Jurnal Teknologi, 45(C) Dis 2006: 57–66

ESTIMATING A LOGISTIC WEIBULL MIXTURE MODELS WITH LONG-TERM SURVIVORS

MOHD. RIZAM ABU BAKAR[1], ISA DAUD[2], NOOR AKMA IBRAHIM[3] & DESI RAHMATINA[4]

Abstract. The mixture model postulates a mixed population with two types of individuals, the susceptible and long-term survivors. The susceptibles are at the risk of developing the event under consideration. However, the long-term survivors or immune individuals will never experience the event. This paper focuses on the covariates associated with individuals such as age, surgery and transplant related to the probability of being immune in a logistic Weibull model and to evaluate the effect of heart transplantation on subsequent survival.

Keywords: Mixture model, simple logistic Weibull model, split Weibull model, split logistic Weibull model

Abstrak. Model gabungan mempostulatkan sebuah populasi yang terdiri daripada dua jenis individu iaitu peka dan kebal. Individu yang peka merupakan individu yang berisiko mengalami peristiwa terhadap kajian yang dibuat. Manakala individu yang kebal adalah individu yang tidak berisiko mengalami peristiwa tersebut. Kertas kerja ini memfokuskan kepada kovariat pada individu seperti umur, pembedahan dan pemindahan yang dihubungkan dengan kebarangkalian wujudnya kebal dalam sebuah model logistik Weibull dan menghuraikan kesan pemindahan jantung terhadap masa hayat seterusnya.

Kata kunci: Model gabungan, model logistik Weibull mudah, model Weibull belahan, model logistik Weibull belahan

1.0 INTRODUCTION

Survival analysis is a class of statistical methods for studying the occurrence and timing of many different kinds of events. There are extensive studies in the context of parametric survival models for which the distribution of the survival times depends on the vector of covariates associated with each individual. See Boag [1], Berkson and Gage [2], and Maller and Zhou [3] for approaches which accommodate censoring and covariates in the ordinary exponential model for survival.

There has been a great deal of recent interest in modeling in survival analysis when immune is a possibility. This paper extends to the case of covariates involved

[1,2,3&4] Department of Mathematics and Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
Email: [1]mrizam@putra.upm.edu.my, [2]isa@putra.upm.edu.my, [3]nakma@putra.upm.edu.my, [4]desirahmatina@yahoo.com

in the mixture models as was considered in Farewell [4]. Currently, such mixture models with immunes and covariates are in use in many areas such as medicine and criminology. See Farewell [5], Schmidt and Witte [6] and Greenberg [7].

In our formulation, the covariates are incorporated into a Weibull mixture model by allowing the proportion of ultimate failures and the rate of failure to depend on the covariates and the unknown parameter vectors via logistic model. Our models and analysis are described in Section 4.

The data set used in this study is the Stanford Heart Transplant data originally taken from Crowley and Hu [8] then reproduced by Kalbfleisch and Prentice [9]. It is on survival from early heart transplant operations at Stanford. This data set consists of the time of admission to study until the time of death (in days), survival status (uncensored and censored), ages of patients (in years) accepted into the study (Age), indicator for previous surgery (Surgery) and time to transplant (Transplant) for 103 patients.

2.0 MIXTURE MODELS

In this section, we will consider mixture model in which the probability of eventual failure is an additional parameter to be estimated, and may be less than one. Mixture models are also known as split-population models and introduced to the criminological literature by Maltz and MacCleary [10], with previous treatments in the statistical literature dating back to Anscombe [11], and they have been further developed as were summarized by Maltz [12] or Schmidt and Witte [13].

Mixture models in the biometrics literature, i.e., part of the population is cured and will never experience the event, and have both a long history (e.g. Boag [1] and Berkson and Gage [2]) and widespread applications and extensions in recent years [7, 14, 15]. The intuition behind these models is that, while standard duration models require a proper distribution for the density which makes up the hazard, mixed population models allow for an existence of a subpopulation which never experiences the event of interest. This is typically accomplished through a mixture of a standard hazard density and a point mass at zero [16]. That is, mixed population models estimate an additional parameter (or parameters) for the probability of eventual failure, which can be less than one for some portion of the data. In contrast, standard event history models assume that eventually all observations will fail, a strong and often unrealistic assumption.

