DOCUMENT RESUME

ED 383 767 TM 023 315

AUTHOR	Schumacker, Randall E.; Denson, Kathleen B.
TITLE	Interpreting Significant Discrete-Time Periods in Survival Analysis.
PUB DATE	Apr 94
NOTE	46p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Educational Research; Interaction; *Prediction; Predictor Variables; *Research Methodology; Simulation; Special Education; *Time Factors (Learning)
IDENTIFIERS	*Discrete Time Survival Analysis

ABSTRACT

Discrete-time survival analysis is a new method for educational researchers to employ when looking at the timing of certain educational events. Previous continuous-time methods do not allow for the flexibility inherent in a discrete-time method. Because both time-invariant and time-varying predictor variables can now be used, the interaction of predictors with time becomes a reality. This article presents an approach to interpreting this interaction which involves testing for significance at each discrete time period. A simulation involving 300 fictitious students in a special education work program illustrates use of the approach. Four tables and six figures illustrate the discussion. An appendix contains a program for discrete-time survival analysis. (Contains 25 references.) (Author/SLD)

INTERPRETING SIGNIFICANT DISCRETE-TIME PERIODS IN SURVIVAL ANALYSIS

Randall E. Schumacker and Kathleen B. Denson

Authors:

Randall E. Schumacker, Ph.D.
Associate Professor
Educational Research
College of Education
UNT Box 13857
University of North Texas
Denton, TX 76203-6857
(817) 565-3962
FAX: 817-565-2185
E-mail: rschumacker@unt.edu

Kathleen B. Denson
Evaluation Specialist
Dallas Independent School District
Dallas, TX

Manuscript: Submitted to Psychological Methods
Date: December 1994

Running Head: Discrete-time intervals

Acknowledgments:

Paper was previously presented at the annual meeting of the American Educational Research Association, Division D
New Orleans, LA
April, 1994

ABSTRACT

Discrete-time survival analysis is a new method for educational researchers to employ when looking at the timing of certain educational events. Previous continuous-time methods do not allow for the flexibility inherent in a discrete-time method. Because both time-invariant and time-varying predictor variables can now be used, the interaction of predictors with time becomes a reality. This article presents an approach to interpreting this interaction which involves testing for significance at each discrete time period.

INTERPRETING SIGNIFICANT DISCRETE-TIME PERIODS

IN

SURVIVAL ANALYSIS

THE HISTORY BEHIND SURVIVAL ANALYSIS

Survival analysis is a statistical technique known by many names, depending on the discipline in which it is used. Sociologists have event history analysis, engineers use failure time analysis, biostatisticians have hazard models, and economists conduct discrete time series analyses. The field of education is just beginning to use this procedure, under the name of discrete-time survival analysis, to answer questions about whether an event will occur, when it is most likely to occur, and what other variables are impacting the occurrence of the event.

Survival analysis can be traced back to the 18th century with the development of the "life table." A *life table* depicts survival/failure conditions mathematically at a particular time among a population (Darden, 1987). It can be thought of as a distribution of the time until an event occurs; death, for example, in the life tables. The method is nonparametric and has been used primarily by demographers (Pollard, Yusef, & Pollard, 1981) and insurance actuaries as a basis for measuring longevity.

The most widely used nonparametric approach to estimating the survival function is the product-limit estimator, referred to as the *Kaplan-Meier* (1958) estimate, for data that is right-censored. Censoring occurs when the event of interest does not occur to all subjects before the conclusion of the study. The product-limit estimator is the maximum likelihood estimator of the survival function when no assumption is made about its functional form (Tuma, 1982). The estimators for each period are then plotted against duration in the event state to produce Kaplan-Meier curves. This technique is commonly reported in sociology and business related literature.

There are problems however with the *life table* and *Kaplan-Meier* estimators. They both lack the ability to adequately address censoring, which generally causes the underestimation of the true expected value (Blossfeld, 1989), and because they are not regression techniques, the inability to estimate relationships of predictor variables (Allison, 1984; Blossfeld, 1989; Singer & Willett, 1991). Tuma (1982) cites the main weakness of the *Kaplan-Meier* estimators as a lack of control for heterogeneity across cases on causal variables.

In the late 1950's and early 1960's the mathematical theory of stochastic processes began to develop. *Panel studies* became popular in sociology during this time, although they were introduced to the field by Lazersfeld in the 1940's. Panel data refer to a collection of records of individuals at two or more points in time, gathered either prospectively or retrospectively (Tuma, 1984). The timing of the collection of data was indicated as "waves." Data from panel studies, analyzed by constructing an n-fold table, could be approached in several ways, each having advantages and disadvantages. A log-linear analysis of a contingency table was easy to perform, but all the variables must be discrete, and it may be difficult to find a sample large enough to fill each cell in the contingency table. A regression strategy allows both qualitative and quantitative variables to be used in the analysis, and is also easy to perform, but Goldberger (1964) found various problems that arise resulting from assuming that a dichotomous dependent variable is linear in the independent variables. These problems included heteroscadasticity and the inefficiency of ordinary least-squares estimators. However, the biggest problem with both the contingency table and regression approach was that the timing of events was ignored as relevant to the identification of the underlying structure causing change.

In 1972, David Cox, a British statistician, published a paper entitled "Regression Analysis and Life Tables" in which he proposed a *proportional hazard* model to express how the hazard rate depended on explanatory variables, namely:

$$h(t) = \alpha(t) + \beta_1 X_1 + \beta_2 X_2 ,$$

where $h(t)$ is the proportional hazard rate, $\alpha(t)$ is any function of time, β_1 and β_2 are parameter estimates, and the *X's* are time-constant variables. However, because *h(t)* theoretically should be greater than 0, the typical approach was to take the natural log of *h(t)* before setting it equal to the explanatory variables. The hazard model could then be written as: $\log h(t) = \alpha(t) + \beta_1 X_1 + \beta_2 X_2$. Because $\alpha(t)$ does not have to be specified, the model is considered to be partially parametric or semiparametric. It is called the *proportional hazards* model because, for any two individuals at any point in time, the ratio of their hazards is a constant. Basically, for any time t, the ratio of $h_i(t)/h_j(t) = c$, where i and j refer to distinct individuals and c may depend on explanatory variables but not on time (Allison, 1984).

Cox developed a partial likelihood method that was similar to the maximum likelihood method already in use with the *proportional hazards* model. A detailed description of the mathematics of partial likelihood estimation can be found in Allison (1984), but the general properties are as follows:

> "The method relies on the fact that the likelihood function for data arising from the proportional hazards model can be factored into two parts: One factor contains information only about the coefficients β_1 and β_2; the other factor contains information about β_1, β_2, and the function $\alpha(t)$. Partial likelihood simply discards the second factor and treats the first factor as if it were an ordinary likelihood function. The first factor depends only on the order in which events occur, not on the exact times of occurrence" (p. 37).

These estimators are asymptotically unbiased and normally distributed, but are not fully efficient due to the information lost by ignoring the timing of the event's occurrence. Efron (1977) found that this loss was so small that it had little bearing on the efficiency, assuming that censoring was not a consequence of the event studied.

Unfortunately, violations of the proportional hazards assumption occurred in several ways. The first involved the inclusion of time-varying variables in the equation, whereby hazards were no longer proportional, but became non-proportional. If there was an interaction between time and one or more of the explanatory variables, the proportional hazard assumption was also violated. The interaction model was written as:

$$\log h(t) = \alpha(t) + \beta x + cxt ,$$

where the product of x and t is one of the explanatory variables. If c is positive, the effect of time on the hazard increases linearly as x increases. When the hazards were not proportional, the effect of some variable on the hazard was different at different points in time.

Violations of this proportionality assumption can be checked both graphically and statistically. By stratifying the sample according to the categories of a variable, assuming that the influence of other covariates are identical for all categories, and transforming the survivor function, the plotted curves should differ only by a constant factor, β. If there is a change in the distance between the two plots, the proportionality assumption may be violated. A statistical test for proportionality would demonstrate that the coefficient β would not be significantly different from zero and the hazard functions of the two categories of the variable should differ only by the constant factor $\exp(\beta)$ (Blossfeld, 1989).

