The Seoul Journal of Medicine
Vol. 33, No. 2: 97-103, June 1992

An Evaluation of the Weibull and the Logistic Models for Cox's Proportional Hazards Model

Moo-Song Lee[1], Youngjo Lee[3], Keun-Young Yoo[1], Dong-Young Noh[2] and Kuk-Jin Choe[2]

Department of Preventive Medicine[1] and Department of Surgery [2],
Seoul National University College of Medicine,
Seoul, 110-799, Korea and
Department of Statistics [3], College of Natural Science, Hallym University,
Chuncheon, Korea

= Abstract =**Cox's proportional hazards model has been widely used in medical researches to evaluate the relationship between prognostic factors of a disease and the occurrence of outcome event. On a theoretical basis, regression coefficient estimated from Cox's proportional hazards model could be approximated by using the Weibull and the logistic model. Breast cancer cases (n=86) diagnosed at the Seoul National University Hospital were selected to evaluate the possibility of some accelerated models as an approximate model to Cox's proportional hazards model. Age at operation, tumor size and lymph node metastasis were the variables concerned in this study. Parameter estimates of two variables from the Weibull model, which seemed not to violate the proportionality assumption of Cox's model, showed almost identical values to those from Cox's proportional hazards model. However, there was a substantial degree of discrepancy in the parameter estimate of another variable, which showed an apparent unproportionality. This study confirmed that both the Weibull and the logistic models could be used as approximate methods to the estimates from Cox's proportional hazards model. Particularly noteworthy was the fact that the PC-SAS system could be successfully applied to survival analysis when the parameters were going to be estimated using Cox's model.**

Key Words: *Cox's proportional hazards model, Logistic model, PC-SAS system, Survival analysis, Weibull model*

INTRODUCTION

A survival analysis is a statistical technique that quantifies the relationship between time to death or the recurrence of a disease and prognostic factors presumed to be associated with the disease (Miller 1981). It has been widely used to analyze survival data in medical fields, especially in clinical trials or in oncological studies. The technique has its own domain · compared to the other statistical methods because of its unique peculiarity of data, with which the technique is dealing; out-

Received December 1991, and in final form May 1992.
서울대학교 의과대학 예방의학교실 : 이무송, 유근영
서울대학교 의과대학 외과학교실 : 노동영, 최국진
한림대학교 자연과학대학 통계학과 : 이영조

come variable of survival analysis is always composed of two separate variables: 1) survivorship of an individual and 2) the time to death or recurrence.

Whenever confounding variable(s) is suspected to distort the relationship between an outcome and prognostic factors, multivariate analysis should be required to yield more valid results. Cox's proportional hazards model can be applied to the survival data in order to adjust for the confounding effect (Cox 1972). The model has been preferred to other statistical methods because of easy understanding of the results, as well as its availability in several computer softwares. Otherwise, accelerated failure time model using some parametric distributions, such as the exponential, Weibull, log-logistic, or gamma distribution, is also available for the purpose.

Several statistical softwares provide procedures for Cox's proportional hazards model. The BMDP-2L program in the BMDP statistical package (Dixon 1987) and the COX procedure in the EPILOG software (Epicenter Software 1988) are commonly used programs. The LIFEREG procedure in the PC-SAS system, Version 6.03, is available for parametric methods using accelerated failure time models (SAS Institute Inc. 1988). Although the PC-SAS system is the most widely used program in many scientific fields, the system contains no procedure for fitting survival data to Cox's proportional hazards model. This study was conducted to develop a program which may enable the LIFEREG procedure in the PC-SAS system to fit to Cox's proportional hazards model. Approximation of the results from the parametric method using the Weibull and the logistic model to Cox's proportional hazards model was evaluated.

MATERIALS AND METHODS

1. Theoretical consideration

1) Cox's proportional hazards model

A functional relationship between prognostic factors, X_i, that may be related to the hazard function, and the hazard of a disease in Cox's proportional hazards model can be formulated as follows (Cox 1972);

$$\lambda_1(t) = \exp(\beta_i \cdot X_i) \cdot \lambda_0(t)$$

where, $\lambda_0(t)$ is a baseline hazard of an individual without covariates, X_i, and $\lambda_1(t)$ is the hazard for an individual with the covariates. β_i is an unknown regression coefficient which refers to the functional relation between the covariate and the hazards. Cox's proportional hazards model is based on two assumptions; proportionality assumption and log-linear assumption. The former assumes that a hazard of an individual with covariates, $\lambda_1(t)$, is proportional to the baseline hazard of the individual, $\lambda_0(t)$. The latter assumes that the proportionality is in log-linear relation. A data to be analyzed is, therefore, recommended to be checked for the proportionality assumption whenever Cox's proportional hazards model is applied for the survival analysis (Lee *et al.* 1991). Compatibility of a data to the assumption can be assessed by graphic method using the log-negative-log survival function.

