GROUPED FAILURE TIMES
TIED FAILURE TIMES

TWO CONTRIBUTIONS TO THE ENCYCLOPEDIA OF BIOSTATISTICS

Mei-Jie Zhang, Ph.D.

Technical Report 24

February 1997

Division of Biostatistics
Medical College of Wisconsin
8701 Watertown Plank Road
Milwaukee WI 53226
Phone: (414)456-8280

GROUPED FAILURE TIMES

In many investigations for life times, data are grouped prior to their statistical analysis. The grouped survival data consists of occurrence and exposure data over given time intervals and possible covariate strata. For grouped failure times there is an assumed continuous underlying hazard function in contrast to discrete failure time data (Fahrmeir [13]) with an intrinsically discrete time variable, discrete hazards, survival functions etc.

One of the primary reasons for grouping can be found in studies involving large sample sizes such as epidemiologic studies (Breslow [6]). Such studies typically involve the follow-up of large population groups over certain time periods to assess the cause and rate of death and/or to compare death rates among different population groups. Grouping data from such large sample sizes into tabular presentations (life tables) often provides a convenient format for presenting and summarizing life information. Also grouping could be done intentionally, e.g. to economize on data transmission and storage, to reduce computation, to protect the privacy of individual records, or to account for the limitations of a measurement instrument. Moreover, some large data sets are publicly released only in grouped form, as discussed by Haitovsky ([19], [20]). Some examples that illustrate such grouped survival data are: the American Cancer Society study of 1,000,000 men and women (Hammond [18]) to determine the dose-time-response relationships between smoking and lung cancer or heart disease and the life span study of over 100,000 Japanese atom bomb survivors in Hiroshima and Nagasaki (Beebe [4]).

Another important reason for grouping data is that it is often difficult or even impossible to obtain exact life time, because ethical, physical or economic restrictions in research design allow the subjects in the follow-up study to be monitored only periodically. Thus, this type of study only provides the grouped information, i.e., the exact failure time is unknown and the only available information is whether the event of interest occurred between two inspection times. The following study illustrate situations where periodic inspection is used: The National Labor Survey of Youth (NLSY) study of time to weaning of breast-fed newborns in which 927 first-born children of mothers who chose to breast feed their children were interviewed yearly.

Similar to continuous data in survival analysis, grouped survival data can involve censored data (right censoring, left censoring or double censoring) and/or truncated data. Moreover, the exact censoring or truncation times may be unknown for grouped data. For example, in the study of time to weaning of breast fed newborns, some infants are lost follow-up and some infants were withdrawn from the study without being weaned. Also grouped survival data can involve covariates (explanatory variables). Some parametric hazard models and the well-known Cox's proportional hazards model are often fitted to grouped survival data (Prentice and Gloeckler [37]).

The vast literature on grouped survival data involves: deriving the estimators of the hazard function and survival function under nonparametric or parametric models, test statistics for comparing the survival probabilities among different population groups, and large sample properties for these estimators and test statistics. Most estimates are derived based on maximum likelihood methods. Some references to such studies will be given later. The Bayesian approach to analyzing grouped survival data has also been studied in the literature (see Cornfield and Detre [8]; Johnson and Christensen [27]).

Notation of Grouped Survival Data

Let time be partitioned into a fixed sequence of intervals $\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_m$ with $\mathcal{T}_j = (t_{j-1}, t_j]$ and $0 = t_0 < t_1 < \cdots < t_m \leq \infty$. For grouped failure time data the only available information is:

n_j = number of subjects entering $\mathcal{T}_j$ not having experienced the event,

d_j = number of individuals experiencing the event in $\mathcal{T}_j$,

w_j = number of individuals lost to follow-up or withdrawn during $\mathcal{T}_j$,

μ_j = number of individuals left truncated during $\mathcal{T}_j$,

Y_j = total time of individuals at risk during $\mathcal{T}_j$.

