International Journal of Engineering Inventions
ISSN: 2278-7461, ISBN: 2319-6491, www.ijeijournal.com
Volume 1, Issue 10 (November2012) PP: 01-04

Survival Analysis Using Split Plot in Time Models

Omar Hikmat Abdulla[1], Khawla Mustafa sadik[2]
[1,2]*Department of Mathematical, Education college, Mosul University, Iraq*

Abstract:—Several experimental situations give rise to analyzing time to response on observational units (survival data) using split plot in time models. The general structure of such experiments is that the observation of the time of occurrence of an event (called a death, failure, or response) is of interest, the observational units are grouped into whole units and the treatments are randomized to whole units if time to the occurrence of an event T is continuous random variable then whole units would be considered as subsamples. If time response was grouped into intervals in the above setting, then the sufficient statistics in this case would be the counts of observed occurrences of an event (number of deaths failure) within intervals. The experiment can then be viewed as a split plot over time where time intervals periods are subunits and whole units would be the same as in continuous time setting, and the response variable is some function of the counts.

Keywords:—survival analysis, experimental situations, random variable, response variable .

I. INTRODUCTION

In life testing and medical follow up, observation of the time of occurrence of the event (called death, failure, or response) is of interest. Sometimes these occurrences may be prevented for some of the items of the sample by the occurrence of some other event (called loss or censoring). Kaplan and Meier (1958) assumed that the life time is independent of the potential loss time, and they provided, for random samples of size N, the product-limit (PL) estimate that can be defined as follows. List and label the N observed lifetimes (whether to death or loss) in order so that one has $0 \leq t_1' \leq t_2' \leq \ldots \leq t_N'$. Then $\hat{P}(t) = \prod_r [(N-r)/(N-r+1)]$, where r assumes those values for which $t_r' \leq t$, and for which t_r' measures the time to death. This is the distribution-free estimator which maximizes the likelihood function.

Cox (1972) considered the analysis of censored failure times. He suggested a regression model for the failure time T of an individual when values of one or more explanatory variables were available. For T continuous, the hazard function is given by

$$\lambda(t, Z) = \lambda_0(t, Z) \exp\left(\beta' Z\right),$$

Which is known as proportional hazard function. It is also known as the multiplicative from of the hazard function with β being the vector of the unknown parameters, and $\lambda_0(t)$ is the underlying hazard function when Z=0. For T discrete, the logistic model was suggested. A conditional likelihood and maximum likelihood estimates were obtained. However, Cox (1972) proportional hazard regression model does not handle grouped survival data or large data sets with many ties (many individuals failed at the same time).

Kalbfleisch and Prentice (1973) obtained a marginal likelihood for the regression parameters by restricting the class of models presented by Cox (1972) to those that possessed a strictly monotone survivor function or, equivalently, to those for which the hazard function $\lambda_0(t)$ was not identically zero over an open interval. The invariance of this restricted class under the group of monotone increasing transformations on T was exploited to derive a marginal likelihood function for β. If no ties occur their results and the results of Cox (1972) are the same with a simple justification. But if ties occur in the data the results obtained by Kalbfleisch and prentice (1973) are different from those suggested by Cox (1972).

We have generalized the Cox (1972) model to include main unit variability to be able to get the split plot in time model.

II. SPLIT PLOT AND VARIANCE COMPONENT LITERATURE

Our model for survival analysis is based on using a split plot in time model, and therefore we need to consider the related literature.

What we need in the variance component analysis is a method for split plot models with unequal sub-plot variances. We must mention here that we could not find any work in the literature that has been done for this particular study. However, a list and a presentation of the literature that has been done in both split plot model and variance component areas separately and combined will be considered. Some of the listed literature might not be of direct relation to our study, and some are related in the sense that they gave us an idea on the approach that we have used for variance component estimation.

A common assumption in split plot experiments is that the error variance for subplot treatments are the same. Curnow (1957) provided tests of significance for the departure from equality of the variances for different subplot treatments. Also, an estimate of tie ratio of a pair of such variances was provided in this paper. For the balanced two-way layout split plot design, Li and Klotz (1978) compared maximum likelihood estimators and restricted maximum likelihood estimators with minimum variance unbiased estimators of variance components. Performance was compared in terms of man squared error for the three estimators. For a general mixed-effects model Brown (1978) viewed the problem of estimating variance components in the context of linear model theory.

