Logistic Loglogistic With Long Term Survivors For Split Population Model

Desi Rahmatina desirahmatina@gmail.com
Department of Accounting
Universitas Maritim Raja Ali Haji,
Tanjungpinang,Indonesia.

Abstract

The split population model postulates a mixed population with two types of individuals, the susceptibles and long-term survivors. The susceptibles are at the risk of developing the event under consideration, and the event would be observed with certainty if complete follow-up were possible. However, the long-term survivors will never experience the event. We known that populations are immune in the Stanford Heart Transplant data. This paper focus on the long term survivors probability vary from individual to individual using logistic model for loglogistic survival distribution. In addition, a maximum likelihood method to estimate parameters in the split population model using the Newton-Raphson iterative method.

Keywords: Split Population Model, Logistic Loglogistic Model, Split Loglogistic Model.

1. INTRODUCTION

Split population models are also known as mixture model. The data used in this paper is Stanford Heart Transplant data. Survival times of potential heart transplant recipients from their date of acceptance into the Stanford Heart Transplant program [3]. This set consists of the survival times, in days, uncensored and censored for the 103 patients and with 3 covariates are considered Ages of patients in years, Surgery and Transplant, failure for these individuals is death. Covariate methods have been examined quite extensively 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 [6] for approaches which accommodate censoring and covariates in the ordinary exponential model for survival.

Currently, such mixture models with immunes and covariates are in use in many areas such as medicine and criminology. See for examples [4][5][7]. In our formulation, the covariates are incorporated into a split loglogistic 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. Within this setup, we provide simple sufficient conditions for the existence, consistency, and asymptotic normality of a maximum likelihood estimator for the parameters involved. As an application of this theory, the likelihood ratio test for a difference in immune proportions is shown to have an asymptotic chi-square distribution. These results allow immediate practical applications on the covariates and also provide some insight into the assumptions on the covariates and the censoring mechanism that are likely to be needed in practice. Our models and analysis are described in section 5.

2. PREVIOUS METHODOLOGY AND DISCUSSION

[2] was the first to publish in a discussion paper of the Royal Statistical Society. He used the method of maximum likelihood to estimate the proportion of cured breast cancer patients in a population represented by a data set of 121 women from an English hospital. The follow-up time for each woman varied up to a maximum of 14 years. [2] approach was to assume a lognormal

distribution as the survival distribution of the susceptibles, although, curiously he noted that an exponential distribution is in fact a better fit to the particular set of data analyzed, and to treat deaths from causes other than the cancer under consideration as a censoring mechanism.

[8] used a model consisting of a mixture of the exponential distribution and a degenerate distribution, to allow for a cured proportion, they fitted this model to a large data set consisting of 2682 patients from the Mayo clinic who suffered from cancer of the stomach. The follow-up time on some of their patients was as much as 15 years. They noted that a correct interpretation of the existence of patients cured of the disease should be that the death rates for individual with long follow-up drop to the baseline death rate of the population.

[9] applied a Weibull mixture model with allowance for immunes to a prospective study on breast cancer. Information on various factors was collected at a certain time for approximately 5000 women, of whom 48 subsequently developed breast cancer. [9] wished to estimate that proportion and to investigate how it may be influenced by risk factors, as well as to investigate how risk factors might affect the time to development of the cancer, if this occurred.

Similarly above, we will allow the covariates associated with individuals, relate the loglogistic with long term survivors, and relate to the probability of being immune in logistic model.

3. SPLIT MODELS

In this section, we will consider ‘split population models’ (or simply ‘split models’) in which the probability of eventual death is an additional parameter to be estimated, and may be less than one. Split 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 [2] and widespread applications and extensions in recent years [4]. The intuition behind these models is that, while standard duration models require a proper distribution for the density which makes up the hazard (i.e., one which integrates to one; in other words, that all subjects in the study will eventually fail), split population models allow for 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 [6]. That is, split 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 does 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” 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 deaths 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 survivor are present in population. In contrast, the remaining individuals are at the risk of developing the event and therefore, they are called *susceptibles*.

Using the notation of [7], we can express a split model as follows. Suppose that $F_R(t)$ is the usual cumulative distribution function for death 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-ω), which is sometimes described as the rate of termination. This second group of immune individuals will never reoffend. 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 an improper distribution, in the sense that, for $0 < \omega < 1$, $F_S(\infty) = \omega < 1$.