The cumulative survival functions from Cox's proportional hazards model can be expressed as follows;

$$S_1(t|X) = S_0(t)^{\exp(-\beta_i \cdot X_i)}$$

where, $S_0(t)$ is the cumulative survival function for an individual with a baseline hazard, and $S_1(t|X)$ is that for an individual with the covariates, X_i.

2) The accelerated failure time model

The accelerated failure time model, which is widely used in engineerings, assumes that the effect of independent variables on an event-time distribution is multiplicative on the event time. In the accelerated failure time model, time to failure (T) of an individual with covariates, Xi, is defined as follows (Cox and Oakes 1984);

$$T = \exp(\delta_i \cdot X_i) \cdot T_0^{\sigma}$$

where, T_0 is a failure time of an individual without covariates, δ_i is unknown regression coefficient, and σ is a scale parameter of the accelerated failure time model. The distribution can be taken from any distribution, for example, the exponential, Weibull, log-logistic, and gamma distributions (Cox and Oakes 1984).

The cumulative survival functions from the accelerated failure time model can be expressed as follows:

$$S_1(t|X) = S_0\,[\exp(-\delta_i \cdot X_i) \cdot t]$$

3) Cox's proportional hazards model and the Weibull model

Considering only a dichotomous variable in a univariate setting for simplicity, the cumulative survival functions of Cox's proportional hazards model and the accelerated failure time model can be converted by log-negative-log transformation as follows:

$$\log[-\log S_1(t)] = -\beta_i \cdot X_i + \log[-\log S_0(t)]$$

for Cox's proportional hazards model, and

$$\log[-\log S_1(t)] = \log\{-\log S_0[\exp(-\delta_i \cdot X_i) \cdot t]\}$$
$$= g_0(-\delta_i \cdot X_i + \tau)$$

for the accelerated failure time model

where, $g_0(\tau) = \log[-\log S_0(t)] = \log[-\log S_0(\exp \tau)]$, and $\tau = \log(t)$. If the distribution of the time to failure follows the Weibull distribution, the cumulative survival functions will be as follows (Cox and Oakes 1984):

$$S_0(t) = \exp(-\alpha \cdot t^{r})$$

where, r is an inverse of the scale parameter of the accelerated failure time model. The cumulative survival functions of the two models can then be modified into:

$$\log[-\log S_1(t|x)] = -\beta_i \cdot X_i + r \cdot \tau + \log \alpha$$

for Cox's proportional hazards model, and

$$\log[-\log S_1(t|x)] = g(-\delta_i \cdot X_i) + r \cdot \tau + \log \alpha$$

for the accelerated failure time model.

Finally, a functional relation between β_i and δ_i can be drawn from those expressions as follows:

$$\beta_i = r \cdot \delta_i$$

From the mathematical identity, if the baseline distribution is the Weibull, a regression coefficient of Cox's proportional hazards model, β_i, can be estimated from the regression coefficient of the Weibull model, δ_i, by multiplying an inverse of the scale parameter, r. Also the Weibull distribution is the only distribution such that Cox's proportional hazards model and the accelerated failure time model is equivalent(Cox and Oakes 1984).

4) Cox's proportional hazards model and the logistic model

The logistic model, most widely used in categorical data analysis with a dichotomous outcome variable, is defined as follows (Cornfield 1962):

$$\text{logit } P = \omega_i \cdot X_i + \varepsilon$$

where, ω_i is an unknown regression coefficient of the logistic model, and ε is an error term. The logistic model assumes that the logit of the probability of death in an interval, conditional that death has not occurred prior to that interval, is a linear function of the covariates and a constant term specific to the interval. When the conditional probabilities are small for each interval, both the discrete proportional hazards model and the logistic model tend to yield similar results (Hosmer and Lemeshow 1989).

When the time to failure is divided into relatively short time intervals in which the failure rate is constant, the logistic model can then be converted into the following expression by incorporating interval term, ηT.

$$\text{logit } P = \omega_i \cdot X_i + \eta T + \varepsilon$$

This formula can provide a basis on an alternative approach to estimate parameters of Cox's proportional hazards model using the logistic model.