In standard survival analysis, data come in the form of failure times that are possibly censored, along with covariate information on each individual. It is also assumed that if complete follow-up were possible for all individual, each would eventually experience the event. Sometimes however, the failure time data come from a population where a substantial proportion of the individuals do not experience the event at the end of the observation period. In some situations, there is reason to believe that some of these survivors are actually "cured" or "long–term survivors" in the sense that even after an extended follow-up, no further events are observed on these individuals. Long-term survivors are those who are not subject to the event of interest. For example, in a medical study involving patients with a fatal disease, the patients would be expected to die of the disease sooner or later, and all death could be observed if the patients had been followed long enough. However, when considering endpoints other than death, the assumption may not be sustainable if long-term survivors are present in population.

Using the notation of Schmidt and Witte [13], we can express a mixture model as follows. Suppose that $F_R(t)$ is the usual cumulative distribution function for failure only, and ω is the probability of being subject to reconviction, which is also usually known as the eventual death rate. The probability of being immune is $(1-\omega)$, which is sometimes described as the rate of termination. This group of immune individuals will never fail. Therefore, their survival times are infinite (with probability one) and so their associated cumulative distribution function is identically zero, for all finite $t > 0$. If we now define $F_S(t) = \omega F_R(t)$, as the new cumulative distribution function of failure for the split-population, then this is an improper distribution, in the sense that, for $0 < \omega < 1$, $F_S(\infty) = \omega < 1$.

Let Y_i be an indicator variable, such that

$$Y_i = \begin{cases} 0; \text{ if } i^{\text{th}} \text{ individual is immune} \\ 1; \text{ if } i^{\text{th}} \text{ individual is subject to failure} \end{cases} \tag{1}$$

and follows the discrete probability distribution

$$\Pr[Y_i = 1] = \omega \tag{2}$$

and

$$\Pr[Y_i = 0] = (1 - \omega) \tag{3}$$

For any individual belonging to the group of eventual failure, we define the density function of eventual failure as $F_R(t)$ with corresponding survival function $S_R(t)$, while for individual belonging to the other (immune) group, the density function of failure is identically zero and the survival function is identically one, for all finite time t.

Suppose the conditional probability density function for those who will eventually fails is

$$f(t|Y=1) = f_R(t) = F_R'(t) \tag{4}$$

wherever $F_R(t)$ is differentiable. The unconditional probability density function of the failure time is given by

$$f_s(t) = f(t \mid Y = 0)\Pr[Y = 0] + f(t \mid Y = 1)\Pr[Y = 1]$$

$$= 0\,(1\text{-}\omega) + f_R(t)\,\omega = \omega f_R(t) \qquad (5)$$

Similarly, the survival function for the failure group is defined as

$$S_R(t) = \Pr\left[T > t \middle| Y = 1\right] = \int_t^{\infty} f\left(u \middle| Y = 1\right) du$$

$$= \int_t^{\infty} f_R(u)\, du = 1 - F_R(t) \qquad (6)$$

The unconditional survival time is then defined for the split-population as

$$S_S(t) = \Pr[T > t] = \int_t^{\infty} \left\{ f\left(u \middle| Y = 0\right)\Pr[Y = 0] + f\left(u \middle| Y = 1\right)\Pr[Y = 1] \right\} du$$

$$= (1 - \omega) + \omega S_R(t) \qquad (7)$$

which corresponds to the probability of being a long-term survivor plus the probability of death at some time beyond t. In this case,

$$F_S(t) = \omega F_R(t) \qquad (8)$$

is again an improper distribution function for $\omega < 1$.

3.0 THE LIKELIHOOD FUNCTION

The corresponding likelihood function for the general model can be written as

$$L(\omega,\theta) = \prod_{i=1}^{n} \left[\omega f_R(t_i)\right]^{\delta_i} \left[(1-\omega) + \omega S_R(t_i)\right]^{1-\delta_i} \qquad (9)$$

and the log-likelihood function as

$$l(\omega,\theta) = \ln L(\omega,\theta)$$

$$= \sum_{i=1}^{n} \left\{ \delta_i \left[\ln\omega + \ln f_R(t_i)\right] + (1-\delta_i)\ln\left[(1-\omega) + \omega S_R(t_i)\right] \right\} \qquad (10)$$

where δ_i is an indicator of the censoring status of observation t_i, and θ is the vector of all unknown parameters for $f_R(t)$ and $S_R(t)$. The existence of these two types of release, one type that simply does not fail and another that eventually fail according to some distribution, leads to what may be described as simple split-model. When we modify both $f_R(t)$ and $S_R(t)$ to include covariate effects, $f_R(t \mid z)$ and $S_R(t \mid z)$ respectively, then these will be referred to as split models.

4.0 MODEL WITH COVARIATES

We now consider models with covariates. This is obviously necessary if we are to make predictions for individuals, or even if we are to make potentially accurate predictions for groups which differ systematically from our original sample. Furthermore, in many applications in economics or criminology, the coefficients of the explanatory variables may be of obvious interest.