Although Cox's *proportional hazards* model still seems the most widely used, there are some important limitations. The first, and most significant, is the basic assumption that cancels the interaction of the variables with a time variable not in the equation. Singer and Willett (1991) state that "TIME itself is the fundamental time varying predictor", and it should not be left out. The other major limitation is the lack of a term to represent unobserved heterogeneity in the model, which has been found to be especially significant when dealing with repeated events. Hence, the emergence of discrete-time analyzes which include a time-varying predictor variable.

DISCRETE-TIME SURVIVAL ANALYSIS

Logistic regression is the method of survival analysis coming to the forefront in the 1990's, although it has been in the literature since the 1970's (Allison, 1982). A new approach to survival analysis using discrete-time measurement and logistic regression has been developed (Willett & Singer, 1991). To illustrate this approach and the ability to interpret significant time periods, a simulated data example was used.

Sample Data

Information was generated for 300 fictitious students enrolled in a special education work program. These students, all ages 20-22, were measured when they began their first job for competitive wages. Employment sites were coded based on the employer's previous experience with an employee who had a handicap (PREVHAND). Data were taken once a month for a year to see if the students were still employed at their sites, and if there had been a job coach supporting them for more than half of their on-clock time (SUPPORT).

Before using logistic regression to conduct a discrete-time survival analysis, the data structure must be transformed from the standard one-person, one-record data set (person data set) into a one-person, multiple-period data set (person-period data set). Singer and Willett (1992) have developed a SAS program that will array the data in such a fashion (Appendix). The records in the restructured person-period data set show what happened to each student during each discrete-time period when the event of interest (leaving the job) could have occurred, until it did occur, or until data collection ended (whichever came first).

The restructured data set yields one record per month per person. Each person-period record contains period-specific values of five different types of predictors: (1) the time-invariant variable, PREVHAND, whose values are constant across records for each person; (2) the time-varying predictor, SUPPORT, whose values may fluctuate from month to month (S1 - S12); (3) OCCASION, dummy variables E1 - E12, specifying the discrete-time interval to which the record refers; (4) a new dummy variable, PE1 - PE12, which reflects the effects of PREVHAND over time; and (5) another new dummy variable, SE1-SE12, which reflects the effects of SUPPORT over time.

In discrete-time survival analysis, a researcher uses the person-period data set to model the relationship between the occurrence of the event of interest (leaving the job) and the selected predictors. Because the outcome is dichotomous, logistic regression is used to model the log-odds of leaving the job (Willett & Singer, 1991).

Research Questions

The type of research questions that can be answered using the discrete-time survival method include:

1) Does the employer having previous experience with a worker who has a handicap have an effect on the length of time an employee with a handicap remains on their first job?

2) When is a student at the greatest risk of losing a job?

3) Does the presence of a job coach play a part in maintaining employment?

4) Is there interaction between the two predictor variables - PREVHAND and SUPPORT?

5) When is it most essential for the job coach to be present to maintain the employment of the student?

An interesting aspect of this method is that it uses two different types of variables. In the method, the variable PREVHAND is a time-invariant predictor, meaning that the information remains constant over time. Other examples of time-invariant predictors include sex, age of first pregnancy, and race. The other variable, SUPPORT, measured S01-S12, is a time-varying variable, meaning that over time, the presence of the job coach varied.

Statistical Model

Relationships between entire hazard profiles and one or more predictors are hypothesized in a hazard model (Willett & Singer, 1991). The predictors for this analysis were SUPPORT and PREVHAND. SUPPORT from a job coach was coded 1 if a coach was present, 0 if not, for each of the 12 monthly periods, S01-S12. PREVHAND was a dummy variable taking on two values 1 for experience and 0 for no experience. In this model, the hazard function was the outcome (employed at time t = 1 and unemployed at time t = 0), with PREVHAND and SUPPORT as potential predictors of that outcome.

Because the variables included in the analysis were measured at different levels, the sample hazard profiles must be transformed logarithmically to put all variables on the same level of measurement (Ferguson & Takane, 1989). Time is measured in discrete intervals, rather than continuous, so that a logistic transformation is appropriate. If *p* represents a probability, then logit *(p)* is the natural logarithm of [p /(1-p)]; so in this case, logit (p) can be interpreted as the conditional log-odds of leaving the job.

The Baseline Model

If h_j represents the entire log hazard profile, the relationship of the log-transformed hazard profile to the variable TIME is: $\log(h_j) = \beta_0(t)$, where $\beta_0(t)$ is termed the baseline log hazard profile, and represents the values of the outcome (the entire log-hazard function) in the population without other predictor variables. It is written as a function of time because the outcome itself, $\log(h)_j$, is an entire temporal profile (Singer & Willett, 1991). This equation can be expanded to account for specific measurements of monthly intervals to:

$$\text{logit}_e(h)_j = [\alpha_1 T_1 + \alpha_2 T_2 + \ldots + \alpha_{12} T_{12}]$$

and therefore,

$$h_j = 1 \backslash 1 + e^{-[\alpha_1 T_1 + \alpha_2 T_2 + \ldots - \alpha_{12} T_{12}]}$$

The alpha parameters are multiple intercepts, one per time period and represent the baseline logit-hazard function because it captures the time-period by time-period conditional log-odds that individuals whose covariate values are all zero will experience the event in each time period, given that they have not already done so (Singer & Willett, 1993).

Adding Predictor Variables

When predictor variables are included to control for observed heterogeneity, the equation expands, as in regression, to include them. The relationship of the log-transformed hazard profile to the predictor variable PREVHAND is:

$$\text{logit}_e(h)_j = [\alpha_1 T_1 + \alpha_2 T_2 + \ldots + \alpha_{12} T_{12}] + \beta_1 \mathit{PREVHAND}$$

and therefore,

$$h_j = 1 \backslash 1 + e^{-([\alpha_1 T_1 + \alpha_2 T_2 + \ldots + \alpha_{12} T_{12}] + \beta_1 \mathit{PREVHAND})}$$

The β_1 parameter is the slope parameter that represents the magnitude of the shift up or down between the two lines, or the vertical shift in logit-hazard associated with a one-unit difference in the predictor (Willet & Singer, 1991).

The inclusion of a time-varying predictor, such as SUPPORT, can be included in the equation as follows:

$$\text{logit}_e(h)j = [\alpha_1T1 + \alpha_2T2 + .. + \alpha_{12}T12] + \beta_1\textit{PREVHAND} + \beta_2\textit{SUPPORT}(t)$$

and therefore,

$$hj = 1 \backslash 1 + e^{-([\alpha_1T1 + \alpha_2T2 + .. + \alpha_{12}T12] + \beta_1PREVHAND + \beta_2SUPPORT(t))}$$

Because SUPPORT is a time-varying predictor, it is distinguished by the *(t)* in the variable name. The model postulates that, although the values of SUPPORT may fluctuate over time, its effect on log-hazard remain constant. The model is set up so that the time-varying predictor has a time-invariant effect (Singer & Willett, 1993).

Adding Interaction Terms

Statistical interactions can also be included in the hazards model. Cross-product terms are added to the main effects models in the same manner in which interactions are examined in multiple regression. The following equation includes the interaction between the support of a job coach and the experience of the employer:

$$\text{logit}_e(h)j = [\alpha 1T1 + \alpha 2T2 + ... + \alpha 12T12] + \beta 1PREVHAND + \beta 2SUPPORT(t) + \beta 3PREVHAND*SUPPORT$$

and therefore,

$h_j = 1\backslash 1 + e^{-(\alpha 1T1 + \alpha 2T2 + .. + \alpha 12T12) + \beta 1PREVHAND + \beta 2SUPPORT(t) + \beta 3PREVHAND*SUPPORT}$

A complete listing of these statistical models and their equations is in Table 1. Figures 1 through 6 graph each of these models. In model 1 (null model),

Insert Table 1 Here

Insert Figures 1-6 Here

PREVHAND and SUPPORT both equal 0, therefore each hazard is plotted by time period. In model 2 (main effects for PREVHAND), if PREVHAND = 0 is higher than PREVHAND = 1, then students working for employers who have not had previous experience with handicapped students are at greater risk of leaving their job. In model 3 (main effects for SUPPORT), the presence of a job coach (SUPPORT = 1) indicates that students without a job coach support are at a greater risk of leaving their job. In model 4 (main effects of PREVHAND and SUPPORT), the combined effect of both having an employer with previous experience and job coach support decreases the risk of losing a job. These form the basic models for discrete-time survival analysis.