Let Y_i be an indicative variable, such that

$$Y_i = \begin{cases} 0; & \text{ith individual will never fail} \\ 1; & \text{ith individual will eventually fail} \end{cases}$$

and follows the discrete probability distribution

$$\Pr[Y_i = 1] = \omega$$

and

$$\Pr[Y_i = 0] = (1 - \omega).$$

For any individual belonging to the group of death, 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 fail (death) is

$$f(t \mid Y = 1) = f_R(t) = F_R^{'}(t)$$

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

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

$$= 0\ (1 - \omega) + f_R(t)\ \omega = \omega\ f_R(t).$$

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

$$S_R(t) = \Pr[T > t \mid Y = 1] = \int_t^\infty f(u \mid Y = 1)du$$

$$= \int_t^\infty f_R(u)du = 1 - F_R(t).$$

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

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

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

which corresponds to the probability of being a long-term survivor plus the probability of being a recidivist who reoffends at some time beyond t.

In this case,

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

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

4. THE LIKELIHOOD FUNCTION

The likelihood function can then be written as

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

and the log-likelihood function becomes

$$l(\omega,\theta)=\ln L(\omega,\theta)=\sum_{i=1}^{n}\{\delta_i[\ln\omega+\ln f_R(t_i)]+(1-\delta_i)\ln[(1-\omega)+\omega S_R(t_i)]\}$$

where δ_i is an indicator of the censoring status of observation t_i, and θ is 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 reoffend and another that eventually fails 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*.

We fit split models to our data using the same three distributions as were considered in the section (exponential, Weibull and loglogistic). The likelihood values achieved were -511.21, -495.60 and -489.17, respectively. The loglogistic model fits the estimation better than other two distributions, while the exponential model better than weibull model. The value of the 'splitting parameter' ω implied by our models were 0.81, 0.84 and 0.78 for the exponential, Weibull and loglogistic distributions, respectively.

5. MODEL WITH EXPLANATORY VARIABLES

We now consider models with explanatory variables. 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. Futhermore, 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 loglogistic distribution. The model in its most general form is a split model in which the probability of eventual death follows a logistic model, while the distribution of the time until death is loglogistic, with its scale parameter depending on explanatory variables. The estimate are based on the usual MLE method.

To be more explicit, we follow the notation of section 3. For individual *i*, there is an unobservable variable Y_{i} which indicates whether or not individual *i* will eventually return to prison. The probability of eventual failure for individual *i* will be denoted ω_{i} so that $P(Y_{\mathrm{i}}=1)=\omega_{\mathrm{i}}$. Let Z_{i} be a (row) vector of individual characteristics (explanatory variables), and let α be the corresponding vector of parameters. Then we assume a logistic model for eventual death:

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

Next, we assume that the distribution of time until death is loglogistic , with scale parameter λ and shape parameter κ .

The likelihood function for this model is

$$l(\omega_{\mathrm{i}}, \theta) = \ln L(\omega_{\mathrm{i}}, \theta)$$
$$= \sum_{i=1}^{n} \{\delta_i [\ln \omega_{\mathrm{i}} + \ln f_R(t_i)] + (1-\delta_i) \ln[(1-\omega_{\mathrm{i}}) + \omega_{\mathrm{i}}\, S_R(t_i)]\}.$$

We can now define special cases of this general model. First, the model in which $\omega_{\mathrm{i}} = 0,$ but in which the scale parameter depends on individual characteristics, will be called *Loglogistic model* (with explanatory variables) , it is not a split model. Second, the model in which ω_{i} is replaced by a single parameter ω will be referred to as the *split Loglogistic model* (with explanatory variables). In this model the probability of eventual death is a constant, though not necessarily equal to one, while the scale parameter of the distribution of time until death varies over individuals or depend on individual characteristic Z_{i}, so that $\lambda_i = \exp(\beta^T z_i)$.