The approach was to estimate the unknown vector of a parameters β by some vector b and thus obtain a vector of residuals e=Y-Xb. A vector of the squares and cross products of the residuals was then obtained, the expectation of which was a known linear transformation of the variance components.

III. GROUPED TIME, MULTIPLICATIVE AND ADDITIVE HAZARD CONDITIONAL ON MAIN UNIT WITH NORMAL MAIN UNIT ERROR

As presented later, the structure for design that will be considered is that we have J main units per treatment combination according to a CRD, n_{ij} observation units in each main unit and time to response on each observational units is measured. Time to response is grouped into intervals where the points defining the time intervals are denoted by $0 = t_0 < t_1 < t_2 < \cdots < t_K$. The number of failures or deaths in time interval k, k=1,….,k is the number of failures or deaths in time interval $(t_{k-1},t_k]$.

Define

n_{ij} : number of individuals assigned to main unit j of trt i,

r_{ijk} : number of individuals failed on trt i, main unit j during time interval k,

S_{ijk} : number of individuals survived interval k for trt i and main unit j, and

n_{ijk} : number of individuals at risk for trt i, main unit j and time interval k.

For the no censoring case we have

$$n_{ij} = n_{ij\,1}\,,\ and\ \ n_{ijk} = S_{ij\,(k-1)}\ \ for\ k > 1.$$

For the censoring case we have to define C_{ijk} as the number censoring during the kth interval, then

$$n_{ij\,1} = n_{ij} - c_{ij\,1}\,,\ and$$
$$n_{ijk} = S_{ij\,(k-1)} - c_{ijk} = n_{ij\,(k-1)} - r_{ij\,(k-1)} - c_{ijk}\ \ for\ k > 1.$$

Also, define

P_{ijk} : conditional conditional probability that an individual on trt i and main unit j fails in interval k given that it survived $k-1$ time intervals, and

$q_{ijk} = 1 - P_{ijk}$: conditional probability of surviving interval k given survival of k-1 time intervals for an individual on trt i and main unit j.

Now, let $F_{ij}(t)$ be the cumulative distribution function for the continuous response time random variable T for a given main unit j. Define $S_{ij}(t) = 1 - F_{ij}(t)$ to be the survival function for trt i and main unit j.

By definition, P_{ijk} : conditional can be written as

$$P_{\text{ijk}} = \frac{\Pr\left(\begin{matrix}\text{failing in time interval } \mathbf{k} \text{ for an individual} \\ \text{on tretment } \mathbf{i} \text{ and main unit } \mathbf{j}\end{matrix}\right)}{\Pr\left(\begin{matrix}\text{surviving (k}-1\text{)time in intervals for} \\ \text{an individual on treatment i and main unit j}\end{matrix}\right)}$$

$$= \frac{F_{ij\,(t_k)} - F_{ij\,(t_{k-1})}}{1 - \mathrm{F}_{\mathrm{ij}\,(\mathrm{t}_{\mathrm{k}-1})}}$$

$$= \frac{\left(1 - F_{ij\,(t_{k-1})}\right) - \left(1 - F_{ij\,(t_k)}\right)}{\left(1 - F_{ij\,(t_{k-1})}\right)}$$

$$= 1 - \frac{\left(1 - F_{ij\,(t_k)}\right)}{\left(1 - F_{ij\,(t_{k-1})}\right)}$$

$$q_{ijk} = 1 - p_{ijk} = \frac{\left(1 - F_{ij\,(t_k)}\right)}{\left(1 - F_{ij\,(t_{k-1})}\right)} = \frac{S_{ij\,(t_k)}}{S_{ij\,(t_k - 1)}} \ldots\ldots\ldots (3.1)$$

Define the hazard function λ(t) as the limiting conditional probability of failing in an interval given surviving until that interval as the interval shrinks, to be

$$\lambda(\mathrm{t}) = \lim_{\Delta \mathrm{t} \to 0} \frac{\Pr(\mathrm{t} \le \mathrm{T} < t + \Delta t / \mathrm{T} \ge \mathrm{t})}{\Delta \mathrm{t}} = \frac{f(t)}{\mathrm{S(t)}}\ ,$$

Where f(t) and S(t) are the density function and the survival function, respectively, for the continuous response time random variable T. Cox (1972) suggested a regression model for the failure time T of an individual when values of one or more explanatory variables are available. For T continuous the hazard function is of form

$$\lambda(\mathrm{t}, \mathrm{z}) = \lambda_0(\mathrm{t})\ \exp(\acute{\beta}\,\mathrm{z})\ \ ,$$

Which is Known as the proportional hazard function, also known as a multiplicative from of the hazard function, where $\lambda_0(t)$ is the underlying hazard function when Z≡0. β is the vector of unknown parameters.