Models 5 and 6 are interaction models that test the proportional hazards assumption. The proportional hazards model assumes that the magnitude of the slope between the two lines remains constant, or proportional over time. As in multiple regression, the interaction term can be removed from the model if the proportional hazards assumption is not violated. Singer and Willett (1991, 1993) found that the proportional hazards assumption is frequently violated, and if a violation is detected, the interaction with time remains in the model to ensure the appropriate estimation of predictor effects. It is these interactions, and at what point in time the differences between the hazard lines become significant, that are the main focus of this article.

The Hazard Functions

The hazard functions, rather than the survival functions, become the "cornerstone" of survival analysis. Singer and Willett (1993) discuss three properties of the hazard function that make it so appealing. First, it indicates whether events occur, and if so, when. The risk of the event occurring during that time period can be directly assessed; the higher the hazard, the higher the risk. Second, both censored and noncensored data are included in the calculations. Third, and what makes discrete-time survival analysis so promising and different, information on variation in the timing of events is not ignored as in other previously mentioned methods.

Parameter Estimates and Goodness-of-Fit test

Besides graphs, logistic parameter estimates, standard errors and goodness-of-fit statistics are also generated when predicting the dichotomous outcome of leaving the job or not using the time indicators and predictors. Allison (1982) demonstrated that these estimates are consistent, asymptotically efficient, and asymptotically distributed. Wright (1993) affirms that the logistic regression model works because it is a Rasch model: "Willett and Singer's technique and rationale provide support and insight for Rasch practitioners. Manual calculation and a *Facets* Rasch analysis confirm Singer and Willett's results. Linearity is assured for fitting data because their models incorporate the necessary and sufficient conditions for constructing linear measures. What is not assured is the extent to which their data cooperate in constructing this linearity, i.e., fit their model" (p. 307). Singer and Willett (1991) have also found that even though the person-period data set increases the sample size, the estimated standard errors are consistent estimators of the true standard errors.

Table 2 gives the parameter estimates and goodness-of-fit statistics for fitted discrete-time hazard models 1, 2 and 3 (see Table 1). The estimate of the α's (E01-E12) lead to fitted hazard probabilities for each discrete-time period and allow reconstruction of fitted hazard and survivor plots. These estimates are maximum likelihood estimates and also constitute the discrete limit of the better known Kaplan-Meier estimate of continuous-time hazard rate (Singer & Willett, 1993). Interpreting the parameter estimates is similar to estimating unstandardized regression coefficients. That is, if b is the coefficient, compute exp(b) (take the anti-log), which means raising the number e to the b power. The interpretation is then as follows: For each unit increase in an explanatory variable, the hazard is multiplied by its exponentiated coefficient. Further, computing $100(\exp(b)^{-1})$ gives the percentage change in the hazard with one unit change in the explanatory variable (Allison, 1984).

Insert Table 2 here

The likelihood-ratio chi-square test is a procedure that is very similar to testing for the significance of increments of R^2 when additional explanatory variables are added to a multiple regression equation. This test should be used whenever one model includes all the variables in another model, but also includes additional variables. The test statistic is constructed from a product of the maximum likelihood estimation, the maximized value of the log-likelihood function (Allision, 1984). To compare the fit of the two models, one calculates twice the positive difference between their log-likelihoods, although most computer printouts report -2 times the log-likelihood, or -2LL. This statistic will have an asymptotic chi-square distribution under the null hypothesis. In most cases, the associated degrees of freedom will be the difference between the number of variables in the two models. As with multiple regression, each model can be assessed until the best model is found.

Although many other statistics are reported on the SAS printout, one is especially noteable. The Odds Ratio column contains the antilog, or e^x. This value is the effect size and can be thought of as the ratio of 1 to the antilog value, e.g. $e^{-.104}$ and $e^{1.310}$ respectively for PREVHAND and SUPPORT in Table 2. The odds ratio for each model that would be reported on the SAS printout are as follows:

Model 2 - the addition of PREVHAND 1: 0.901
Model 3 - the addition of SUPPORT 1: 3.706

PROPORTIONALITY AND DISCRETE-TIME INTERVALS

Having postulated the discrete-time hazard model, Singer and Willett (1993) made three assumptions. The first is linearity. It is similar to linearity in regression with the addition that vertical displacements in logit hazard are linear per unit of difference in each predictor. This assumption can be checked by exploratory data analysis or statistical inference. The second assumption is that of no unobserved heterogeneity. All of the error is assumed to be accounted for by the inclusion of predictors in the model. Thus it becomes very important to choose the correct predictors and not omit relevant predictors. Model proportionality, described by Cox's Proportional Hazards model, is the third assumption. Logit-hazard profiles of various predictor models should maintain the approximate shape of the baseline profile, but shift it up or down, depending upon the sign of the b value. Other models make no allowance for violation of this proportionality assumption, but violation does occur frequently, and across many disciplines. Violations of the proportionality assumption are the rule, rather than the exception. If data is not checked, either graphically or statistically, for nonproportionality, results may be biased.

In discrete-time survival analysis, it is relatively easy to ascertain if the proportionality assumption has been violated. Their SAS program can be used to create new dummy variables, in this case PREVTIME (PE1-PE12), and SUPPTIME (SE1-SE12), both reflecting the effects of the predictors over time. These new variables are cross-products in the person-period data set between the time indicators (a_1T_1, a_2T_2, ... $a_{12}T_{12}$) and the predictor. The model equation for the effect of SUPPORT across time is:

$$\text{logit}_e(h)j = [\alpha 1T1 + \alpha 2T2 + \ldots + \alpha 12T12] * \beta 2SUPPORT(t)$$

and therefore, $h_j = 1\backslash 1 + e^{-[\alpha 1T1 + \alpha 2T2 + \ldots + \alpha 12T12] + \beta 2SUPPORT(t)}$.

This procedure will allow the data to be checked both graphically and statistically. By looking at the graphs of Model 5 and Model 6 (Figures 5 and 6), it can be seen that the proportionality assumption appears to have been violated. There are no longer proportional distances between the baseline, which represents TIME only, when PREVHAND = 0 and when the line representing the values of the employer who had previous experience with an employee with handicaps, PREVHAND = 1. There are similar differences in Model 6 between the baseline, SUPPORT = 0, and the presence of a job coach, SUPPORT = 1.

Graphically, one can see that the two regression lines are no longer proportional; but statistically, where are the differences? During which time periods is there a statistically significant difference in the distance between the two lines? When the answers to these questions are known, one can see, not only when and if events occur, but also when are critical times for the presence of predictors, and when specific intervention might encourage or discourage events to occur.

TESTING THE SIGNIFICANCE OF DISCRETE-TIME INTERVALS

The proportionality assumption in discrete-time survival analysis is analagous to the homogeneous regression assumption for the analysis of covariance (violation of this assumption in ANCOVA presents problems in interpretation because the magnitude of the treatment effects is not the same at different levels of X). If homogeneity of regression slopes is not tested, or data is not plotted, false conclusions can be drawn such that means are equal and there is no treatment effect. In discrete-time survival analysis erroneous conclusions can also be drawn if the effects of a predictor over time are not assessed, such as the constant effect of a predictor rather than a time-varying one.

In regression, the Johnson-Neyman technique (Huitema, 1980; Johnson & Fay, 1950; Johnson & Neyman, 1936; Pedhazur, 1982) is applied to ANCOVA designs with heterogeneous regression slopes to identify the values of X that are associated with significant group differences in Y. Limits of the regions of nonsignificance are computed using the quadratic equation. Pothoff (1964) has extended the Johnson-Neyman technique to the establishment of simultaneous regions of significance for all possible points.

Aiken and West (1991) have applied the Johnson-Neyman technique to find differences between regression lines at a specific point. They recommend the use of the Bonferoni procedure to adjust obtained values for the number of tests undertaken. Huitema (1980) has an extensive discussion of both the Johnson-Neyman technique and the use of the Bonferoni procedure in this context. Pohlmann (1993) has also discussed the examination of group differences on the dependent variable at specific point values of the covariate with a Johnson-Neyman analysis using SAS REG programs.

A similar approach can be used to test the significance of discrete-time intervals, or the nonproportionality of hazard profiles in discrete-time analysis. This is done by computing a t using the b and the standard error of the b. Since $t^2=F$, when df=1 (Ferguson & Takane, 1989), an F is computed and compared to the critical value of the Bonferoni F.

