A LIFE TABLE REGRESSION ANALYSIS FOR COMPLEX SURVEY DATA

Kevin F. O'Brien, East Carolina University

C. M. Suchindran, University of North Carolina

1. Introduction

Much of demographic research is based on data obtained through survey samples. Frequently, the event under study has an associated waiting time and can be examined using life table methods (see Shryock, Seigel and Associates, 1971). Recently the work of Cox (1972) and others on life table regression models has allowed the statistical examination of relationships between a set of covariables and the occurrence of the event of interest. However, a problem exists since an adequate set of arguments which provide a sound basis for the use of such analytic techniques for survey data has not been given. Recently a large amount of interest has been given to design based versus model based inference regarding survey data (Tomberlin, 1980; Cassel, Sarndell and Wretman, 1978 and Koch, Gillings and Stokes, 1981). The focus of such discussions being whether one relies only upon the randomization inherent in the sampling design or bases data analysis upon an assumed underlying stochastic model. Related topics such as likelihood based inference for survey have been discussed by Basu (1969, 1975), and actual estimation based upon an assumed population structure has been studied by Royall (1970). Most of this work, however, is not directed at the applied statistician and is piecemeal in the sense that specific issues are covered. The entire analysis is never viewed from start to finish. The result is that in practice many of the issues are ignored, or not considered, and analyses are performed with possibly invalid outcomes.

The purpose of this paper is to present a methodology for the use of life table regression models for analysis of survey data. The given methodology will encompass both estimation of parameters and hypotheses testing. The methodology is general enough to encompass a variety of models other than the stated life table regressions. Issues covered include the probability structure of the data, likelihood construction and estimation of parameters, and variance or mean square error (MSE) estimates. The proposed methodology is illustrated using data from the 1973 National Survey of Family Growth.

2. Probability Structure and Likelihood

The issue here is that the life table implies on underlying stochastic process governing the occurrence of events (Chiang, 1968). Such an assumed stochastic process runs contrary to the well established design based or fixed population approach to survey data. This design based approach claims that the values of the variables under study are fixed, yet unknown, for each unit in the population. The only probabilities are those generated by the randomization inherent in the selection process. Such an approach excludes consideration of life table regression models, and any analysis approach which assumes a probabilistic structure other than the sample design. However, an alternative view of survey data is available which allows considerable flexibility regarding survey data analysis. This alternative approach is the superpopulation or model based approach. Here the value of the variable under study is treated as the outcome of a random variable with a specified stochastic structure (see Cassel, Sarndell and Wretman, 1978).

The superpopulation view of survey data is the framework within which life table analysis of survey data must be couched. The main argument for this is the implied stochastic process underlying the time to events in the life table.

The occurrence of events in a life table is governed by what is termed a hazard rate. The hazard rate for a life table can be defined in terms of a non-negative continuous random variable T with values t. The definition of the hazard rate is

$$\lambda(t)=\lim_{\Delta t\to o} P(t<T<t+\Delta t|T>t)/\Delta t \ .$$

Cox (1972) proposed a hazard rate model which incorporated covariables. The Cox model for the hazard rate is

$$\lambda(t:\underset{\sim}{z}) = \exp(\underset{\sim}{z}\underset{\sim}{\beta})\lambda(t) \qquad (1)$$

where

$\lambda(t)$ is an underlying hazard; treated as a nuisance function

$\underset{\sim}{\beta}$ is a vector of unknown regression coefficients to be estimated

and

$\underset{\sim}{z}$ is a vector of known covariates which can depend on time.

The main focus of attention in this paper will be directed at using the Cox model for survey data. However, other regression methods will be mentioned.

The first assumption necessary is that events occur independently among the population units. This is a standard life table assumption, but in practice requries examination since events like infectious disease occurrences may not be independent. A related concern is with respect to intraclass correlation resulting from 'cluster sampling'. The primary effect of such correlation is reduced sample precision through a redundancy of information (i.e., individuals within clusters have similar covariate values leading to similar times to occurrence of the event under study). The intraclass correlation is a function of the sampling design alone which here is assumed independent of the stochastic process of the life table. Within this proposed framework such a correlation structure has no effect upon the independence of the events within the population and is not a factor in estimating parameters. However, the sample design will affect inference since the properties such as expected values or variances are defined in terms of the sampling design probability function. This topic is discussed further in section 3.