We begin by fitting a parametric model based on the Weibull distribution. The model in its most general form is a mixture model in which the probability of eventual failure follows a logistic model, while the distribution of the time until failure is Weibull, with its scale parameter depending on covariates. The estimates are based on the usual MLE method.

To be more explicit, we follow the notation of Section 2. For individual i, there is an unobservable variable Y_i which indicates whether or not individual i will eventually fail. The probability of eventual failure for individual i will be denoted ω_i so that $P(Y_i = 1) = \omega_i$. Let Z_i be a (row) vector of individual characteristics (covariates), and let α be the corresponding vector of parameters. Then we assume a logistic model for eventual failure as

$$\omega_i = \frac{\exp\left(\alpha^T z_i\right)}{\left[1+\exp\left(\alpha^T z_i\right)\right]} \tag{11}$$

The likelihood function for this model is

$$\begin{aligned} l(\omega_i,\theta) &= \ln L(\omega_i,\theta) \\ &= \sum_{i=1}^{n}\left\{\delta_i\left[\ln\omega_i + \ln f_R(t_i)\right] + (1-\delta_i)\ln\left[(1-\omega_i)+\omega_i S_R(t_i)\right]\right\} \end{aligned} \tag{12}$$

or

$$l(\omega_i,\lambda_i,\kappa) = \sum_{i=1}^{n}\left\{\begin{aligned} &\delta_i\left[\ln\omega_i + \ln\left(\lambda_i\kappa(\lambda_i t)^{\kappa-1}\exp\left(-(\lambda_i t)^{\kappa}\right)\right)\right] \\ &+(1-\delta_i)\ln\left[(1-\omega_i)+\omega_i\exp\left(-(\lambda_i t)^{\kappa}\right)\right] \end{aligned}\right\} \tag{13}$$

We can now define special cases of this general model in (13). First, we consider the ordinary model in which $\omega_i = 1$ but the scale parameter depends on individual covariates, by substituting

$$\lambda_i = \exp\,(\beta^T z_i) \tag{14}$$

into (13), and will be known as Weibull model (with covariates). This model has been considered by Kalbflesch and Prentice [9] and Lawless [17]. Schmidt and

Witte [6] have used a very similar model in their analysis. It is not a mixture model, the log-likelihood function for this model is

$$l(\beta,\kappa) = \ln L(\beta,\kappa)$$

$$= \sum_{i=1}^{n}\left\{\delta_i\left[\ln\kappa + (\kappa-1)\ln t_i + \kappa\beta^T z_i\right] - \left(t_i \exp\left(\beta^T z_i\right)\right)^{\kappa}\right\} \quad (15)$$

The result for this case can be seen in Table 1.

The second is the model in which ω_i is replaced by a single parameter ω which will be referred to as the split Weibull model (with covariates). In this model, the probability of eventual failure is a constant, though not necessarily equal to one, while the scale parameter of the distribution of time until failure varies over individuals or depend on covariates. The likelihood function for this model is

$$l(\omega,\beta,\kappa) = \sum_{i=1}^{n}\left\{\begin{array}{l}\delta_i\left[\ln\omega + \ln\kappa + \kappa\beta^T z_i + (\kappa-1)\ln t_i - \left(t_i\exp\left(\beta^T z_i\right)\right)^{\kappa}\right] \\ +(1-\delta_i)\ln\left((1-\omega) + \omega\exp\left(-\left(t_i\exp\left(\beta^T z_i\right)\right)^{\kappa}\right)\right)\end{array}\right\} \quad (16)$$

The third model is where λ_i is replaced by a single parameter λ. We will consider this model as the simple logistic Weibull model. In this model, the probability of eventual failure varies over individual, while the distribution of time until failure does not depend on covariates. The likelihood function for this model is form by substituting Equation (11) into (13) and can be written as

$$l(\alpha,\lambda,\kappa) =$$
$$\sum_{i=1}^{n}\left\{\begin{array}{l}\delta_i\left[\ln\left(\dfrac{\exp\left(\alpha^T z_i\right)}{1+\exp\left(\alpha^T z_i\right)}\right) + \ln\kappa + \kappa\ln\lambda + (\kappa-1)\ln t_i - (\lambda t_i)^{\kappa}\right] \\ +(1-\delta_i)\ln\left(\dfrac{1+\exp\left(\alpha^T z_i\right)\exp\left(-(\lambda t_i)^{\kappa}\right)}{1+\exp\left(\alpha^T z_i\right)}\right)\end{array}\right\} \quad (17)$$