Note that all $\mu_j = 0$ when no left truncation occurs. Also, when the subjects are monitored periodically, the total time at risk Y_j is unknown. It is often approximated by $Y_j \approx [n_j - (d_j + w_j)/2](t_j - t_{j-1})$ for right censored data.

Life Table

The life table is one of the oldest and most commonly used methods of presenting lifetime data. It is a table for presenting and summarizing data, and estimating the survival function, the probability density function and the hazard function along with the variance of these estimators. For more details on the life table, see Gehan [17], Breslow[7] and Hoem [22].

Interval Censored Grouped Data

For the interval (doubly) censored grouped data, Turnbull ([40], [41]) proposed an "self-consistency" procedure, developed by Efron [11], to estimate the survival function $S(t)$. The Turnbull Estimator is a nonparametric maximum likelihood estimator (NPMLE). Frydman [16] discussed derivation and asymptotic properties of the Turnbull Estimator. Sun [39] discussed some alternative approaches to maximizing the NPMLE.

Log-Rank Test

Comparison of the survival probabilities with treatment groups or covariate strata in the grouped data can be done through rank tests. In the continuous data case Fleming and Harrington [14] studied a class of weighted log-rank tests. These weighted log-rank tests can be extended to the grouped failure time data. The usual log-rank test (or evenly weighted log-rank test) is most commonly and widely used in practice. Here we discuss the grouped data version of the log-rank test. First, let's consider the two sample case. Let n_{ij} and d_{ij}, $j = 1, \cdots, m, i = 1, 2$, be the number at risk at begining of jth interval and observed failures in jth interval, respectively, in sample i. Take n_j and d_j to be the corresponding values in the combined sample. The data can be summarized as

	Sample		
Failure	1	2	Total
Yes	d_{1j}	d_{2j}	d_j
No	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
Total	n_{1j}	n_{2j}	n_j

corresponding to the jth time interval. The grouped data based two sample log-rank test can be computed as

$$Q = \left\{ \sum_{j=1}^{m} (d_{1j} - E_{1j}) \right\}^2 \Bigg/ \left\{ \sum_{j=1}^{m} V_{1j} \right\},$$

where E_{1j} and V_{1j} are the expected value and variance of d_{1j}, given by

$$E_{1j} = \frac{d_j n_{1j}}{n_j}, \quad \text{and} \quad V_{1j} = \frac{d_j n_{1j} n_{2j} (n_j - d_j)}{n_j^2 (n_j - 1)}.$$

Under the hypothesis of $S_1(t) = S_2(t)$, the two-sample log-rank test statistic Q has approximately the chi-squared distribution with 1 degree of freedom when the sample sizes are moderately large for each sample.

We can extend the two-sample log-rank test to the k-sample comparison. The k-sample log-rank test has a quadratic form with $(d_{1j} - E_{1j})$ replaced by the corresponding values from $(k - 1)$ samples and with V_{1j} replaced by the corresponding covariance matrix, where the (hl)th element is

$$\hat{\sigma}_{hl} = \frac{d_j n_{hj}}{n_j} \left(\delta_{hl} - \frac{n_{hj}}{n_j} \right) \frac{(n_j - d_j)}{(n_j - 1)},$$

and δ_{hl} is a Kronecker delta, i.e., $\delta_{hl} = 1$ if $h = l$, and 0 otherwise.

Parametric Models and Regression Analysis

In survival analysis some parametric models have been studied extensively. The common parametric distributions considered are Exponential, Gamma, Weibull, Log Normal and Gompertz distributions. These parametric models are often fitted to grouped data as well. The parameters are usually estimated by maximizing the full (unconditional) likelihood or the conditional likelihood. That is the likelihood function for the interval $(t_{j-1}, t_j]$ conditional on surviving till t_{j-1}. Many authors have given grouped data version MLE, see Elandt-Johnson and Johnson [12], Lawless [30] and Deddens and Koch [10]. Turnbull [42] studied a likelihood ratio statistic for testing goodness of fit for grouped failure data with possible doubly censoring.

