The relationship between the parameterizations of the Weibull model and the Cox hazards model by using application and simulation

م. علي لطيف عارض
جامعة ذي قار / كلية العلوم

الخلاصة :

الوقت لحدوث أي حادث فردي يدعى بوقت البقاء ويعتبر توزيع ويبل النموذج ألمعلمي الملائم لهذا النوع من البيانات , والذي يعتبر أنموذجا مرنا يسمح لإدراج المتغيرات اللازمة لوقت البقاء . كذلك فان أنموذج وجه الخطر النسبي والذي يسمح لإدراج المتغيرات يعتبر أيضا من نماذج وقت البقاء . في هذا البحث تمت مقارنة مقدرات الميل الخاصة بالأنموذجين وباستعمال بيانات حقيقية تمثل (38) مريض يعانون من مرض الصداع المزمن .تم تقدير المعلمات الخاصة بأنموذج ويبل باستعمال طريقة الإمكان الأعظم بينما استعملت طريقة بريسلو أو الإمكان الأعظم الجزئي للأنموذج الثاني .وعندما تكون معلمة الشكل الخاصة بنموذج ويبل معلومة نجد إن هناك علاقة ارتباط بين تقديرات الأنموذجين وهي ($-\beta\alpha_1 = \gamma$) , كذلك وباستعمال المحاكاة لتوليد متغيرات عشوائية لتوزيع ويبل تم إثبات نفس العلاقة , و لتحليل النتائج استعمل البرنامج الإحصائي (SAS 9.1) .

Abstract :

The time for an event to take place in an individual is called a survival time. familiar parametric model for this type of data is the Weibull model, which is flexible model that allows for the inclusion of covariates of the survival times , Cox proportional hazards is another model also allows for the inclusion of covariates of survival times ,the research compares estimates of the slope of the covariate in the tow models with real data for (38) patients to estimate the parameters of the models by used the maximum likelihood method for the Weibull and the Breslow method for the semi-parametric Cox proportional hazards model. When the shape parameter is known, the estimate of tow models are related with the relation ($-\beta\alpha_1 = \gamma$) .Also by using the simulation to proving this relation the data was analysis by using the program (SAS 9.1) .

Introduction

Survival analyses are concerned with the occurrence of events over time. and refers to the techniques used to study the time to occurrence of some event in a population, and is often called time - to event analysis and concerned with studying the time between entry to a study and a subsequent event. For example, Cancer studies are frequently concerned with the risk of relapse or death. In demographic studies, the interesting event might be household migration or the birth of a child. ends before all subjects have been observed some observations are cut off or censored.

Originally the analysis was concerned with time from treatment until death applications of life time distribution methodology range from investigations of the durability of manufactured items to study of human diseases and their treatment. And so, life time distribution methodology is widely used in the biomedical and engineering sciences. The analysis tries too predicting relapse from a set of independent variables describing treatment type, initial disease status, demographics, etc.A fundamental concept for understanding survival models is the hazard function h(t).and means that the patient faces the greatest risk early more survival implies less risk. And survival analysis concerned with studying the time between entry to a study and a subsequent event. To study survival analysis, we must introduce some notation and concepts for describing the distribution of time to event for a population of

individuals. Let T be denote the time to the event of our interest. T is non - negative random variable $(T \geq 0)$ which has to be unambiguously defined ; that is we must be very specify about the start and end with the length of the time period in between corresponding to T.[1] [2]

Weibull distribution

This distribution is often used to model the time until failure of many different purpose . The parameters in the distribution provide a great deal of flexibility to model systems in which the number of failures. The random variable x with probability density function [5]

$$f(t,\beta,\theta) = \frac{\beta}{\theta} t^{\beta-1} \exp(-\frac{x^{\beta}}{\theta}) \quad \text{for } t > 0 \quad \text{---------- (1)}$$

Is Weibull random variables with Scale parameter $\theta > 0$, and Shape parameter $\beta > 0$.
The cumulative distribution function is often used to compute probabilities. The following result can be obtained.