For our problem we generalize Cox's (1972) model to include the extra variability involved. In other words we will try to model the continuous time variable in a way related to Cox's (1972) model to include the random component ε_{ij}.

The multiplicative hazard function for trt i and main j that will be considered is as follows

$$\lambda_{ij}(\mathrm{t}) = \lambda_0(\mathrm{t})\ exp\left(\acute{\beta}\mathrm{X_i}(\mathrm{t}) + \varepsilon_{\mathrm{ij}}\right),$$

Where $\lambda_0(t)$ is the underlying hazard when $X_i(t)=0$ and $\varepsilon_{ij}=0$, β is the vector of unknown parameters and $X_i(t)$ are the variables influencing failure times. Also, the survival function for trt i given main unit j is given by

$$S_{ij}(t) = 1 - F_{ij}(t) = \exp\left(- \int_0^t \lambda_{ij}(u)\, du \right)$$

and therefore

$$S_{ij}(t) = \exp\left(- \int_0^t \lambda_0(u) \exp(\beta' x_i(u) + \varepsilon_{ij})\, du \right) \quad \text{…….. (3.2)}$$

Substituting (3.2) in (3.1) we get

$$q_{ijk} = \exp\left(- \int_{t_{k-1}}^{t_k} \lambda_0(u) \exp(\beta' x_i(u) + \varepsilon_{ij})\, du \right).$$

Now, let us assume that $X_i(t)$ is constant on interval k, i.e. , let X_{ik} = value of $X_i(t)$ on interval k. Then we have

$$q_{ijk} = \exp\left(- \exp(\beta' x_{ik} + \varepsilon_{ij}) \int_{t_{k-1}}^{t_k} \lambda_0(u) du \right)$$

and this lead to

$$\text{Log}(-\text{Log} q_{ijk}) = \beta' x_{ik} + \varepsilon_{ij} + \text{Log} \int_{t_{k-1}}^{t_k} \lambda_0(u)\, du. \quad \text{…….. (3.3)}$$

Let

$T_k = Log \int_{t_{k-1}}^{t_k} \lambda_0(u)\, du$, *then*

$Log(-Log\, q_{ijk}) = \beta' X_{ik} + \varepsilon_{ij} + T_k$, where $\beta \in \mathbb{R}^P$, $T_k \in \mathbb{R}$, and

$Log(-Log\, \hat{q}_{ijk}) = Log(-Log\, q_{ijk}) + \delta_{ijk}$, where $\hat{q}_{ijk} = \frac{S_{ijk}}{n_{ijk}}$, and δ_{ijk} is a random error defined by $\delta_{ijk} = Log(-Log\, \hat{q}_{ijk}) - Log(-Log\, q_{ijk})$.

For the additive from of the hazard function we generalize the model presented by Elandt-Johnson (1980) to include the random component ε_{ij} in an additive fashion. Now we drive the model that will be used later in analysis using the additive hazard model which is given by

$$\lambda_{ij}(t) = \lambda_0(t) + \beta' x_i(t) + \varepsilon_{ij}$$

The survivor function is then given by

$$S_{ij}(t) = 1 - F(t)$$

$$S_{ij}(t) = \exp\left(- \int_0^t \lambda_0(u) + \beta' x_i(u) + \varepsilon_{ij} \right) du \quad \text{…….. (3.4)}$$

$$= \exp\left(- \int_0^t \lambda_{ij}(u)\, du \right)$$

Substituting (3.4) in (3.1) we get

$$q_{ijk} = \exp\left(- \int_{t_{k-1}}^{t_k} (\lambda_0(u) + \beta' x_i(u) + \varepsilon_{ij}) \right) du$$