It will be recalled that Model 5 and 6 (Figures 5 and 6) are the models set up to test the assumption that predictors vary proportionally with the baseline model across time. The SAS program will compute parameter estimates and standard error values for the variables E1-E12, which are the α values from the null model when the value for the predictor variables PREVHAND equals 0 and SUPPORT equals 0. The same estimates and standard errors are computed for PE1-PE12, which is the new dummy variable computed as cross products between PREVHAND and TIME. The parameter estimates for PE1-PE12 reflect the value of b, which is the shift in distance between the two plots.

A t value can be computed by dividing the parameter estimate by its corresponding standard error. An F is calculated by squaring t (t^2). It will be noted that the degrees of freedom for each variable is 1, thereby making $t^2 = F$ appropriate. Each F is evaluated by consulting a Bonferoni F table for the critical value of F based on p dependent variables, J-1, and N-J-C degrees of freedom, where N=number of subjects, J=number of groups, and C=number of comparisons (Huitema, 1980). The critical value of F for this data was approximately 8.3; df = 1,286. See Table 3 for a comparison of values of β, SE β, t and F for Model 5. Based on this procedure, only three periods were found to be critical – months 1 and 2, and month 12. It may be concluded from this data that the most critical time to have an employer, who has had previous experience with a handicap employee, help an employee is at the beginning of the job and after one year.

The same procedure was applied to the parameter estimates and standard errors for Model 6 where SE1-SE12 was the cross-products of SUPPORT across TIME. In Table 4, it can be seen that significant F's (Bonferoni $F_{cv} = 8.3$ with df 1, 286) were obtained at Months 1, 2, 3, 5, 6, and 7. These results may lead to the conclusion that, for this data, the most critical times for the presence of a job coach are during the first six months on the job for the student.

The SAS computer printout also provides information that can be used to get a different check on these critical periods. A column on the printout will give the probabilities of the Wald Chi-Square indicating that time periods with significant F's also have a chi-square with a probability less than .004. This is the value of .05 divided by 12, or the probability of the Bonferoni F with 12 comparisons. This provides an adjustment for experiment-wide error rate.

Insert Tables 3 and 4 Here

Discrete-time survival analysis affords a valuable tool to educational researchers whose research questions don't fit a proportional-log hazard model that does not allow the fluctuation of the significance of the predictor variables across time. The testing of predictor variables across time for significance helps in the interpretation of when intervention should occur. This method for evaluating the significance of interactions between predictors and TIME, is "not simply nuisances, (but) can lead to richer substantive interpretation" (Singer & Willett, 1993).

REFERENCES

Aiken, L.S. & West, S.G. (1991). Multiple regression: Testing and interpreting interactions. Newbury Park: Sage.

Allison, P.D. (1984). Event history analysis: Regression and longitudinal data. Beverly Hills: Sage Publication.

Allison, P.D. (1982). Discrete-time methods for the analysis of event histories. In S. Leinhardt (Ed.), Sociological methodology. San Francisco: Jossey-Bass.

Blossfeld, H., Hamerle, A., & Mayer, K. (1989). Event history analysis. Hillsdale, NJ: Lawrence Erlbaum.

Cox, D.R. (1972). Regression models and life tables. Journal of the Royal Statistical Society. B, 34, 187-202.

Darden, W.R., Hampton, R.D. & Boatwright, E.W. (1987). Investigating retail employee turnovers: An application of survival analysis. Journal of Retailing, 63, 69-88.

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

Ferguson, G.A. & Takane, Y. (1989). Statistical analysis in psychology and education. New York: McGraw-Hill Book Company.

Goldberger, A.S. (1964). Econometric theory. New York: John Wiley & Sons.

Huitema, B.E. (1980). The analysis of covariance and alternatives. New York: John Wiley and Sons.

Johnson, P.O. & Fay, L.C. (1950). The Johnson-Neyman technique, it's theory and application. Psychometrika, 15, 349-367.

Johnson, P.O. & Neyman, J. (1933). Tests of certain linear hypotheses and their application to some educational problems. Statistical Research Memoirs, 1, 72-93.

Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457-481.

Lazersfeld, P.F. (1940). Panel studies. Public Opinion Quarterly, 4, 122-128.

Pedhazur, E.J. (1982). Multiple regression in behavioral research (2nd Ed.). New York: Holt, Reinhart and Winston.

Pollard, A.H., Yusef, F., & Pollard, G.N. (1981). Demographic techniques. Sydney: Pergamon.

Pohlmann, J.T. (1993). Some nonstandard applications of the analysis of covariance model. Multiple Linear Regression Viewpoints, 20, 20-24.

Pothoff, R.F. (1964). On the Johnson-Neyman technique and some extensions thereof. Psychometrika, 29, 241-256.

Singer, J.D. & Willett, J.B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. Journal of Educational Statistics, 18, 155-195.

Singer, J.D. & Willett, J.B. (1992). Using survival analysis in educational research: New methods for studying duration and educational transitions. Presentation at AERA, 1992 Conference.

Singer, J.D. & Willett, J.B. (1991). Modeling the days of our lives: Using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. Psychological Bulletin, 110, 268-290.

Tuma, N.B. & Hannan, M.T. (1984). Social dynamics: Models and methods. New York: Academic.

Tuma, N.B. (1982). Nonparametric and partially parametric approaches to event history analysis. In S. Leinhardt (Ed.), Sociological methodology. San Francisco: Jossey-Bass.

Willett, J.B. & Singer, J.D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. Review of Educational Research, 61, 407-450.

Wright, B.D. (1993). Rasch Measurement Transactions, 7:3, 307.

Table 1

POSSIBLE MODELS

Model 1 - Null Model	$logite(h)j = [\alpha 1T1 + \alpha 1T2 + ... + \alpha 12T12]$
Model 2 - Main effects-PREVHAND	$logite(h)j = [\alpha 1T1 + \alpha 2T2 + ... + \alpha 12T12] + \beta 1 \mathit{PREVHAND}$
Model 3 - Main effects-SUPPORT	$logite(h)j = [\alpha 1T1 + \alpha 2T2 + ... + \alpha 12T12] + \beta_2 \mathit{SUPPORT(t)}$
Model 4 - PREVHAND + SUPPORT	$logite(h)j = [\alpha 1T1 + \alpha 2T2 + ... + \alpha 12T12] + \beta 1 \mathit{PREVHAND} + \beta_2 \mathit{SUPPORT(t)}$
Model 5 - Assumption - PREVHAND * TIME	$logite(h)j = [\alpha 1T1 + \alpha 2T2 + ... + \alpha 12T12] * \beta 1 \mathit{PREVHAND}\ (t)$
Model 6 - Assumption - SUPPORT * TIME	$logite(h)j = [\alpha 1T1 + \alpha 2T2 + ... + \alpha 12T12] * \beta_2 \mathit{SUPPORT(t)}$

Table 2

Parameter estimates and goodness of fit statistics for three fitted discrete-time hazard models for special education employees data

Predictor	Model 1	Model 2	Model 3
E1	2.683	2.727	1.982
E2	2.247	2.289	1.576
E3	1.946	1.987	1.437
E4	1.775	1.815	1.308
E5	1.645	1.681	1.179
E6	1.404	1.439	0.958
E7	1.398	1.432	0.806
E8	1.314	1.345	0.454
E9	1.038	1.067	0.355
E10	1.042	1.071	0.575
E11	0.894	0.924	0.477
E12	0.001	0.045	-0.567
PREVHAND		-0.104	
SUPPORT			1.310
-2LL	4671.81	2687.87	2529.33
change in -2LL (df)		1983.94 (1)	2142.48 (1)
p		=.0001	=.0001

Model 1 - Null Model
Model 2 - Main Effects of PREVHAND
Model 3 - Main Effects of SUPPORT

Table 3
Calculation of F Values from Parameter Estimates
Model 5 - Interaction between PREVHAND and TIME

Month	Parameter Estimate	Standard Error	t	F
1	-1.5213	.3436	-4.427	**19.59**
2	-0.9372	.2908	-3.222	**10.38**
3	-0.7230	.2692	-2.685	7.21
4	-0.5861	.2531	0.333	.11
5	-0.3889	.2662	-1.460	2.13
6	-0.1229	.2729	-0.450	.20
7	0.1258	.2942	.427	.18
8	0.4083	.3605	1.132	1.28
9	0.5687	.4031	1.410	1.99
10	0.3891	.4099	.949	.90
11	0.7861	.5695	1.380	1.90
12	2.5813	.6100	4.231	**17.90**

Bonferoni F cv≅8.3 with df 1, 286
Significant F's are in bold print.