2. Materials

Breast cancer cases (n = 86) were selected among those who had been confirmed histologically at Seoul National University Hospital during 1984 to 1988. Survivorship of each study subject was followed at an endpoint of March, 1991. Data on the prognostic factors (covariates) and individual characteristics were collected from clinical record of the hospital. Age at operation, tumor size and lymph node metastasis were the variables concerned in this analysis (Table 1).

Table 1. Prognostic factors analyzed in the data

Prognostic factors	Values	
Age at operation (NAGE)	1: less than 50 years	2: greater than 50 years
Tumor size* (NT)	1: T=1	2: T=2-4
Lymph node* (NN)	1: N=0	2: N=1-3

* By UICC, AJCC classification

3. Methods

All the covariates were dichotomized for the convenience of illustration. The proportionality assumption was tested by graphic demonstration of the LLS plot (log-negative-log survival function against the log month) of the survival function of each covariate. Regression coefficients from Cox's proportional hazards model were estimated by the BMDP-2L program. Regression coefficients and the scale parameter from the Weibull model were estimated by the LIFEREG procedure of the PC-SAS system (SAS Institute Inc. 1988). Regression coefficients from the Weibull model divided by the scale parameter from the model were then compared with the parameter estimated from Cox's proportional hazards model.

Regression coefficients estimated from the logistic model were also compared with the values from Cox's proportional hazards model. In the model-building procedure, a term indicating time to failure was incorporated into the model as an indicator variable. The failure time of each variable was divided into 2, 5, and 10 months-interval. The LOGISTIC procedure of the PC-SAS system (SAS Institute Inc. 1990) was used to fit the data to the logistic model.

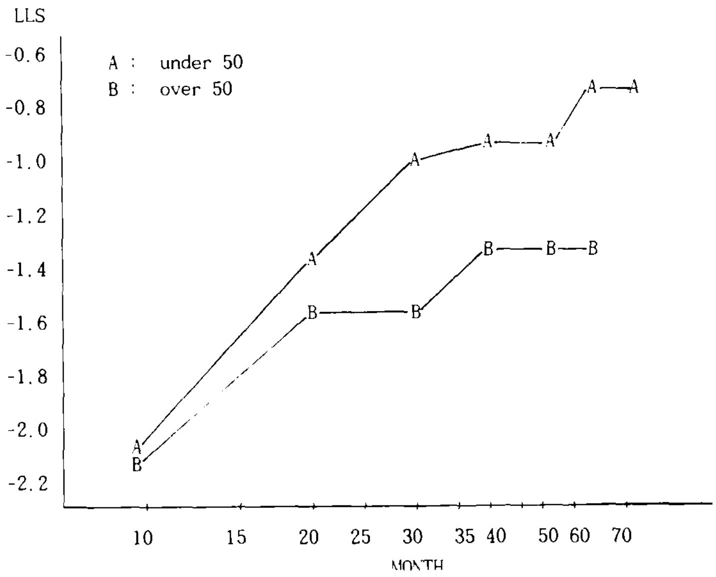

Fig. 1. Plots of the log of the negative log of the estimated survival functions against log month by NAGE.

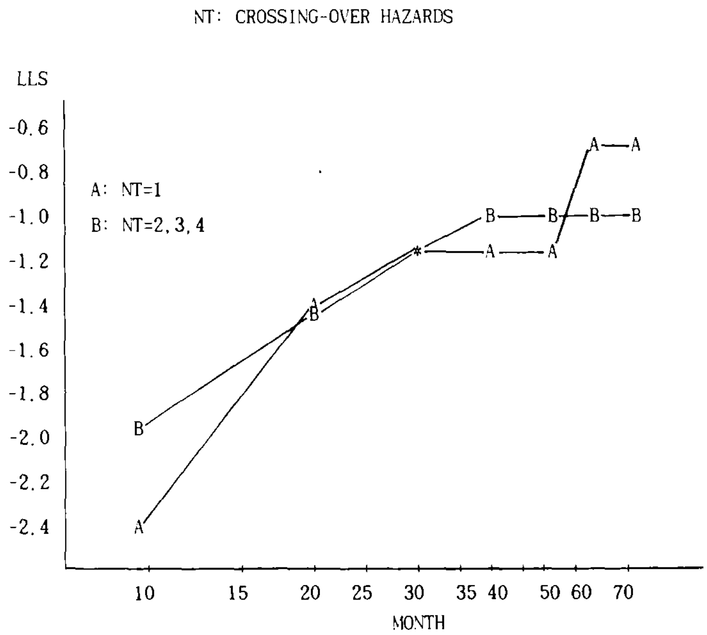

Fig. 2. Plots of the log of the negative log of the estimated survival functions against log month by NT