It is important to assess the effects of covariates that may be associated with the lifetimes in many applications of survival analysis. The regression model for the conditional hazard function $\lambda(t|\boldsymbol{z})$ of the failure time given covariate $\boldsymbol{z}$ could be used to examine the covariate effects. Continuous covariates are often grouped into a fixed number strata and the value for the strata is approximated by the midpoint of the covariate in the stratum. For simplicity we consider a one dimensional covariate case. The methods and results discussed here can be extended to multidimensional cases. Let the cells into which the data are grouped be denoted $\mathcal{C}_{rj} = \mathcal{T}_r \times \mathcal{I}_j$, where $\mathcal{T}_1, \ldots, \mathcal{T}_{L_n}$ and $\mathcal{I}_1, \ldots, \mathcal{I}_{J_n}$ are the respective calendar periods (time intervals) and covariate strata. Grouped failure time data consist of the total number of failures (occurrence) and the total time at risk (exposure) in each cell $\mathcal{C}_{rj}$, given by d_{rj} and Y_{rj}. In the literature, most early work has been done under the piecewise exponential model, i.e., the hazard function is assumed to be piecewise constant within each grouping cell. The natural estimate of the unknown hazard rate λ_{rj} is $\hat{\lambda}_{rj} = d_{rj}/Y_{rj}$ (occurrence/exposure rate). Deddens and Koch [10] showed that the maximum likelihood is approximately equivalent to maximizing the piecewise exponential likelihood function

$$L = \prod_{r,j} \lambda_{rj}^{d_{rj}} \{\exp(-\lambda_{rj} Y_{rj})\}.$$

The occurrence/exposure rate estimator can also be obtained by solving the equations of $\partial \text{log} L / \partial \lambda_{rj} = 0$.

The counting process approach and martingale techniques are applicable in grouped failure time data analysis. We assume that the counting process N_i, where $N_i(t)$ is the number of failures of the ith individual during time period $[0, t]$, has intensity

$$\lambda_i(t) = Y_i(t) \lambda(t, Z_i(t)),$$

where $Y_i(t)$ is a predictable $\{0,1\}$-valued process indicating that the ith individual is at risk with $Y_i(t)=1$, and $Z_i(t)$ is a predictable covariance process. The occurrence and exposure in each cell $\mathcal{C}_{rj}$ can be written as

$$d_{rj} = \sum_i \int_{\mathcal{T}_r} I\{Z_i(t) \in \mathcal{I}_j\} dN_i(t) \quad \text{and} \quad Y_{rj} = \sum_i \int_{\mathcal{T}_r} I\{Z_i(t) \in \mathcal{I}_j\} Y_i(t) dt.$$

When the censoring processes are independent of the survival time, we can show that $M_i(t) = N_i(t) - \int_0^t \lambda_i(u)du$ are local martingales. Under the piecewise constant model ($\lambda(t,z) = \lambda_{rj}$, for $(t,z) \in \mathcal{C}_{rj}$),

$$\hat{\lambda}_{rj} = \frac{d_{rj}}{Y_{rj}} = \frac{M_{rj}}{Y_{rj}} + \lambda_{rj}\frac{Y_{rj}}{Y_{rj}},$$

where $M_{rj} = \sum_i \int_{\mathcal{T}_r} I\{Z_i(t) \in \mathcal{I}_j\} dM_i(t)$ is the martingale part of d_{rj}. Since each $t \in \mathcal{T}_r$, Y_{rj} is not predictable, the martingale techniques are not applicable directly. However, under the *iid* cases and some mild conditions, we can show that there exists a piecewise constant function f_{rj} bounded away from zero such that $n^{-1}Y_{rj}$ converges to f_{rj} in probability. Then we can replace M_{rj}/Y_{rj} by M_{rj}/nf_{rj} with the difference of $o_P(1)$. It follows that

$$\hat{\lambda}_{rj} = \frac{M_{rj}}{nf_{rj}} + \lambda_{rj} + o_P(1),$$

and the predictable variation process of M_{rj}/f_{rj} is

$$\left\langle \frac{M_{rj}}{f_{rj}} \right\rangle = \frac{\lambda_{rj} Y_{rj}}{f_{rj}^2}.$$