$$F(t,\beta,\theta) = 1 - \exp(-\frac{t^{\beta}}{\theta}) \quad \text{---------- (2)}$$

When $\beta = 1$, the Weibull distribution is identical to the exponential distribution. and the hazard rate remains constant as time increases, and when $\beta = 2$ it is the Rayleigh distribution.
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 It can be shown that the mean and variance are:

$$Mean = \theta^{1/\beta}\Gamma(1+(1/\beta)) \quad \text{---------- (3)}$$

$$Variance = (\theta^{1/\beta})^2(\Gamma(1+(2/\beta)) - \Gamma^2(1+(1/\beta))) \quad \text{---------- (4)}$$

The hazard function for Weibull distribution with the two parameters is

$$h(t) = \frac{\beta}{\theta} t^{\beta-1} \quad \text{---------- (5)}$$

Where ***t*** represent the random variable for time to occurrence of failure
And survival time is

$$S(t) = \exp(-\frac{t^{\beta}}{\theta}) \quad \text{---------- (6)}$$

Methods of estimation

There are several methods to estimation the parameters of Weibull distribution some of them are traditional dependents upon the parameters are constants but did not variables such as

1. Maximum Likelihood (ML).
2. Method of moments (MM).
3. Least squared method (LSM).

The second method dependents upon assumption advance information about the parameters that we want to estimates to might formulation as function called (prior pdf) such as

1. Baysian method.
2. Shrunken method.
3. Weighted Baysian.
4. White method.

The method of maximum likelihood estimated for weibull model is: [4]

$$L(t_1,\ldots\ldots,t_n,\theta,\beta)=\prod\frac{\beta}{\theta}t^{\beta-1}e^{-\frac{t^\beta}{\theta}}$$

$$=\frac{\beta^n}{\theta^n}e^{-\frac{\sum t_i^\beta}{\theta}}\prod t_i^{\beta-1}$$

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

$$\frac{\partial\ln L}{\partial\beta}=\frac{n}{\hat{\beta}}-\frac{\sum_{i=1}^{n}t_i^{\hat{\alpha}}\ln t_i}{\theta\wedge}+\sum_{i=1}^{n}\ln t_i=0 \qquad \text{---------- (7)}$$

$$\frac{\partial\ln L}{\partial\theta}=-\frac{n}{\hat{\theta}}+\frac{\sum_{i=1}^{n}t_i^{\hat{\alpha}}}{\hat{\theta}^2}=0 \qquad \text{---------- (8)}$$

Survival and Hazard function

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

$S(t) = P\ (T > t\ ).$ ---------- (9)

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

$S(t)$ = 1 – P (an individual fails before time t)

= 1 – $F(t)$. ---------- (10)

From this, it is easy to see that $S(t)$ is no increasing and has the following properties

$S(t) = 1$ *for* $t = 0$

$S(t) \to 0$ *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 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 It is defined as [3]

$$h(t)=\frac{f(t0}{1-F(t)}=\frac{f(t)}{S(t)} \qquad \text{---------- (11)}$$

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

Hence, in terms of the survival function,

$$h(x)=-\frac{d}{dx}\log S(x) \qquad \text{---------- (12)}$$

Thus

$$\log S(x)=-\int_0^t h(x)dx$$ And since S(0)=1

$$S(x)=\exp(\int_0^t h(x)dx) \qquad \text{---------- (13)}$$

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

$$S(t) = h(t)\exp(\int_0^t h(x)dx)$$ ---------- (14)

Cox proportional Hazard model

One of the most important statistical models in medical research is the Cox proportional hazards model, Cox 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 [8]

$$h(t,x) = h_0(t)\exp(\gamma x) \quad , x > 0$$ ---------- (15)

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.

$$HR_{i,j} = \frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp(\gamma_1 x_{i1} + \ldots\ldots\ldots\ldots + \gamma_k x_{ik})}{h_0(t)\exp(\gamma_1 x_{j1} + \ldots\ldots\ldots\ldots + \gamma_k x_{jk})} = \exp(\gamma_1(x_{i1} - x_{j1}) + \ldots..\gamma_k(x_{ik} - x_{jk}))$$ ---------(16)

Hence, the hazard ratio is a constant .and should be strictly parallel [6].