RESULTS

1. Test for the proportionality assumption

The compatibility of the data to the proportionality assumption of Cox's model was assessed by graphic method using the log negative log survival function (the LLS plot) for each covariate (Fig. 1-3). For age at operation (NAGE), the LLS plots showed a linear pattern, which referred to the compatibility of the vari-

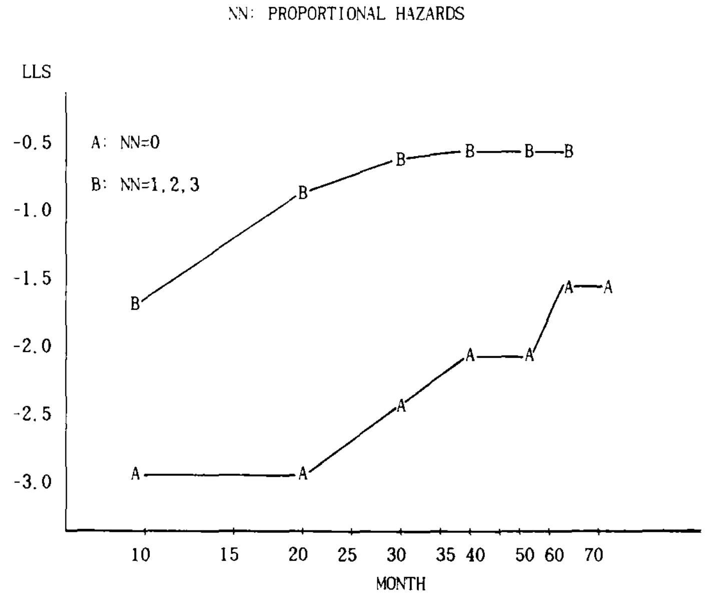

Fig. 3. Plots of the log of the negative log of the estimated survival functions against log month by NN.

able with the Weibull model. Meanwhile, there was a parallelism between the stratum-specific plots of the NAGE stratified by age subcategories in the early phase of follow-up period. The pattern turned out not to be parallel in the later period, which suggested somewhat disproportional hazards. For tumor size (NT), the LLS plots were not only crossing over each other but also nonlinear. The crossing-over plots suggested that the variable was not compatible to Cox's proportional hazards model. The nonlinear plots suggested that the variable was compatible to the Weibull model, neither. The LLS plots of the last covariate, lymph node metastasis (NN), showed a linear and parallel patterns (proportional hazards).

2. Comparison of parameters estimated from the Weibull model to those from Cox's proportional hazards model

Regression coefficients estimated from the Weibull model were compared to the parameters from Cox's proportional hazards model (Table 2). The regression coefficients of NAGE and NN, derived from regression parameters divided by scale parameters of the Weibull model, were almost identical with those estimated from Cox's proportional hazards model. However, the covariate, NT, showed a slightly different value compared to that from Cox's model. The relative risks and their 95% confidence intervals of each variable showed almost identical patterns (Table 3), as can be seen in Table 2. Statistical significance of each covariate was changed, neither. Lymph node metastasis was the only covariate, significant for predicting prognosis of breast cancer (relative risk = 4.18(1.54-11.4) for Cox's model; 4.14 (1.54-10.9) for the Weibull model).

Table 2. Regression coefficients and scale parameters by the Weibull Model and Cox's proportional hazards model

Covariates	Weibull model			Cox model
	β coeff.	scale parameter	derived β coeff.*	β coeff.
NAGE	-0.68	1.42	-0.48	-0.47
NT	-0.01	1.43	-0.01	-0.04
NN	1.93	1.36	1.42	1.43

$* \text{ derived } \beta \text{ coefficient } = \dfrac{\beta \text{ coefficient}}{\text{scale parameter}}$

Table 3. Relative risks and corresponding 95% confidence intervals by the Weibull model and Cox's proportional hazards model

Covariates	Weibull model		Cox model	
	Relative risk	Confidence interval	Relative risk	Confidence interval
NAGE	0.62	0.25-1.54	0.63	0.25-1.57
NT	0.99	0.41-2.39	0.96	0.40-2.32
NN	4.14	1.54-10.9	4.18	1.54-11.4

3. Comparison of parameters estimated from the logistic model to those from Cox's proportional hazards model

Time intervals were categorized as 2, 5, and 10 months for the fitting to the logistic model. The regression coefficients and their standard errors estimated from the logistic model were compared to the estimates from Cox's proportional hazards model (Table 4).