Therefore, $\hat{\lambda}_{rj}$ is an asymptotic unbiased estimator and the variance can be consistently estimated by

$$\hat{\sigma}_{rj} = \widehat{\mathrm{Var}}(\hat{\lambda}_{rj}) = \frac{d_{rj}}{(Y_{rj})^2}.$$

For the general nonparametric model where the hazard function is unspecified, Holford [23] noted that this estimator is inconsistent unless the grouping becomes finer as the sample size increases.

The useful models for many applications are the multiplicative and additive risk model. The model equations are given by

$$\lambda_{rj} = \lambda_{r0} \exp(\beta z_j) \quad \text{and} \quad \lambda_{rj} = \lambda_{r0} + \beta z_j,$$

where λ_{r0} is the baseline hazard rate of the rth time period. The parameters λ_{r0} and β are readily estimated by the MLE. Berry [5] and Frome [15] provide explicit

MLE for this approach. For the multiplicative risk model the hazard function can be written as $\lambda_{rj} = \exp(\alpha_i + \beta z_j)$ which has a log-linear form. It is often called the log linear piecewise constant model. Holford [24] derived the log likelihood for this model:

$$L = \sum_r \alpha_r d_{r\cdot} + \sum_{r,j} d_{rj} \beta z_j - \sum_{r,j} Y_{rj} \exp(\alpha_r + \beta z_j),$$

where $d_{r\cdot} = \sum_{j=1}^{J_n} d_{rj}$ is the number of failures in the rth calendar period. Taking derivatives of L with respect to α_r and β and setting them equal to zero, the MLE estimator of β is given by solving the following equation:

$$\sum_{r,j} z_j d_{rj} - \sum_r \frac{\sum_j Y_{rj} z_j \exp(\beta z_j)}{\sum_j Y_{rj} \exp(\beta z_j)} d_{r\cdot} = 0.$$

As we discuss later, this MLE estimator of β also can be obtained by maximizing the grouped data version of Cox's partial likelihood.

The more general models are: Cox's proportional hazards model (Cox [9]) where $\lambda(t, z) = \lambda_0(t) \exp(\beta z)$, and Aalen's additive risk model (Aalen [1]) where $\lambda(t, z) = \lambda_0(t) + \beta(t) z$.

Cox's proportional hazards model has so far been the most popular model in survival analysis. The parameter estimator $\hat{\beta}$ is obtained by maximizing Cox's partial likelihood function. Andersen and Gill [3] provide an excellent proof that $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{P} N(0, V)$, where V^{-1} is consistently estimated by $-n^{-1} \partial U(\hat{\beta})/\partial\beta$ and U is the partial likelihood score function $U(\beta) = \partial \log L(\beta)/\partial\beta$. The grouped data based estimator $\hat{\beta}_g$ can be obtained by maximizing the following approximation to the partial likelihood:

$$L_g(\beta) = \prod_{r,j} \left\{ \frac{e^{\beta z_j}}{\sum_k Y_{rk} e^{\beta z_k}} \right\}^{d_{rj}}$$