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

$$L(\gamma) = \prod \frac{\exp(\gamma x_i)}{\sum \exp(\gamma x_i)}$$ ---------- (17)

Where the product is taken over all uncensored times $t_1 < t_2 < \ldots\ldots < t_k$, The estimate of β does not depend on the actual survival times[7]

Relationship between the estimations

The parameters in the Weibull model will be estimated by uses the maximum likelihood estimates. And the parameters in the Cox proportional hazards model will be estimated by 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 different parameterizations. The coefficients that are estimated by the two procedures are not the same, but they are related. uses the model

$$h(t,x) = h_0(t)\exp(\gamma x)$$

where $h(x$)is the hazard function and $h_0(x)$ is the baseline hazard function

uses the model

$$T = T^* e^{\alpha_0 + \alpha_1 x}$$ ---------- (18)

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

$$S(t)^* = \exp(-t^\beta)$$ ---------- (19)

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

$$S(t) = S^*(te^{-(\alpha_0+\alpha_1 x)}) = e^{-(te^{-(\alpha_0+\alpha_1 x)})^\beta} = (e^{-t^\beta e^{-\beta\alpha_0}})^{e^{-\beta\alpha_1 x}}$$ ----------(20)

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

$$S(t) = (e^{-t^\beta e^{-\beta\alpha_0}})^{e^{\gamma x}}$$ ----------(21)

It follows that the relationship between the parameterizations of the Weibull model and Cox proportion hazard as :

$$-\beta\alpha_1 = \gamma$$ ----------(22)

If $\hat{\alpha}_1$ and $\hat{\beta}$ are estimates of the slope and shape parameters and $\hat{\gamma}$ is the estimate of the Cox proportion then $-\hat{\beta}\hat{\alpha}_1$ and $\hat{\gamma}$ are estimates of the same parameter which we call "PH-slope".

Application:

.The example was Survival data is given in Table (1) from Thirty patients are different pain relievers are assigned The outcome reported is the time in minutes until headache relief. Age is a continuous covariate and Censor indicates censoring where Censor = 1 is a censored observation.

Table 1 : Data for 38 patients in headache relief.

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**	**64**
9	**0**	**65**	**35**	**1**	**63**

Table 2: shows the estimate of Weibull distribution

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	3.0743	0.1258	2.8277	3.3210	596.90	<.0001
age	**1**	**0.0023**	0.0031	-0.0038	0.0084	0.54	0.4638
Scale	1	0.1795	0.0350	0.1225	0.2632		
Weibull Shape	**1**	**5.5697**	1.0870	3.7994	8.1650		

Table 3: shows the estimate of Cox proportional hazards model.

Analysis of Maximum Likelihood Estimate						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	**-0.02036**	**0.01730**	1.3839	0.2394	0.980

The estimate of the slope and shape parameters from (Table 2) are $\hat{\beta}$ = 0.0023 And $\hat{\alpha}_1$ = 5.5697, respectively. Using the relationship above, the estimate of PH-slope is

$(-\hat{\beta}\hat{\alpha}_1)$ = -(0.0023)×(5.5697) = -0 .01

This compares to $\hat{\gamma}$ = -.02036, with standard error 0.01730

Simulation

Simulation studies represent an important statistical tool to investigate the performance, of statistical models, test statistics and estimation techniques considering is pre-specified conditions. The Cox model and the corresponding partial likelihood are intensively investigated by means of simulation studies to get information about efficiency of the estimated regression coefficients for a variety of situations, in particular when fundamental model assumptions are violated.