Again assume that $X_i(t)$ is constant on interval k. In this case we have

$$q_{ijk} = \exp\left(- \int_{t_{k-1}}^{t_k} (\lambda_0(u) + \beta' x_i(u) + \varepsilon_{ij}) \right) du,$$

and this leads to

$$log\ (q_{ijk}) = \acute{\beta} X_{ik}\ (t_{k-1} - t_k) + \varepsilon_{ij}\,(t_{k-1} - t_k) + (-\int_{t_{k-1}}^{t_k} \lambda_0(u)\ du\ ) \dots\dots\dots (3.5)$$

Define $Z_{ik} = X_{ik}\ (t_{k-1} - t_k)\,, \acute{\varepsilon}_{ij} = \varepsilon_{ij}\,(t_{k-1} - t_k),\ \ and\ T_k = -\int_{t_{k-1}}^{t_k} \lambda_0(u)\ du.$

Then

$Log(q_{ijk}) = \acute{\beta} z_{ik} + \acute{\varepsilon_{ij}} + T_k$, where $\beta \in \mathbb{R}^p$, $T_k \in \mathbb{R}$

$Log\ (\hat{q}_{ijk}) = Log\ (q_{ijk}) + \delta_{ijk}$, where $\hat{q}_{ijk} = \frac{S_{ijk}}{n_{ijk}}$, and δ_{ijk} is a random error defined by $\delta_{ijk} = Log\ (\hat{q}_{ijk}) - Log(q_{ijk})$.

Our grouped time model given by the :

$$Response = \mu + \alpha_i + \varepsilon_{ij} + \beta_k + (\alpha\beta)_{ik} + \delta_{ijk}$$

where i=1, 2, … , I , j=1, 2, …, J , and k=1, 2, …, K . is similar to these continuous models in the sense of having similar set of parameters. Therefore we can start with continuous setting for response time T and still and up with grouped time model that we considered for analysis although in our case response time T is discrete random variable.

It is appropriate here to mention that the proportional hazards model is convenient, e.g., the log(-log) model is to be preferred over the log model for the following two reasons:

1) Using the proportional hazards model leads to work with log(-log) model specified by the equation.

$$log(-log\ q_{ijk}) = \acute{\beta} X_{ik} + \varepsilon_{ij} + log \int_{t_{k-1}}^{t_k} \lambda_0(u)\ du\ .$$

However, using the additive from for the hazard leads to work with log model specified by the equation

$$log\ (q_{ijk}) = \acute{\beta} X_{ik}\ (t_{k-1} - t_k) + \varepsilon_{ij}\,(t_{k-1} - t_k) + (-\int_{t_{k-1}}^{t_k} \lambda_0(u)\ du\ ).$$

Therefore, inference with log(-log) transform is directly related to the parameters of the continuous time interpretation. The log(-log) model is to be preferred since β is invariant to time grouping.

2) The log model has a restricted range$\hat{q}_{ijk}$'s are observed proportions and thus$0 \le \hat{q}_{ijk} \le 1$, which implies that $log(\hat{q}_{ijk}) < 0$.

REFERENCES

1. Aranda-ordaz, F. J. "An extension of proportional - hazards model for grouped data" Biometrics, 1983, 39, 109-117 .
2. Arnold, S. F. "The Theory of Linear Models and Multivariate Analysis" ,1981, John Wiley & sons, Inc., New York.
3. Cox, D. R. "Regression models and - life tables (with discussion)", Journal of the Royal Statistical Society Series B 34, 1972, 187 – 220 .
4. Elandt-Johnson, Regina C. "Some Prior Distribution In Survival Analysis A critical Insight On Relationships Derived From Cross-Sectional data"J. R. statistic. soc., B, 1980, 42, 96 - 106 .
5. Henderson, C. R. "Estimation of variance and covariance components" Biometrics, 1953, 9, 226-252 .
6. Kaplan, E. L. and Meier, P. "Non-parametric estimation from incomplete observations", Journal of the American Statistical Association 53, 1958, 457– 481 and 562–563.
7. Kay, R. "proportional hazard regression models and the analysis of censored survival data" Appl. statist., 1977, 26, 227-237 .
8. Krane, S. A. "Analysis of survival data by regression techniques" Technometics, 1963, 5, 161-174 .
9. Manton, K. G. Wood bury,M. A. and Stallard, E. "A variance components approach to categorical data models with heterogeneous cell population with analysis of spatial gradients in lung-cancer mortality rates in north Carolina counties", Bimetrics, 1981, 37, 359-269.