Table 4. Regression coefficients and standard errors by the logistic model and Cox's proportional hazards model

Covariates	Intervals used	LR Model*		Cox Model	
		β_L coeff.	s.e.**	β_C coeff.	s.e.
NAGE	2 months	-0.47	0.47	-0.47	0.47
	5 months	-0.48	0.48		
	10 months	-0.35	0.49		
NT	2 months	-0.13	0.43	-0.04	0.45
	5 months	-0.02	0.46		
	10 months	-0.03	0.47		
NN	2 months	1.42	0.50	1.43	0.51
	5 months	1.41	0.51		
	10 months	1.43	0.52		

* logistic regression model

** standard error

For the covariates, there was not so significant difference between parameters estimated from the logistic model and Cox's model.

DISCUSSION

This study confirmed that both the Weibull model and the logistic model can be used as approximate methods in order to estimate parameters from Cox's proportional hazards model. Particularly noteworthy was that the PC-SAS system could be successfully applied in survival analysis when the parameters were going to be estimated from Cox's model. Approximation to Cox's model was theoretically considered and illustratively presented in this paper using actual data on prognostic evaluation of breast cancer.

For the covariates, NAGE and NN, regression coefficients estimated from the Weibull model was almost identical with those estimated from Cox's proportional hazards model. Such a good estimation was certainly due to the fact that each hazard stratified by the variable was suitable to the proportional model, as seen in graphic illustrations. However, the covariate, NT, showed a discrepancy in the parameter estimate compared to that from Cox's model, since the variable had a substantial degree of unproportionality, which was destined to violate the proportionality assumption (Cox 1972; Shibata *et al.* 1989). These findings suggest that such an estimation of regression coefficients of Cox's proportional hazards model may not be suitable, where violation of the proportionality assumption is evident. If such violation can be confirmed apparently in a survival data, one may use another approach using a parametric model, which may be best fitted to the data. The exponential function or the gamma function is generally recommended for such an alternative choice for analysis on survival data (SAS Institute Inc. 1988). Such inappropriateness due to violation of the assumption might be alleviated through increasing the sample size of the study group (Shibata *et al.* 1989). Relative risks and their 95% confidence intervals showed almost identical results for those variable suitable to the proportionality assumption in this study. In survival analysis, statistical significance may be affected by real difference in survivorship of each prognostic factor, as well as by the fitness to Cox's proportional hazards model (Lee *et al.* 1991).

For the approximation procedure of the parameter from the logistic regression model to Cox's regression model, similar results could be obtained for the covariates, NAGE, NN and NT. Such approximation could be validated, because we handled the survival time as a discrete time interval. It has been suggested that the estimates from the discrete logistic model may be similar to those from an analysis based on the proportional hazards model when the event rates are small in each interval, normally less than 0.1 (Hosmer and Lemeshow, 1989).

REFERENCES

Cornfield J. Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. Federation Proceedings 1962;

21:58-61

Cox DR. Regression models and life tables (with discussion). J R Stat Soc B 1972; 34:187-220

Lee MS, Yoo KY, Noh DY, Choe KJ. Proportionality assumption test of Cox's proportional hazard model in survival analysis. J Korean Cancer Assoc 1991; 23: 852-9

Cox DR, Oakes D. Analysis of survival data. Chapman and Hall, London New York, 1984

Dixon WJ. BMDP Statistical Software. University of California Press, Berkeley, 1987

Epicenter Software. Epilog Plus. Epicenter Software, Pasadena, California, 1988

Miller RG. Survival analysis. John Wiley & Sons, New York, 1981

Hosmer DW, Lemeshow S. John Wiley & Sons, New York Chichester Brisbane Toronto Singapore, 1989

SAS Institute Inc. SAS Guide for Personal Computers, Version 6.03 Edition. SAS Institute Inc., Cary: NC, 1988

SAS Institute Inc. SAS Technical Report P-200, Version 6.04 Edition. SAS Institute Inc., Cary: NC, 1990

Shibata A, Hamajima N, Tamakoshi A, Suzuki S, Sasaki R, Aoki K. Dealing with the proportional hazards assumption when using the proportional hazards model with a single independent variable. Jpn J Clin Oncol 1989; 19:195-201