where the product is over the grouping cells, the sum is over the covariate strata, and z_j is the midpoint of the jth covariate stratum. This estimator has been studied by Kalbfleisch and Prentice [28], Holford [23], Prentice and Gloeckler [37], Breslow [6], Hoem [21], Selmer [38], and Huet and Kaddour [25]. It can be interpreted as the maximum likelihood estimator in a Poisson regression model, as shown by Laird and Olivier [29]. Under slightly stronger regularity conditions proposed in Andersen and Gill [3], it can be shown that $\sqrt{n}(\hat{\beta}_g - \beta_0) \xrightarrow{P} N(0, V)$, when the time intervals and covariate strata shrink at some suitable rate as the sample size increases. It is important to be able to assess estimation bias caused by grouping and to correct it if necessary. In the general grouped data analysis, A 'Sheppard correction' can be used to reduce the bias to a higher order of the interval width, see Lindley [31]. McKeague and Zhang [36] obtained a Sheppard correction for Cox's proportional hazards model, provided a consistent estimator for Sheppard correction, and derived the optimal rate of convergence for $\hat{\beta}_g$. The grouped data based estimator of the baseline hazard function, λ_0, is

$$\hat{\lambda}_0(t) = \frac{\sum_j d_{rj}}{\sum_j Y_{rj} e^{\hat{\beta}_g z_j}} \quad \text{for } t \in \mathcal{T}_r.$$

Aalen's additive risk model provides a useful and sometimes biologically more plausible alternative to the Cox proportional hazards model. For continuous data, Aalen proposed a least squares estimator for the cumulative hazard functions which has been studied by Aalen ([1], [2]), Mau ([32], [33]), and McKeague [34]. McKeague [35] and Huffer and McKeague [26] fit Aalen's additive risk model to the grouped data (when the covariates are observed for each individual and are non-time dependent), and studied asymptotic results for the grouped data version of the least squares estimator and weighted least squares estimator. The estimators can be generalized to the more general grouped data setting where the only available information is d_{rj} and Y_{rj} for each cell $\mathcal{C}_{rj}$. More work is needed.

Finally, fitting parametric and regression models to grouped failure time data is based on d_{rj} and Y_{rj}. As we discussed in the univariate case, Y_{rj} may not be observable in some applications. It is usually approximated by $Y_{rj} \approx (n_{rj} - (d_{rj} + w_{rj})/2) l_r$, where n_{rj} is the number of individuals at risk at beginning of the time period $\mathcal{T}_r$ for the jth covariate stratum, w_{rj} is the number of individuals who withdrew in cell $\mathcal{C}_{rj}$, and l_r is the width of the time interval $\mathcal{T}_r$. This approximation is based on the assumption that, on the average, the individuals failed or withdrew at middle of the each time period. However, in most applications, this assumption does not hold true. The bias introduced by this approximation could be severe. Cautions must be taken when grouping the data so that the number of grouping cells are sufficiently large (the width of time periods and covariate strata are relative small), and each grouping cell contains sufficient individuals at risk.

References

[1] Aalen, O.O. (1980). A model for non-parametric regression analysis of counting processes. In W. Klonecki, A. Kozek, and J. Rosinski, eds., *Leture Notes On Mathematical Statistics And Probability*, Vol. 2, Springer-Verlag, New York.

[2] Aalen, O.O. (1989). A linear regression model for the analysis of life time. *Statistics in Medicine*, **8**, 907–925.

[3] Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, **10**, 1100–1120. New York.

[**4**] Beebe, G.W. (1981). The atomic bomb survivors and problem of low dose radiation effects. *American Journal of Epidemiology*, **114**, 761–783.

[**5**] Berry, G. (1983). The analysis of mortality by the subject-years method. *Biometrics*, **39**, 173–184.

[**6**] Breslow, N.E. (1986). Cohort analysis in epidemiology, In A.C. Atkinson and S.E. Fienberg, eds., *A Celebration of Statistics: the ISI Centenary Volume.* Springer-Verlag, New York, pp. 109–143.

[**7**] Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, **2**, 437–453.

[**8**] Cornfield, J. and Detre, K. (1977). Bayesian life table analysis. *Journal of the Royal Statistical Society Series B*, **39**, 86–94.

[**9**] Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society Series B*, **34**, 187–220.