The above assumptions allow a likelihood to be written in a product form. A final consideration regards the sampling weights. These weights associated with each unit sampled usually impart information concerning actual or proportional frequencies of certain population characteristics. As such, their inclusion in the computation of statistics is generally warranted. Since the sample weights are frequency related, and the likelihood a product of individual terms the method proposed here is to include the sampling weights as exponents in the likelihood. Denoting the sample weights by w_j $j=1,\dots,n$ (n the sample size) examples of including the sample weights as exponents result in

$$\hat{\mu}=\sum_{j=1}^{n} w_j x_j / \sum_{j=1}^{n} w_j$$

as the estimator of the mean for a normal population and

$$\hat{\mu}=\sum_{j=1}^{n} w_j / \sum_{j=1}^{n} w_j x_j$$

for the mean of an exponential population. The above estimators are similar to Horvitz-Thompson (1952) estimators.

If an indicator random variable

$$Y_{ij}=\begin{cases} 1 & \text{the event occurs at the ith time(interval) to the jth unit} \\ 0 & \text{otherwise} \end{cases}$$

is used to denote the event of interest then a likelihood based on hazard model (1) is

$$L\{\beta:Y,\underset{\sim}{Z},\pi(r)\}=\pi(r)\prod_{i=1}^{k}\prod_{j=1}^{n}(1-p_i)^{\exp(Z_j\beta)Y_{ij}w_j} \cdot p_i^{\exp(\underset{\sim}{Z}_j\beta)(1-Y_{ij})w_j} \quad (2)$$

where

$$p_i=\exp\{-\int_{t_{i-1}}^{t_i}\lambda(\tau)d\tau\}$$

is the conditional probability of surviving the i-th time (interval). The function $\tau(r)$ is the probability function for the given sample design, r the specific sample in question; Y denotes the nx1 vector of indicator variables for the sample units. Note that $\pi(r)$ need not be known since it becomes an additive constant in the log-likelihood, and vanishes when derivatives with respect to β are taken. Other likelihoods for the Cox model (1) are available, see for example Kalbfleishch and Prentice (1980). The only unusual aspect of (2) is the inclusion of the sample weights as exponents.

3. Variance/Mean Square Error Estimation

The final issue is that of obtaining as estimate of MSE for the model parameters upon which inferences can be based. The usual definition for variance/MSE is

$$\widetilde{\mathrm{Var}}(\hat{\underset{\sim}{\beta}})=\sum_{r\varepsilon R}(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})'\pi(r),$$

where R is the set of all samples, r, under the specified design, and prime denotes the transpose operator.

A problem with complex sample designs is that the variance or MSE estimates are not easily computed. Methods like balanced repeated replications (BRR), jackknife, or Taylor Series Linearization (TSL) must be used (see Kish and Frankel, 1974). The situation here, with the life table regression coefficients, is that the estimates are found using some iterative method. The cost of finding the parameter estimates is not incidental especially if estimates must be found for a number of 'half-samples' or jackknifed samples as would be required by the BRR or jackknife; application of the TSL in this situation is not clear since no closed for estimator exists. To resolve this problem of having to compute parameter estimates for several subsets of the sample data the score statistic

$$S(\beta:Y,\underset{\sim}{Z})=\frac{\partial}{\partial\beta}\log L\{\beta:Y,\underset{\sim}{Z},\pi(r)\}$$

can be used. The score statistic is known for the life table regression model (1), and has known statistical properties when evaluated at β. Since the score statistic is known, a variance estimate for this statistic can be computed using methods like the BRR or jackknife. Viewing the score statistic as a function of β, the variance of the score statistic, Var(S), can be thought as a function of the variance of β, i.e.

$$\mathrm{Var}(\hat{S})=\left[\frac{\partial S}{\partial\beta}\bigg|_{\underset{\sim}{\beta}=\hat{\underset{\sim}{\beta}}}\right]^2 \mathrm{Var}(\hat{\underset{\sim}{\beta}}). \quad (3)$$

The equation (3) can then be used to find a variance estimate for β since both Var(S) and the partial derivatives are known. The result for a single parameter is

$$\widehat{\mathrm{Var}}(\hat{\beta})=\left[\frac{dS}{d\beta}\bigg|_{\beta=\hat{\beta}}\right]^{-2}\widehat{\mathrm{Var}}(\hat{S}). \quad (4)$$