Table 4

Calculation of F Values from Parameter Estimates
Model 6 - Interaction between SUPPORT and TIME

Month	Parameter Estimate	Standard Error	t	F
1	1.5932	.5104	3.121	**9.74**
2	1.5409	.4462	3.453	**11.92**
3	1.5966	.4771	3.571	**12.57**
4	1.2235	.4571	2.708	7.33
5	1.5950	.5207	3.489	**12.17**
6	2.1905	.6391	3.427	**11.74**
7	3.7716	1.0482	3.598	**12.95**
8	-0.3260	.6284	-0.518	.27
9	-0.2776	.6163	-0.450	.20
10	1.0986	.7497	1.465	2.15
11	-0.3365	.8040	-.418	.17
12	0.5108	1.0165	.461	.21

Bonferoni F cv≅8.3 with df 1, 286
Significant F's are in bold print.

Figure Captions

Figure 1. Null Model

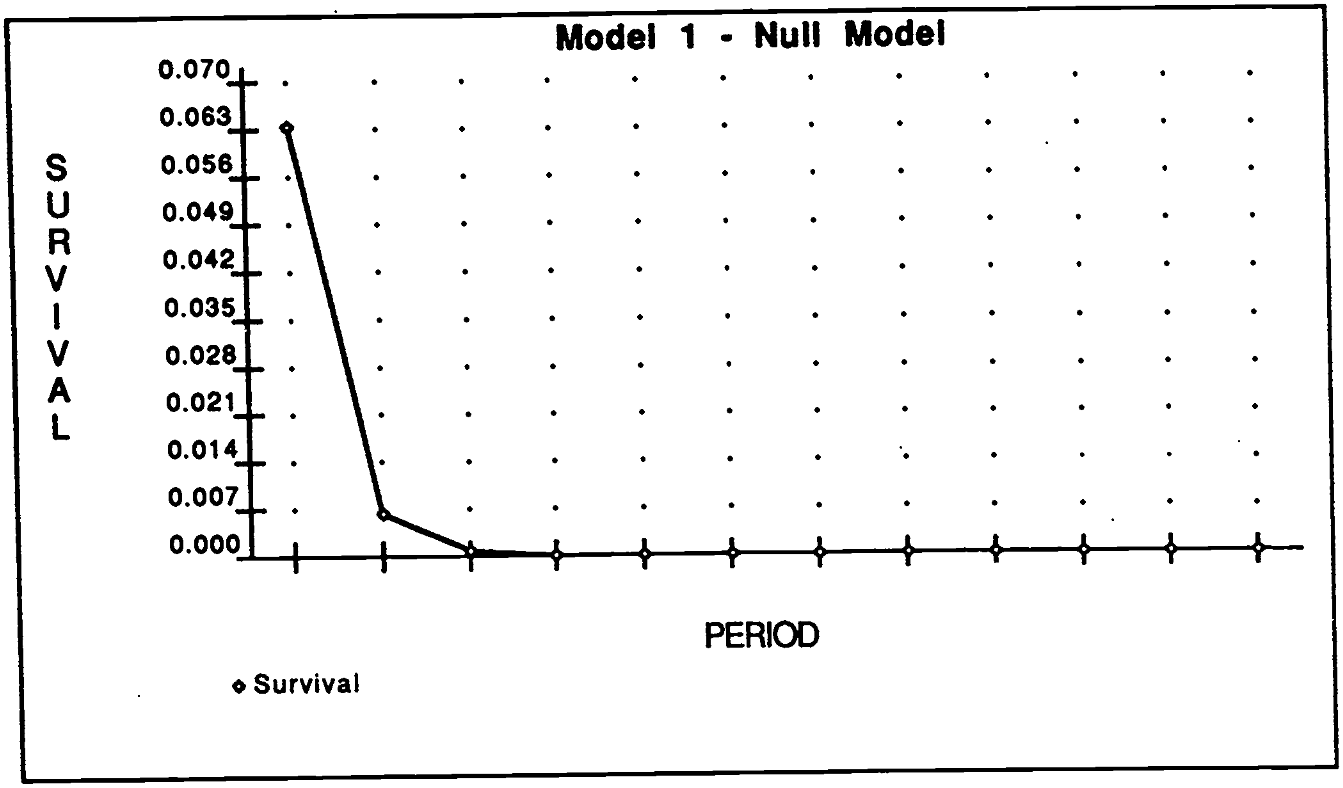
Figure 2. Main Effects of PREVHAND

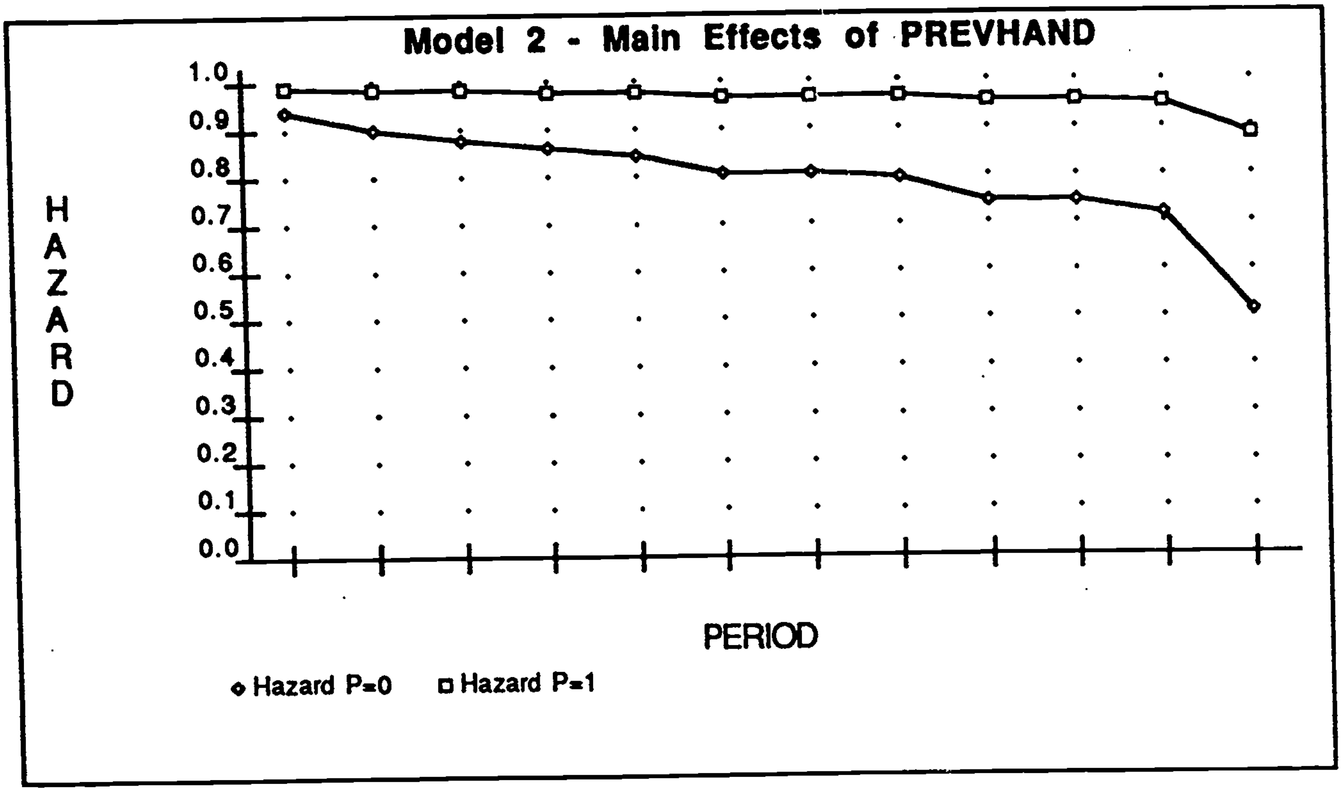
Figure 3. Main Effects of SUPPORT

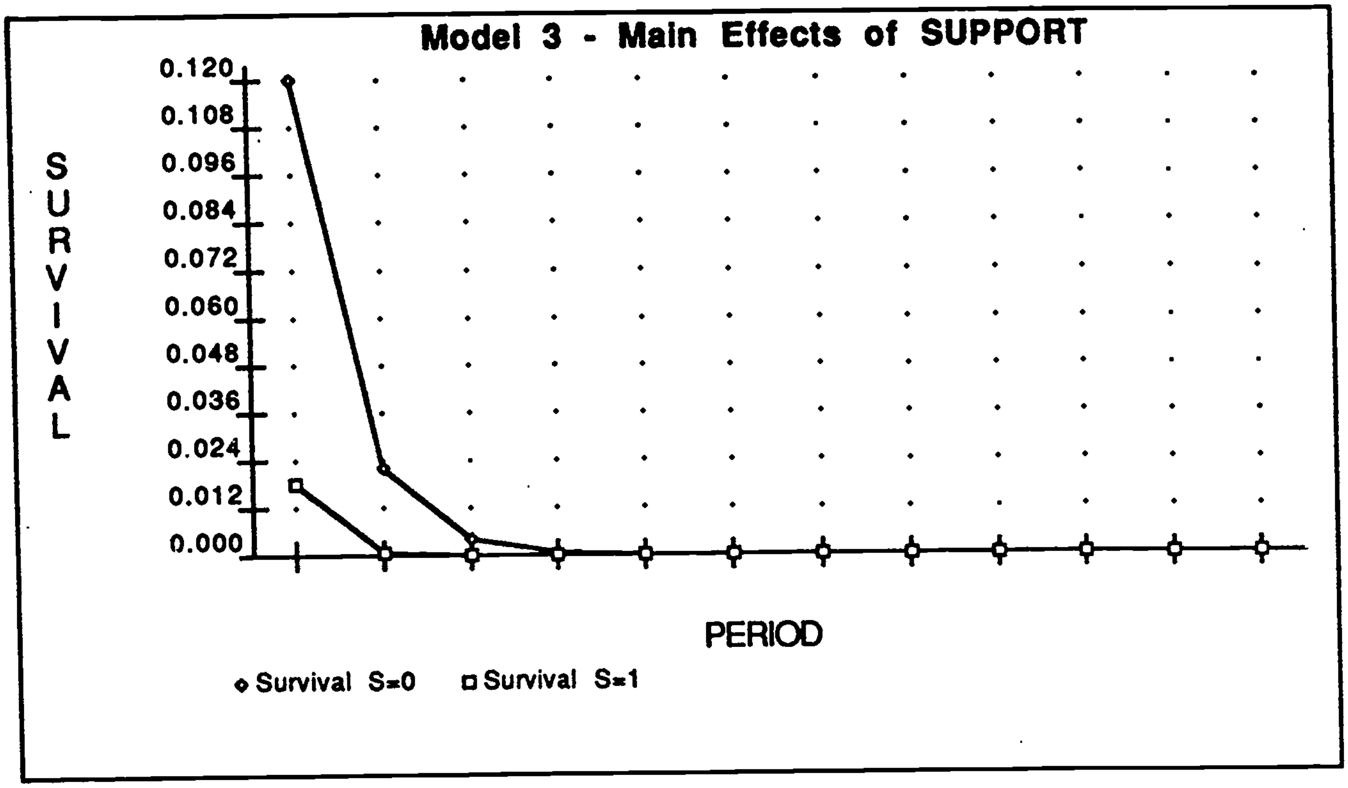
Figure 4. Main Effects of PREVHAND + SUPPORT

Figure 5. Interaction PREVHAND * TIME

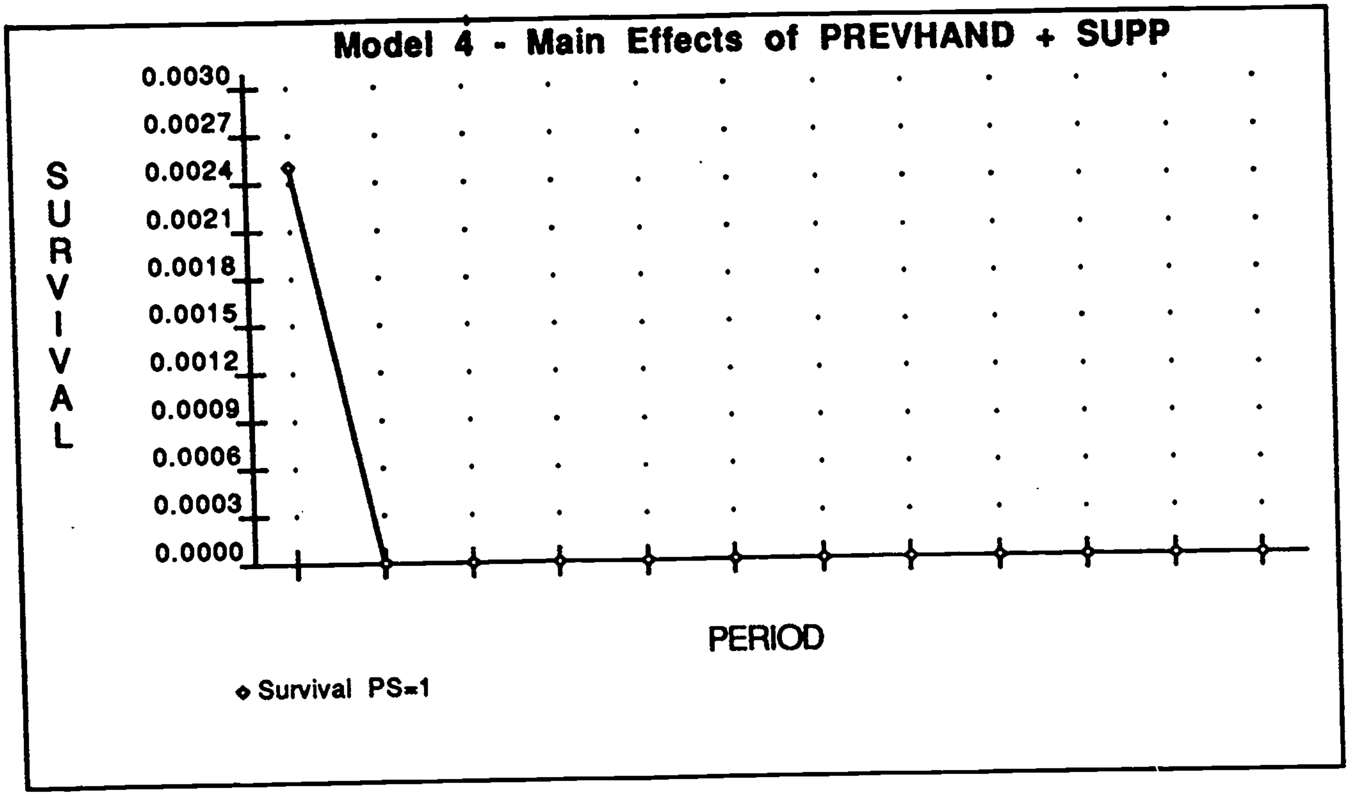
Figure 6. Interaction SUPPORT * TIME


Model 1 - Null Model
SURVIVAL
0.070
0.063
0.056
0.049
0.042
0.035
0.028
0.021
0.014
0.007
0.000
PERIOD
Survival