Finally, we consider the logistic Weibull model, where both the probability of eventual fail and the distribution of time until failure vary over individuals. The likelihood function for this model is obtained by substituting Equations (11) and (14) into (13) and may be written as

$$l(\alpha,\beta,\kappa) = \sum_{i=1}^{n}\left\{\begin{array}{l}\delta_i\left[\ln\left(\dfrac{\exp\left(\alpha^T z_i\right)}{1+\exp\left(\alpha^T z_i\right)}\right)+\ln\kappa+\kappa\beta^T z_i+(\kappa-1)\ln t_i-\left(t_i\exp\left(\beta^T z_i\right)\right)^{\kappa}\right] \\ +(1-\delta_i)\ln\left(\dfrac{1+\exp\left(\alpha^T z_i\right)\exp\left(-\left(t_i\exp\left(\beta^T z_i\right)\right)^{\kappa}\right)}{1+\exp\left(\alpha^T z_i\right)}\right)\end{array}\right\} \tag{18}$$

In Table 1, we can see that both the age and transplant covariates are significant with ρ-values close to zero, while surgery is found marginally significant with ρ-value of 0.0595.

Table 2 gives the result for the split Weibull model and the simple logistic Weibull model. In both respects the split Weibull model dominates the Weibull model and simple logistic Weibull model. For example, the likelihood value of −470.30 for the split Weibull model is noticeably higher than the values for the Weibull and simple

Table 1 Weibull model with covariates

Variable	Weibull model	
	Coefficient	**ρ-value**
Intercept	−7.9722460	0.0000
Age	0.0923678	0.0000
Surgery	−1.2131023	0.0595
Transplant	−2.5389025	0.0000
	κ = 0.6819701	
	ln L = −495.60	

Table 2 Split Weibull and simple logistic Weibull models

Variable	Split Weibull		Simple logistic Weibull	
	Coefficient	**ρ-value**	**Coefficient**	**ρ-value**
Intercept	−6.498745	0.0000	−0.087700	0.9592
Age	0.064654	0.0009	0.100070	0.007
Surgery	−1.243274	0.0427	−1.102607	0.1666
Transplant	−2.562121	0.0000	−3.130009	0.0639
	κ = 0.734900		λ = 0.006324	
	ω = 0.954908		κ = 0.634158	
	ln L = −470.30		ln L = −486.96	

logistic Weibull models (−495.60 and −486.96) but less than the logistic Weibull model (−468.51). Parameter estimates for the logistic Weibull model, in which both the probability of eventual failure and the distribution of time until failure vary according to covariates given in Table 3. They are somewhat more complicated to discuss than the results from previous models, in part because there are simply more parameters, and some of them turn out to be statistically insignificant.

In Table 3, we can see that covariate age is statistically significant on the probability of immune with ρ-value of 0.0352 but marginally significant on the Weibull regression with ρ-value of 0.0564. Surgery just fail to be significant on the probability of immune with ρ-value of 0.9379 but marginally significant on the Weibull regression with ρ-value of 0.0676 and finally, transplant which is not significant on the probability of immune with ρ-value of 0.1555 but most significant in the Weibull regression with ρ-value of 0.0000.

Furthermore, these results are reasonably similar to the results we obtained using a logistic exponential model [18]. They are similar on the probability of immune where age is significant for both Weibull and exponential models with ρ-value of 0.0352 and 0.0359, respectively while surgery is not significant on both the Weibull and exponential models with ρ-value of 0.9379 and 0.0662, respectively. However, there is a slight disagreement for transplant which is not significant for the Weibull model with ρ-value of 0.1555 but highly significant for the exponential model with ρ-value of 0.0000.

On the distribution of lifetime, surgery is not significant on these two models with ρ-value of 0.0676 and 0.8793, respectively. Age is found to be significant with a ρ-value of 0.0184 for logistic exponential model but marginally significant with ρ-value of 0.0564 for the logistic Weibull model, and finally, transplant is significant with ρ-value 0.0000 for logistic Weibull model but marginally significant for logistic exponential with ρ-value of 0.0655. It seems that the logistic Weibull model is a better fit than the logistic exponential model ($\ln L = -473.33$) based on the likelihood ratio test.