The case of several parameters involves matrix operations in both (3) and (4) an expression which emphasizes this fact is

$$\hat{\underset{\sim}{V}}(\hat{\underset{\sim}{\beta}})=\underset{\sim}{H}(\hat{\underset{\sim}{\beta}})\hat{\underset{\sim}{V}}(\hat{\underset{\sim}{S}})\underset{\sim}{H}(\hat{\underset{\sim}{\beta}})', \quad (5)$$

where $\hat{\underset{\sim}{V}}(\hat{\underset{\sim}{\beta}})$ and $\hat{\underset{\sim}{V}}(\hat{\underset{\sim}{S}})$ are the variance/covariance matrices of $\hat{\underset{\sim}{\beta}}$ and the score statistic $\hat{S}$ respectively, and $\underset{\sim}{H}(\hat{\underset{\sim}{\beta}})$ is the inverse matrix of partial derivatives of the score statistic evaluated at β; prime denotes transpose.

An interesting result is obtained when applying the above to the single parameter case. The result is given in terms of the design effects

$$\mathrm{deff}=\frac{\text{Variance under the true design}}{\text{Variance as if a simple random sample}} \quad (6)$$

defined by Kish (1965). Under simple random sampling the correct asymptotic variance for the score statistic, disregarding any sampling fraction, is

$$\hat{V}_{SRS}(\hat{S})=dS/d\beta\bigg|_{\beta=\hat{\beta}}, \quad (7)$$

and the variance of $\hat{\beta}$ is the inverse of (7), (Kendall and Stewart, Vol. 2, 1963). Using equations (4) and (7) and definition (6) the result of the proposed methodology is

$$\hat{V}_{CSD}(\hat{\beta}) = \left[dS/d\beta\,|_{\beta=\hat{\beta}}\right]^{-1} \hat{V}_{CSD}(\hat{S}) \left[dS/d\beta\,|_{\beta=\hat{\beta}}\right]^{-1}$$
$$= \text{deff}\cdot\hat{V}_{SRS}(\hat{\beta}) \qquad (8)$$

where $\hat{V}_{CSD}$ indicates variance under the complex sampling design. It needs to be pointed out that the deff in equation (8) are in terms of the score statistic, not $\hat{\beta}$.

4. Illustration of the Methods

Data from the 1973 National Survey of Family Growth (NSFG) were used to illustrate the methods previously proposed (see French (1978) for a description of the 1973 NSFG). The event 'marital disruption' was examined for 2297 black women whose first marriage occurred at least five years before the survey date. The covariate studied was the woman's age at first marriage.

Since the methods proposed are applicable to statistical procedures other than life table regressions two other regression models were examined; a logistic model and a simple linear regression. If

$$Y_j = \begin{cases} 1 & \text{the marriage is disrupted} \\ 0 & \text{the marriage remains intact} \end{cases}$$

then the three models are:

i. Simple Linear Regression: $E(Y)=\alpha+\beta Z$

ii. Logistic Regression: $E(Y)=\left[1+\exp\{-(\alpha+\beta Z)\}\right]^{-1}$

iii. Life Table Regression: $E(Y)=1-\exp\{-\alpha\exp(\beta Z)\}$

The parameter α for the life table model is the integrated underlying hazard for the five year period immediately following marriage.

The parameters are estimated both using the sample weights in the likelihoods and ignoring the sample weights. Estimates of the variance/covariance matrices were computed under simple random sampling and under the NSFG design using Balanced Repeated Replications. Parameter estimates are given in Table 1 and variance/covariances are presented in Table 2.

The results of Table 1 indicate that for the given data including or excluding the sample weights from the likelihood has little effect on the estimates obtained. However, since the NSFG is a national survey, differences between estimates of as little as 0.001 can be considerable when translated into actual numbers of persons. A general trend noted here is for the parameter estimate to be slightly less when including the weignts than when not using them. The only noted exception here is the life table parameter α. The average absolute difference over all parameters and models was .0133.

Variance/Covariance estimates are presented in Table 2. The variance estimates obtained based upon the complex sample design are larger in every case than those estimated under the assumed simple random sample. Covariances under the complex sample design are larger also except for those of the simple linear regression.

5. Discussion

This paper has dealt with some of the issues surrounding the use of life table regression models for the analysis survey data. Arguments regarding the probability structure of the data and use of the sample weights in the estimation have been covered as well as variance or MSE estimation.

