc lifereg data=c;
model t=x / dist=weibull;
ods select ParameterEstimates;
run;
ods trace off;

ods trace on;
proc phreg data=c;
model t=x;
ods select ParameterEstimates;
run;
ods trace off;

%end;
%mend;

%one;

ods tagsets.excelxp close;

run;
quit;
```

Simulation SAS Code for N=15, Censored

```
title 'Censored Data N=15';
dm "output;clear;log;clear;";
ODS TAGSETS.EXCELXP
file='\\statsrvr\home\asoehlke\My Documents\censored15.xls'
STYLE=sasweb
OPTIONS ( Orientation = 'landscape');
%macro one;
%do k = 1 %to 1000 %by 1;

data c;
do x = 1 to 5 by 1;
        do i = 1 to 6 by 1;
        F=rand('uniform');
        t=(-exp(-x)*LOG(1-F))**(1/2);
            rand_number = rand('uniform');
            If rand_number <=.2 then censor = 1;
            else censor = 0;
            output;
        end;
end;
run;



ods trace on;
proc lifereg data=c;
model t*censor(1) =x / dist=weibull;
ods select ParameterEstimates;
run;
ods trace off;

ods trace on;
proc phreg data=c;
model t*censor(1) = x;
ods select ParameterEstimates;
run;
ods trace off;

%end;
%mend;

%one;

ods tagsets.excelxp close;

run;
quit;
```