The likelihood function for this model is

$$l(\omega_{\mathrm{i}}, \beta, \kappa) =$$
$$\sum_{i=1}^{n} \left\{ \begin{array}{l} \delta_i \left[\ln \omega_{\mathrm{i}} + \beta^T z_i + \ln \kappa + (\kappa - 1) \ln(t_i \exp(\beta^T z_i)) - 2 \ln\left(1 + \left(\exp(\beta^T z_i)\kappa t_i\right)^{\kappa}\right)\right] + \\ (1-\delta_i) \ln \left[\dfrac{(1-\omega_{\mathrm{i}}) \left[1 + \left(\exp(\beta^T z_i)\kappa t\right)^{\kappa}\right]^2 + \omega_{\mathrm{i}}}{\left[1 + \left(\exp(\beta^T z_i)\kappa t\right)^{\kappa}\right]^2} \right] \end{array} \right\}.$$

Third, the model in which λ_{i} is replaced by a single parameter λ will be called the *logistic Loglogistic model*. In this model the probability of eventual death varies over individual , while the distribution of time until death (for the eventual death) does not depend on individual characteristics. The likelihood function for this model is

$$l(\alpha, \lambda, \kappa) =$$
$$= \sum_{i=1}^{n} \left\{ \begin{array}{l} \delta_i \left[\ln\left(\dfrac{\exp(\alpha^T z_i)}{1 + \exp(\alpha^T z_i)} \right) + \ln \lambda + \ln \kappa + (\kappa - 1) \ln(\lambda t_i) - 2 \ln\left(1 + (\lambda \kappa t)^{\kappa}\right) \right] + \\ (1-\delta_i) \ln \left(\dfrac{[1 + (\lambda \kappa t)^{\kappa}]^2 + \exp(\alpha^T z_i)}{\left[1 + \exp(\alpha^T z_i)\right] \left[1 + (\lambda \kappa t)^{\kappa}\right]^2} \right) \end{array} \right\}.$$

Finally, the general model as presented above will be called the *logistic / individual Loglogistic model*. In this model both the probability of eventual death and the distribution of time until death vary over individuals, the likelihood function for this model is

$$l(\alpha,\beta,\kappa) =$$

$$= \sum_{i=1}^{n} \left\{ \begin{array}{l} \delta_i \left[\begin{array}{l} \ln\left(\dfrac{\exp(\alpha^T z_i)}{1+\exp(\alpha^T z_i)} \right) + \beta^T z_i + \ln\kappa + (\kappa-1)\ln\left(t_i \exp(\beta^T z_i)\right) - \\ 2\ln\left(1+\left(\kappa t_i \exp(\beta^T z_i)\right)^{\kappa}\right) \end{array} \right] + \\ (1-\delta_i)\ln\left(\dfrac{[1+\left(\kappa t_i \exp(\beta^T z_i)\right)^{\kappa}]^2 + \exp(\alpha^T z_i)}{\left[1+\exp(\alpha^T z_i)\right]\left[1+\left(\kappa t_i \exp(\beta^T z_i)\right)^{\kappa}\right]^2} \right) \end{array} \right\}.$$

In the tables 1 gives the results for the split loglogistic model and the logistic loglogistic model. The split loglogistic model dominates the logistic loglogistic models. For example, the likelihood value of -18.2007 for the split loglogistic model is noticably higher than the values for the logistic loglogistic models with likelihood value of -19.7716. We now turn to the logistic /individual loglogistic model, in which both the probability of eventual death and the distribution of time until death vary according to individual characteristics. These parameter estimates are given in table 2. They are somewhat more complicated to discuss than the results from our other models, in part because there are simply more parameters, and some of them turn out to be statistically insignificant.

In table 2, we can see that two covariates have significant on the probability of immune, Age and Transplant with (p - value 0.0081 and 0.031, respectively) but different on the loglogistic regression, Age is fail significant with p -value of 0.1932, while Transplant to be significant with p - value of 0.0002. Surgery just fail to be significant on the probability of immune with p - value of 0.9249 but significant on the loglogistic regression with p -value of 0.077.

Furthermore, these results are reasonably similar to the results we obtained using a logistic/individual exponential model [1]. There are similars on the probability of immune that Age and Transplant are significant with (p - value 0.0081 and 0.031, respectively) for logistic /individual loglogistic model and with (p - value 0.0359 and 0.000, respectively) for logistic /individual exponential model, while Surgery did not have significant on both the loglogistic and exponential model with (p -value 0.9249 and 0.0662, respectively). Next, we analyzing statistically significant on the distribution of time until death using both the logistic/individual loglogistic and exponential model. Age did not have significant with p -value of 0.1932 for loglogistic model but significant with p - value of 0.0184 for the exponential model, Surgery is significant with p -value of 0.0077 for loglogistic model but just fail significant with p -value of 0.8793 for the exponential model and finally Transplant is significant with p -value 0.0002 for loglogistic model but marginally significant for exponential model with p -value 0.0655.