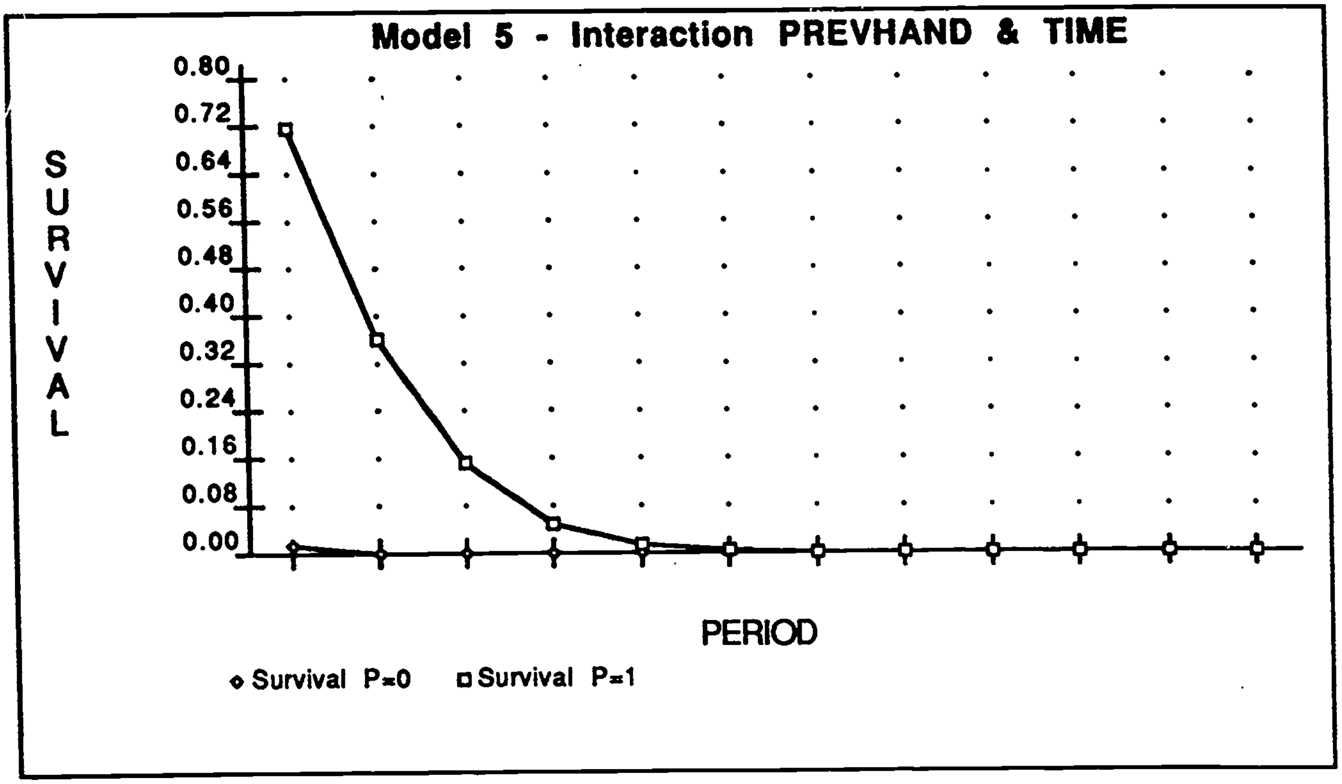

Model 2 - Main Effects of PREVHAND
HAZARD
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PERIOD
Hazard P=0
Hazard P=1

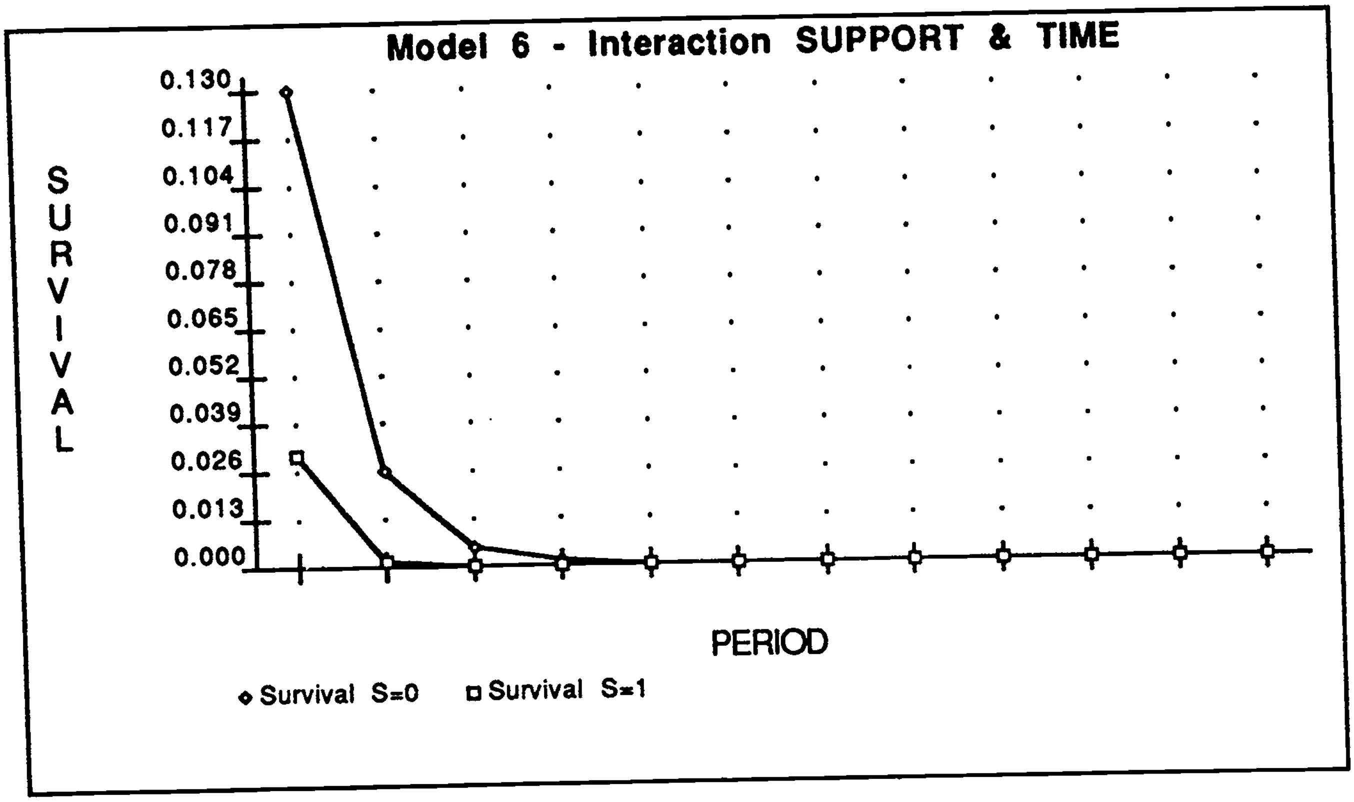


Model 3 - Main Effects of SUPPORT
SURVIVAL
0.120
0.108
0.096
0.084
0.072
0.060
0.048
0.036
0.024
0.012
0.000
PERIOD
Survival S=0
Survival S=1



Model 4 - Main Effects of PREVHAND + SUPP
SURVIVAL
0.0030
0.0027
0.0024
0.0021
0.0018
0.0015
0.0012
0.0009
0.0006
0.0003
0.0000
PERIOD
Survival PS=1


Model 5 - Interaction PREVHAND & TIME
SURVIVAL
0.80
0.72
0.64
0.56
0.48
0.40
0.32
0.24
0.16
0.08
0.00
PERIOD
Survival P=0
Survival P=1



Model 6 - Interaction SUPPORT & TIME
SURVIVAL
0.130
0.117
0.104
0.091
0.078
0.065
0.052
0.039
0.026
0.013
0.000
PERIOD
Survival S=0
Survival S=1

APPENDIX

WILLETT AND SINGER'S SAS PROGRAM FOR CONDUCTING DISCRETE-TIME SURVIVAL ANALYSIS

*** CREATING THE PERSON PERIOD DATA SET;**

```
DATA JOBSURV;
        SET JOBINFO; (Assumes the previous creation of data set JOBINFO)
        ARRAY OCCASION[12]E01-E12;
        ARRAY ASSIGN[12]S01-S12;
        ARRAY PREVTIME[12]PE01-PE12;  (Creates the variable PREVTIME)
        DO PERIOD=1 TO MIN(LASTPD,12);
                IF PERIOD=LASTPD AND CENSOR=0 THEN Y=1;
                        ELSE Y=0;
        DO INDEX=1 TO 12;
                SUPPORT=ASSIGN[PERIOD];
                IF INDEX=PERIOD THEN OCCASION[INDEX]=1;
                        ELSE OCCASION[INDEX]=0;
                PREVTIME[INDEX]=PREVHAND*OCCASION[INDEX];
END;
OUTPUT;
END;
        ARRAY SUPPTIME[12]PE01-PE12;  (Creates the variable SUPPTIME)
        DO PERIOD=1 TO MIN(LASTPD,12);
                IF PERIOD=LASTPD AND CENSOR=0 THEN Y=1;
                        ELSE Y=0;
        DO INDEX=1 TO 12;
                SUPPORT=ASSIGN[PERIOD];
                IF INDEX=PERIOD THEN OCCASION[INDEX]=1;
                        ELSE OCCASION[INDEX]=0;
                SUPPTIME[INDEX]=SUPPORT*OCCASION[INDEX];
END;
OUTPUT;
END;
```