[**10**] Deddens, J.A. and Koch, G.G. (1988). Survival analysis, Grouped data, In N.L. Johnson and S. Kotz, eds., *Encyclopedia of Statistical Sciences*, Vol. 9, Wiley, New York, pp. 129–134.

[**11**] Efron, B. (1967). The two-sample problem with censored data. *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press, **4**, 831–853.

[**12**] Elandt-Johnson, R.C. and Johnson, N.L. (1980). *Survival Models and Data Analysis.* Wiley, New York.

[**13**] Fahrmeir, Ludwig. (1997). Discrete failure time models. In P.K. Andersen and N. Keiding, eds., *Encyclopedia of Biostatistics, Section: Survival Analysis*, Wiley, New York.

[**14**] Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes And Survival Analysis.* Wiley, New York.

[**15**] Frome, E.L. (1983). The analysis of rates using Poisson regression models. *Biometrics*, **39**, 665–674.

[**16**] Frydman, H. (1997). Turnbull esitmator. In P.K. Andersen and N. Keiding, eds., *Encyclopedia of Biostatistics, Section: Survival Analysis*, Wiley, New York.

[17] Gehan, E.A. (1969). Estimating survival functions from the life table. *Journal of Chronic Disease*, **21**, 629–644.

[18] Hammond, E.C. (1966). Smoking in relation to the death rates of one million men and women. *National cancer Institute Monograph*, **19**, 127–204.

[19] Haitovsky, Y. (1973). *Regression estimation from grouped observations.* Charles Griffin, London/Hafner Press, New York.

[20] Haitovsky, Y. (1983). Grouped data. In N.L. Johnson and S. Kotz, eds., *Encyclopedia of Statistical Sciences*, Vol. 3, Wiley, New York, pp. 527–536.

[21] Hoem, J. M. (1987). Statistical analysis of a multiplicative model and it application to the standarization of vital rates: a review. *International Statistics Review*, **55**, 119-152.

[22] Hoem, J. M.. (1997). Life table. In P.K. Andersen and N. Keiding, eds., *Encyclopedia of Biostatistics, Section: Survival Analysis*, Wiley, New York.

[23] Holford, T.R. (1976). Life tables with concomitant information. *Biometrics*, **32**, 587–597.

[24] Holford, T.R. (1980). The analysis of rates and of survivalship using log-linear models. *Biometrics*, **36**, 299–305.

[25] Huet, S. and Kaddour, A. (1994). Maximum likelihood estimation in survival analysis with grouped data on censored individuals and continuous data on failures. *Applied Statistics*, **43**, 325–333.

[26] Huffer, F.W. and McKeague, I.W. (1991). Weighted least squares estimation for Aalen's additive risk model. *Journal of the American Statistical Association*, **86**, 114–129.

[27] Johnson, W. and Christensen, R. (1986). Bayesian nonparametric survival analysis for grouped data. *Canadian Journal of Statistics*, **14**, 307–314.

[28] Kalbfleisch, J.D. and Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, **60**, 267–278.

[29] Laird, N. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, **76**, 231–240.

[30] Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* Wiley, New York.

[31] Lindley, D. V. (1950). Grouping corrections and maximum likelihood equations. *Proceedings. Cambridge Philosophical Society*, **46** 106–110.

[32] Mau, J. (1986). On a graphical method for the detection of time-dependent effects of covariates in survival data. *Applied Statistics*, **35**, 245–255.

[33] Mau, J. (1988). A comparison of counting process models for complicated life histories. *Applied Stochastic Models and Data Analysis*, **4**, 283–298.

[34] McKeague, I.W. (1988a). Asymptotic theory for weighted least squares estimators in Aalen's additive risk model. *Contemporary Mathematics*, **80**, 139–152.

[35] McKeague, I.W. (1988b). A counting process approach to the regression analysis of grouped survival data. *Stachastic Processes and Their Applications*, **28**, 221–239.