Results of the proposed methodology when used to examine data from the 1973 NSFG are encouraging, since including the weights did not substantially change the parameter estimates. A similar result was noted for ratio proportions estimated for the same data set (O'Brien, 1980).

Substantial differences were noted for variances estimates between those computed under simple random sampling and under the complex NSFG design. However, since no known criterion was available for comparison with the results, caution is warranted. Further analytic and empirical investigation are needed to firm the arguments presented and study the proposed methods of including the sample weights and computing variances.

Table 1: Parameter Estimates by Model and Weighting Scheme

	Parameter	
Model	α	β
Linear Regression		
Weighted	.5643	-.0387
Unweighted	.5707	-.0456
Logistic Regression		
Weighted	.2701	-.1603
Unweighted	.3046	-.1920
Life Table Regression		
Weighted	.4842	.0440
Unweighted	.4807	.0482

Table 2: Variance and Covariance Estimates by Model and by Sample Design

	Parameter		
Model	$\widehat{Var}(\hat{\alpha})$	$\widehat{Var}(\hat{\beta})$	$\widehat{COV}(\hat{\alpha},\hat{\beta})$
Linear Regression			
SRS	$6.44\text{x}10^{-4}$	$4.84\text{x}10^{-5}$	$-1.61\text{x}10^{-4}$
Complex	$1.58\text{x}10^{-3}$	$9.37\text{x}10^{-5}$	$-1.05\text{x}10^{-4}$
DEFF	2.46	1.94	0.65
Logistic Regression			
SRS	$1.13\text{x}10^{-2}$	$8.86\text{x}10^{-4}$	$-2.89\text{x}10^{-3}$
Complex	$3.25\text{x}10^{-2}$	$5.01\text{x}10^{-3}$	$-1.26\text{x}10^{-2}$
DEFF	2.88	5.65	4.36
Life Table Regression			
SRS	$2.15\text{x}10^{-3}$	$6.07\text{x}10^{-4}$	$-1.07\text{x}10^{-3}$
Complex	$6.23\text{x}10^{-3}$	$1.12\text{x}10^{-3}$	$-2.61\text{x}10^{-3}$
DEFF	2.90	1.86	2.44

References

Basu, D. (1969). Role of sufficiency and the likelihood principles in sample survey theory, Sankhya 31, 441-454.

Basu, D. (1975). Statistical information and likelihood, Sankhya Ser. A 37, 1-71.

Cassel, C., Sarndall, C. and Wretman, J. H. (1977). Foundations of Inference in Survey Sampling, John Wiley & Sons, New York.

Chiang, C. L. (1968). Introduction to Stochastic Processes in Biostatistics, John Wiley & Sons, New York.

Cox, D. R. (1972). Regression models and life tables (with discussion). JRSS B 34, 187-202.

French, D. K. (1978). National Survey of Family Growth, Cycle 1: Sample design, estimation procedures, and variance estimation. Vital and Health Statistics, Series 2, No. 76, National Center for Health Statistics. U.S. Government Printing Office, Washington, D. C.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, JASA 47, 663-685.

Kalbfleisch, J. D. and Prentice, R. L. (1980). The statistical analysis of failure time data. New York: John Wiley and Sons.

Kendall, M. and Stuart, A. (1963). The Advanced Theory of Statistics, Volume 2, Hafner Press, New York.

Kish, L. (1969). Survey Sampling. John Wiley & Sons., New York.

Kish, L. and Frankel, M. R. (1974). Inference from complex samples. JRSS B 36, 1-37.

Koch, G. G., Gillings, D. and Stokes, M. E.(1980) Biostatistical implications of design, sampling and measurement to the analysis of health science data. Ann. Rev Public Health, 1, pp. 163-225.

O'Brien, K. (1980). Life table analysis of survey data. Institute of Statistics Mimeo Series No. 1337, Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27514.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models, Biometrika 57, 377-387.

Shryock, M. S., Seigel, J. S. and Associates (1977). The Methods and Materials of Demography, U.S. Department of Commerce, U.S. Government Printing Office; Washington, D.C.

Tomberlin, T. J. (1980). A model based approach to the analysis of contingency tables of data from complex samples. Paper presented at the Joint Annual Statistical Meetings, Houston, Texas.