The Cox proportional hazards model is given by

$$h(x) = h_0(x)\exp(\beta x)$$

Where *t* is the time, *x* the vector of covariates, β the vector of regression coefficients and $h_0(t)$ is the so-called baseline hazard function, i.e. the hazard function under *x*=0. Because the model is formulated through the hazard function, the simulation of appropriate survival times for the Cox model is not straightforward. One important issue in simulation studies regarding regression models is the knowledge of the true regression coefficients.. However, in the Cox model, the effect of the covariates have to be translated from the hazards to the survival times, because the usual software packages for Cox models require the individual survival time data, not the hazard function. The translation of the regression effects from hazard to survival time is easy if the baseline hazard function is constant, i.e.. This may be the reason why most simulation studies regarding the Cox model and Another frequently used distribution for survival times is the Weibull distribution . The Weibull parameters can be chosen such that the hazards are proportional and the true hazard ratio (HR) for the comparison of the two groups can be calculated from the Weibull parameters.

a simulation study was done to explains the advantage for 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

1. The data were simulated using generating the random numbers from uniform distribution .
2. Generating the random variable T from Weibull distribution
   $$T = ((-\log(1-U))e^{-x})^{1/\beta} \ .$$
3. Given a value to the shape parameter of Weibull distribution such as β = 2 The values of the covariate are x = 1, 2, 3, 4, and 5.
4. The total sample sizes are 15, 30, and 90 .
5. One-thousand replications of each sample size were run .

Table 4 : shows the estimate of Weibull distribution by using simulation

Analysis of Parameter Estimates						
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-0.4380	0.2210	-0.8711 0.0050	3.93	0.0474
x	1	**-0.4331**	0.0661	-0.5626 -0.3036	42.97	<.0001
Scale	1	0.3557	0.0798	0.2291 0.5520		
Weibull Shape	1	**2.8117**	0.6307	1.8114 4.3642		

Table 5 : shows the estimate of Cox proportional hazards model by using simulation

Variable	DF	Parameter Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
x	1	**0.**9929	0.3545	7.8441	0.0051	2.699

The estimate of the PH-slope and shape parameters are $\hat{\beta}$ = -.04331 And $\hat{\alpha}_1$ =2.8117, respectively.

Using the relationship above, the estimate of PH-slope is ($-\hat{\beta}\hat{\alpha}_1$) = -(-.04331)×(2.8117) = 1.2177

This compares to $\hat{\gamma}$ = .09929, with standard error 0.3545 .

The statements of statistical program SAS are using to estimate the parameters in the two models by using the simulation

do k = 1 to 1000 by 1;	run;
data ali ;	ods trace off;
do x = 1 to 5 by 1;	ods trace on;
do i = 1 to 3 by 1;	Proc phreg data=ali;
F=rand('uniform');	model t=x;
t=(-exp(-x)*LOG(1-F))**(1/2);	ods select Parameter Estimates;
output;	run;
end;	ods trace off;
end;	%mend;
run;	%one
ods trace on;	ods tagsets.excelxp close;
proc lifereg data=c;	run;
model t=x / dist=Weibull;	Quit
ods select Parameter estimates;	

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.

References

1. Allison, P. D. (2002), "Bias in Fixed-Effects Cox Regression with Dummy Variables," unpublished paper, Department of Sociology, University of Pennsylvania.
2. Breslow N. (1974). Covariance analysis of censored survival data. *Biometrics***,** *30,* 89-99.
3. Cox, D.R. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society, 34,* 187–220.
4. Howard, R. Chorles, A. & Lawrence, A. K., (1974), "Maximum Likelihood Estimation with the Weibull Model", JASA, Vol. 69.
5. Hu P, Tsiatis AA, Davidian M. Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics* 1998; **54**: 1407- 1419.
6. Lee, E. (1992). Statistical Methods for Survival Data Analysis. New York: John Wiley & Sons, Inc.
7. Park,P.J. *et al*. (2002) Linking expression data with patient survival times using partial least squares.
   Bioinformatics, **18**, S120–S127.
8. Schemper M. Cox analysis of survival data with non-proportional hazard functions. *Statistician* 1992; **41**: 455-465.