***CREATING THE INITIAL MODEL;**

```
PROC LOGISTIC DATA=JOBSURV NOSIMPLE OUT=ESTIMATE;
TITLE2 "MODEL 1 - INITIAL (NULL) MODEL";
MODEL Y=E01-E12/NOINT MAXITER=100;
```

```
*COMPUTING FITTED HAZARD AND SURVIVAL FUNCTIONS;

DATA NEWEST;
        SET ESTIMATE;
        ARRAY OCCASION[12]E01-E12;
        SURVIVAL=1;
        DO PERIOD=1 TO 12;
                X=OCCASION[PERIOD];
                HAZARD=1/(1+(EXP(X)));
                SURVIVAL=(1-HAZARD)*SURVIVAL;
                OUTPUT;
        END;
        KEEP PERIOD SURVIVAL HAZARD;


*PRINT SURVIVAL AND HAZARD RESULTS;

PROC PRINT;
        VAR PERIOD SURVIVAL HAZARD;
        FORMAT SURVIVAL HAZARD 6.4;
PROC PLOT;
        PLOT(SURVIVAL HAZARD)*PERIOD;

*MODEL 2 - MAIN EFFECT OF PREVHAND;

PROC LOGISTIC DATA=JOBSURV NOSIMPLE OUT=ESTIMATE;
TITLE2 "MAIN EFFECT OF PREVIOUSLY EMPLOYED HANDICAPPED PERSONS";
TITLE3 "MODEL 2";
MODEL Y=E01-E12 PREVHAND/NOINT MAXITER=100;

*COMPUTING FITTED HAZARD AND SURVIVAL FUNCTIONS;

DATA NEWEST;
        SET ESTIMATE;
        ARRAY OCCASION[12]E01-E12;
        DO PREVHAND=1 TO 2;
                SURVIVAL=1;
                DO PERIOD=1 TO 12;
                X=OCCASION[PERIOD]+(PREVHAND-1)*PREVHAND;
                HAZARD=1/(1+(EXP(X)));
                SURVIVAL=(1-HAZARD)*SURVIVAL;
                OUTPUT;
                END;
END;
KEEP PREVHAND PERIOD SURVIVAL HAZARD;
```

```
*PRINT SURVIVAL AND HAZARD RESULTS;

PROC SORT;
        BY PREVHAND;
PROC PRINT;
        BY PREVAND;
        ID PERIOD;
        VAR SURVIVAL HAZARD;
        FORMAT SURVIVAL HAZARD 6.4;
PROC PLOT;
        PLOT(SURVIVAL HAZARD)*PERIOD=PREVHAND;


*MODEL 3 - MAIN EFFECT OF SUPPORT;

PROC LOGISTIC DATA=JOBSURV NOSIMPLE OUT=ESTIMATE;
TITLE2 "MAIN EFFECT OF SUPPORT OF A JOB COACH";
TITLE3 "MODEL 3";
MODEL Y=E01-E12 SUPPORT/NOINT MAXITER=100;

*COMPUTING FITTED HAZARD AND SURVIVAL FUNCTIONS;

DATA NEWEST;
        SET ESTIMATE;
        ARRAY OCCASION[12]E01-E12;
        DO SUPPORT=1 TO 2;
                SURVIVAL=1;
                DO PERIOD=1 TO 12;
                X=OCCASION[PERIOD]+(SUPPORT-1)*SUPPORT;
                HAZARD=1/(1+(EXP(X)));
                SURVIVAL=(1-HAZARD)*SURVIVAL;
                OUTPUT;
                END;
END;
KEEP SUPPORT PERIOD SURVIVAL HAZARD;

*PRINT SURVIVAL AND HAZARD RESULTS;

PROC SORT;
        BY SUPPORT;
PROC PRINT;
        BY SUPPORT;
        ID PERIOD;
        VAR SURVIVAL HAZARD;
        FORMAT SURVIVAL HAZARD 6.4;
PROC PLOT;
        PLOT(SURVIVAL HAZARD)*PERIOD=SUPPORT;

*MODEL 4 - MAIN EFFECTS OF PREVHAND AND SUPPORT;
PROC LOGISTIC DATA=JOBSURV NOSIMPLE OUT=ESTIMATE;
TITLE2 "MAIN EFFECT OF SUPPORT OF A JOB COACH";
TITLE3 "MODEL 4";
MODEL Y=E01-E12 PREVHAND SUPPORT/NOINT MAXITER=100;
```

```
*COMPUTING FITTED HAZARD AND SURVIVAL FUNCTIONS;

DATA NEWEST;
        SET ESTIMATE;
        ARRAY OCCASION[12]E01-E12;
        DO SUPPORT=2;
                SURVIVAL=1;
        DO PREVHAND=2;
                SURVIVAL=1;
                DO PERIOD=1 TO 12;
                X=OCCASION[PERIOD]+(SUPPORT-1)*SUPPORT + (PREVHAND-1)*PREVHAND;
                HAZARD=1/(1+(EXP(X)));
                SURVIVAL=(1-HAZARD)*SURVIVAL;
                OUTPUT;
                END;
                END;
END;
KEEP SUPPORT PERIOD SURVIVAL HAZARD;

*PRINT SURVIVAL AND HAZARD RESULTS;

PROC PRINT;

        ID PERIOD;
        VAR SURVIVAL HAZARD;
        FORMAT SURVIVAL HAZARD 6.4;
PROC PLOT;
        PLOT(SURVIVAL HAZARD)*PERIOD='+';

*MODEL 5 - INTERACTION BETWEEN PREVHAND AND TIME;

PROC LOGISTIC DATA=JOBSURV NOSIMPLE OUT=ESTIMATE;     (This model tests the
TITLE2 "INTERACTION BETWEEN PREVHAND AND TIME";           assumption of proportion)
TITLE3 "MODEL 5";
MODEL Y=E01-E12 PE01-PE12/NOINT MAXITER=100;


*COMPUTING FITTED HAZARD AND SURVIVAL FUNCTIONS;

DATA NEWEST;
        SET ESTIMATE;
        ARRAY OCCASION[12]E01-E12;
        ARRAY PREVHAND[12]PE01-PE12;
        DO PREVHAND=1 TO 2;
                SURVIVAL=1;
                DO PERIOD=1 TO 12;
                X=OCCASION[PERIOD]+(PREVHAND-1)*PREVHAND[PERIOD];
                HAZARD=1/(1+(EXP(X)));
                SURVIVAL=(1-HAZARD)*SURVIVAL;
                OUTPUT;
                END;
END;
KEEP PREVHAND PERIOD SURVIVAL HAZARD;
```

```
*PRINT SURVIVAL AND HAZARD RESULTS;

PROC SORT;
        BY PREVHAND;
PROC PRINT;
        BY PREVAND;
        ID PERIOD;
        VAR SURVIVAL HAZARD;
        FORMAT SURVIVAL HAZARD 6.4;
PROC PLOT;
        PLOT(SURVIVAL HAZARD)*PERIOD=PREVHAND;

*MODEL 6 - INTERACTION BETWEEN SUPPORT AND TIME;

PROC LOGISTIC DATA=JOBSURV NOSIMPLE OUT=ESTIMATE;    (This model tests the
TITLE2 "INTERACTION BETWEEN SUPPORT AND TIME";          assumption of proportion)
TITLE3 "MODEL 6";
MODEL Y=E01-E12 SE01-SE12/NOINT MAXITER=100;


*COMPUTING FITTED HAZARD AND SURVIVAL FUNCTIONS;

DATA NEWEST;
        SET ESTIMATE;
        ARRAY OCCASION[12]E01-E12;
        ARRAY SUPPORT[12]SE01-SE12;
        DO SUPPORT=1 TO 2;
                SURVIVAL=1;
                DO PERIOD=1 TO 12;
                X=OCCASION[PERIOD]+(SUPPORT-1)*SUPPORT[PERIOD];
                HAZARD=1/(1+(EXP(X)));
                SURVIVAL=(1-HAZARD)*SURVIVAL;
                OUTPUT;
                END;
END;
KEEP SUPPORT PERIOD SURVIVAL HAZARD;

*PRINT SURVIVAL AND HAZARD RESULTS;

PROC SORT;
        BY SUPPORT;
PROC PRINT;
        BY SUPPORT;
        ID PERIOD;
        VAR SURVIVAL HAZARD;
        FORMAT SURVIVAL HAZARD 6.4;
PROC PLOT;
        PLOT(SURVIVAL HAZARD)*PERIOD=SUPPORT;
```