[36] McKeague, I.W. and Zhang, M.J. (1996). Fitting Cox's proportional hazards model using grouped survival data. In N.P. Jewell, A.C. Kimber, M.L.T. Lee and G.A. Whitmore, eds., *Lifetime Data: Models in Reliability and Survival Analysis*, Kluwer Academic Publishers, pp. 227–232.

[37] Prentice, R.L. and Gloeckler, L.A. (1978). Regression analysis of grouped Survival data with application to breast cancer data. *Biometrics*, **34**, 57–67.

[38] Selmer, R. (1990). A comparison of Poisson regression models fitted to multiway summary tables and Cox's survival model using data from a blood pressure screening in the city of Bergen, Norway. *Statistics in Medicine*, **9**, 1157–1165.

[39] Sun, Jianguo (1997). Interval censoring. In P.K. Andersen and N. Keiding, eds., *Encyclopedia of Biostatistics, Section: Survival Analysis*, Wiley, New York.

[40] Turnbull, B.W. (1974). Nonparametric estimation of a survivaorship function with double censored data. *Journal of the American Statistical Association*, **69**, 169-173.

[41] Turnbull, B.W. (1976). The empirical distribution function with arbiltrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society Series B*, **38**, 290–295.

[42] Turnbull, B.W. and Weiss L. (1978). A likelihood ratio statistics for testing goodness of fit with randomly censored data. *Biometrics*, **34**, 367–375.

TIED FAILURE TIMES

Tied failure times frequently occur in survival studies. Although theoretically a lifetime is a continuous variable, in practice it is often measured to a degree of fineness due to measurement limitation, the way failure times are recorded, or the expense of more accurate measurements may outweigh the value of added information. If the number of ties are substantial, discrete failure time models may need to be considered. Therefore, discrete failure time methods or grouped data techniques such as life tables should be used. However, if there are only a few ties, the regular procedures in handling continuous data may be used with some adjustment for tied observations. In the literature adjustment for ties has been proposed and studied for various statistical procedures in survival analysis. See Miller [9], Lawless [8], Kalbfleisch and Prentice [6], Peto and Peto [10], Andersen et al [1], and Klein and Moeschberger [7]. Here we will only discuss adjustment for ties for some common statistical procedures.

Consider the method of handling ties in the Kaplan-Meier or product-limit (PL) estimator of the survival function. If only one individual fails (no ties are present) at time t, then the factor for the single death in the PL estimator is $(1 - 1/Y(t))$ where $Y(t)$ counts the number of individuals at risk at time $t-$. For tied uncensored observations, suppose d failures occur at time t. Split the times of the d failures infinitesimally so that the factor for the d failures in the PL estimator is

$$\left(1 - \frac{1}{Y(t)}\right)\left(1 - \frac{1}{Y(t)-1}\right)\cdots\left(1 - \frac{1}{Y(t)-d+1}\right) = 1 - \frac{d}{Y(t)}.$$

If censored and uncensored observations are tied at time t, consider the uncensored individuals as having failed just before the censored observations.

In the k-sample test, the weighted log rank test statistic is

$$Z_h(t) = \int_0^t K(s)dN_h(s) - \int_0^t K(s)\frac{Y_h(s)}{Y_.(s)}dN_.(s),$$

for $h = 1, 2, \cdots, (k-1)$, where K is the weight function, $N_h(s)$ and $Y_h(s)$ are the number of failures during time period $[0, s]$ and number of individuals at risk prior to time s for hth sample, respectively, and $N_. = \sum_h N_h, Y_. = \sum_h Y_h$. The covariance of $(Z_h(t), Z_j(t))$ may be estimated consistently by

$$\hat{\sigma}_{hj} = \int_0^t K^2(s)\frac{Y_h(s)}{Y_.(s)}\left(\delta_{hj} - \frac{Y_j(s)}{Y_.(s)}\right)dN_.(s),$$

where δ_{hj} is a Kronecker delta, i.e., $\delta_{hl} = 1$ if $h = l$, and 0 otherwise. In the presence of tied observations, the covariance of $(Z_h(t), Z_j(t))$ needs to be adjusted to

$$\hat{\hat{\sigma}}_{hj} = \int_0^t K^2(s)\frac{Y_h(s)}{Y_.(s)}\left(\delta_{hj} - \frac{Y_j(s)}{Y_.(s)}\right)\frac{Y_.(s) - \Delta N_.(s)}{Y_.(s) - 1}dN_.(s).$$