Table 3 Logistic Weibull model

Variable	Equation for immune		Equation for duration, given eventual failure	
	Coefficient	**ρ-value**	**Coefficient**	**ρ-value**
Intercept	−0.405679	0.7868	−5.406052	0.0000
Age	0.085003	0.0352	0.040808	0.0564
Surgery	2.678153	0.9379	−1.617680	0.0676
Transplant	−1.700581	0.1555	−2.197031	0.0000
		$\kappa = 0.745974$		
		$\ln L = -468.514$		

5.0 CONCLUSION

In this section, we will summarise the results as shown in Table 4. For a split Weibull model, all covariates are significant but not so for the logistic Weibull model where surgery and transplant fail to be significant with ρ-value of 0.9379 and 0.1555, respectively. Age is found to be significant with ρ-value of 0.007 in simple logistic Weibull, but marginally significant for logistic Weibull model with ρ-value of 0.0564. However, surgery fails to be significant in simple logistic Weibull with ρ-value of 0.1666 but marginally significant in the logistic Weibull model with ρ-value of 0.0676. Finally, transplant is marginally significant for simple logistic Weibull model with ρ-value of 0.0639 but most significant with ρ-value of 0.000 in the logistic Weibull model.

The analysis suggest that for logistic Weibull model, only age is the important covariate in determining the lifetime distribution while transplant is the only covariate significantly contributes to fully recovery status from the heart failure.

Table 4 Results of the covariates for the Stanford Heart Transplant data

Variable	**ρ-value**		
	Split Weibull	**Simple logistic Weibull**	**Logistic Weibull model**
β_0 (*Intercept*)	0.0000	-	0.7868
β_1 (*Age*)	0.0009	-	0.0352
β_2 (*Surgery*)	0.0427	-	0.9379
β_3 (*Transplant*)	0.0000	-	0.1555
α_0 (*Intercept*)	-	0.9592	0.0000
α_1 (*Age*)	-	0.007	0.0564
α_2 (*Surgery*)	-	0.1666	0.0676
α_3 (*Transplant*)	-	0.0639	0.0000
λ	-	0.0000	-
κ	0.0000	0.0000	0.0000
ω (*Population split*)	0.0000	-	-

ACKNOWLEDGEMENTS

The authors would like to acknowledge the research grant funded by IRPA no. 54429 and no. 54207.

REFERENCES

[1] Boag, J. W. 1949. Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. *Journal of the Royal Statistical Society, Series B*. 11(1): 15-44.

[2] Berkson, J. and R. P. Gage. 1952. Survival Curve for Cancer Patients Following Treatment. *Journal of the American Statistical Association*. 47: 501-515.

[3] Maller, R. and S. Zhou. 1994. Testing for Sufficient Follow-Up and Outliers in Survival Data. *Journal of the American Statistical Association*. 89: 1499-1506.

[4] Farewell, V. T. 1977. A Model for a Binary Variable with Time Censored Observations. *Biometrika*. 64: 43-46.

[5] Farewell, V. T. 1982. The Use of Mixture Models for the Analysis of Survival Data with Long-term Survivors. *Biometrics*. 38: 1041-1046.

[6] Schmidt, P. and A. D. Witte. 1984. *An Economic Analysis of Crime and Justice: Theory, Methods, and Applications*. Orlando, FL.: Academic Press.

[7] Greenberg, D. F. 1991. Modeling Criminal Careers. *Criminology*. 29(1): 17-46.

[8] Crowley, J. and M. Hu. 1977. Covariance Analysis of Heart Transplant Survival Data. *Journal of the American Statistical Association*. 72: 27-36.

[9] Kalbfleisch, J. D. and R. L. Prentice. 1980. *The Statistical Analysis of Failure Data*. New York: Wiley.

[10] Maltz, M. D. and R. MacCleary. 1977. The Mathematics of Behavioral Change: Recidivism and Construct Validity. *Evaluation Quarterly*. 1: 421-438.

[11] Anscombe, F. J. 1961. Estimating a Mixed-exponential Response Law. *Journal of American Statistical Association*. 56: 493-502.

[12] Maltz, M. D. 1984. *Recidivism*. New York: Academic Press.

[13] Schmidt, P. and A. D. Witte. 1988. *Predicting Recidivism Using Survival Models*. New York: Springer-Verlag.

[14] Aalen, O. O. 1988. Heterogeneity in Survival Analysis. *Statistics in Medicine*. 7: 1121-1137.

[15] Kuk, A. Y. C. and C. H. Chen. 1992. A Mixture Model Combining Logistic Regression with Proportional Hazards Regression. *Biometrika*. 79: 531-541.

[16] Maller, R. and S. Zhou. 1996. *Survival Analysis with Long-Term Survivors*. 1st edition. New York: Wiley.

[17] Lawless, J. F. 1982. *Statistical Models and Methods for Lifetime Data*. New York: Wiley.

[18] Abu Bakar, M. R. and D. Rahmatina. 2005. Split Population Model For Survival Analysis. Proceedings of the Simposium Kebangsaan Sains Matematik XIII. Universiti Utara Malaysia. 51-54.