Variable	Split loglogistic		Logistic loglogistic	
	Coefficient	p - value	Coefficient	p - value
intercept	9.683928	0.0000	-0.207918	0.8972
Age	-2.240139	0.0000	0.097019	0.0039
Surgery	-8.762389	0.2153	-0.983625	0.187
Transplant	-6.673942	0.1429	-3.074790	0.0468
	κ = 0.094981		λ = *0.021881*	
	ω **=** 0.808882 *ln L = -18.2007*		κ = *0.566241* *ln L= -19.7716*	

TABLE 1: Split Loglogistic Model and Logistic Loglogistic Model.

Variable	**Equation for Pr(never fail)**		**Equation for duration, given eventual failure (Loglogistic regression)**	
	Coefficient	p - value	Coefficient	p - value
intercept	-0.529806	0.6852	-4.387064	0.0012
Age	0.087290	0.0081	0.037826	0.1932
Surgery	0.130583	0.9249	-2.250424	0.0077
Transplant	-2.254443	0.031	-2.064279	0.0002
κ = 0.767896 ln L = -468.533				

TABLE 2: Logistic / Individual Loglogistic Model.

6. CONCLUSION

In this section, we will summaries the result above about significantly covariates in the data for those models which we presented in section 4, and we shown in table 3. As we can see in table 3, There are similars on both the split loglogistic and logistic/ individual loglogistic model that Age is significant with (p -value 0.9379 and 0.1555, respectively) and Surgery just fail to be significant with (p -value 0.2153 and 0.9249, respectively), while the different that Transplant did not have significant with p -value of 0.1429 for split loglogistic but to be significant with p -value of 0.031 for logistic/ individual loglogistic model.

* not relevant

Variable	p -value		
	Split Loglogistic	*Logistic* Loglogistic	Logistic/ individual Loglogistic model.
β_0 *(intercept)*	0.0000	*	0.6852
β_1 *(Age)*	0.0000	*	0.0081
β_2 *(Surgery)*	0.2153	*	0.9249
β_3 *(Transplant)*	0.1429	*	0.031
α_0 *(intercept)*	*	0.8972	0.0012
α_1 *(Age)*	*	0.0039	0.1932
α_2 *(Surgery)*	*	0.1870	0.0077
α_3 *(Transplant)*	*	0.0468	0.0002
λ	*	0.0044	-
κ	0.0000	0.0000	0.0000
ω *(Population split)*	0.0000	-	-

TABLE 3: Significantly Covariates for the Stanford Heart Transplant Data.

Now, we can see that Age and Surgery have different significant effects on both the logistic loglogistic and logistic/ individual loglogistic model. Age is found to be the significant with a p-value of 0.0039 for logistic loglogistic but not on the logistic/ individual loglogistic model where p -value of 0.1932. Surgery just fail significant for logistic loglogistic with p-value of 0.1870 but significant on the logistic/ individual loglogistic model with p-value of 0.0077, and finally Transplant have similar significant on both the logistic loglogistic and logistic/ individual loglogistic model with (p -value 0.0468 and 0.0002, respectively).

6.REFERENCES

[1] M.R Bakar and D.R Tina. “Split Population Model For Survival Analysis”. Seminar Kebangsaan Sains Kumulatif XIII, Universiti Utara Malaysia, pp. 51-54. 2005.

[2] J. W Boag, “Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy”. Journal of the Royal Statistical Society, series B . vol 11(1), pp. 15-44. 1949.

[3] J Crowley. and M Hu.” Covariance Analysis of Heart Transplant Survival Data”. Journal of the American Statistical Association, vol. 72, pp. 27-36. 1977.

[4] Farewell.” The use of mixture models for the analysis of survival data with long-term survivors”. Biometrics*,* vol. *38,* pp. 1041-1046. 1982.

[5] D. F Greenberg. “Modeling criminal careers”. Criminology*, vol. 29 ,* pp*. 17-46.* 1991.

[6] R. Maller and X. Zhou. Survival Analysis with Long-Term Survivors, 1st Ed., New York: Wiley, 1996, pp. 112-25

[7] P Schmidt and A. D Witte. “Predicting Recidivism Using Survival Models.” Research in Criminolog*,* Springer- Verlag, New York. 1988.

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

[9] Farewell, V. T. (1977a). “A model for a binary variable with time censored observations.” Biometrika, 64, 43-46.