Clearly, when there are no tied observations, $\hat{\sigma}_{hj}$ and $\hat{\hat{\sigma}}_{hj}$ coincide.

Cox's partial likelihood has been commonly used to estimate the coefficients, $\boldsymbol{\beta}$, in Cox's proportional hazards model. Let $t_1 < t_2 < \cdots < t_k$ be the k ordered event times. Let the set $\mathcal{D}_i$ consist of the d_i individuals who failed at the time t_i and $\mathcal{R}_i$ be the risk set prior to t_i. Denote $\boldsymbol{s}_i = \sum_{l \in \mathcal{D}_i} \boldsymbol{z}_l$. If there are ties among event times, the following adjusted partial likelihoods have been proposed:
Breslow [2] suggests a partial likelihood of

$$L_1(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{s}_i)}{\left[\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}'\boldsymbol{z}_l)\right]^{d_i}}.$$

Efron [5] proposed an alternative partial likelihood of

$$L_2(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{s}_i)}{\prod_{j=1}^{d_i}\left[\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}'\boldsymbol{z}_l) - \frac{j-1}{d_i}\sum_{l \in \mathcal{D}_i} \exp(\boldsymbol{\beta}'\boldsymbol{z}_l)\right]}.$$

The third partial likelihood due to Cox [3] is based on a discrete time hazard rate model. The discrete logistic likelihood is

$$L_3(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{s}_i)}{\sum_{\boldsymbol{q} \in \mathcal{Q}_i} \exp(\boldsymbol{\beta}'\boldsymbol{s}_q^*)},$$

where $\mathcal{Q}_i$ is the set of all subsets of d_i individuals who could be selected from the risk set $\mathcal{R}_i$ and $\boldsymbol{s}_q^* = \sum_{j=1}^{d_i} \boldsymbol{z}_{q_j}$.
The fourth alternative partial likelihood is (see DeLong et al [4])

$$L_4(\boldsymbol{\beta}) = \prod_{i=1}^{k} \left\{ \int_0^{\infty} \prod_{j=1}^{d_i} \left[1 - \exp\left(-\frac{\exp(\boldsymbol{\beta}'\boldsymbol{z}_j)}{\sum_{l \in \mathcal{R}_i^*} \exp \boldsymbol{\beta}'\boldsymbol{z}_l)} t\right)\right] \exp(-t)dt \right\},$$

where $\mathcal{R}_i^* = \mathcal{R}_i \backslash \mathcal{D}_i$ is the set of individuals whose event or censored times exceed t_i or whose censored times are equal to t_i. It is often called exact likelihood.

Note that when the number of ties is small, Breslow's and Efron's likelihoods are quite close. Of course, if no ties occur at the event times, all four likelihood functions reduce to the regular Cox's partial likelihood.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

[2] Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*, **30**, 89-99.

[3] Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society Series B*, **34**, 187-220.

[4] DeLong, D.M., Guirguis, G.H. and So, Y.C. (1994). Efficient computation of subset selection probabilities with application to Cox regression. *Biometrika*, **81**, 607-611.

[5] Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72**, 557-565.

[6] Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, New York.

[7] Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis: Applied Methods and Examples.* Springer-Verlag, New York.

[8] Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* Wiley, New York.

[9] Miller, R.G. (1981). *Survival Analysis.* Wiley, New York.

[10] Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society Series A*, **135**, 185-206.