BIOMETRICS *31*, 863–872
December 1975

# ON THE USE OF INDICATOR VARIABLES FOR STUDYING THE TIME-DEPENDENCE OF PARAMETERS IN A RESPONSE-TIME MODEL

CHARLES C. BROWN

*Biometry Branch, National Cancer Institute, Bethesda, Maryland 20014 U.S.A.*

## 1. INTRODUCTION

The study of the dependence of response-time data on a multivariate regressor variable in the presence of arbitrary censoring has been approached in a number of ways. The exponential regression model proposed by Feigl and Zelen [1965] and extended by Zippin and Armitage [1966] and by Mantel and Myers [1971] to the case of arbitrarily right censored data relates the reciprocal of the exponential parameter, i.e. the expected survival time, to a linear function of the regressor variables. Later, Glasser [1967] proposed an exponential model in which the logarithm of the exponential parameter was assumed to be a linear function of the regressor variables. In both formulations the rather stringent assumption of a constant hazard may be dropped by the assumption of a more general response-time distribution such as the Weibull, gamma or Gompertz, each of which contains the exponential as a special case. The nonparametric model proposed by Cox [1972] admits an arbitrary response-time distribution and, for discrete data, becomes a logistic regression model. An alternative version of Cox's discrete model has been proposed by Kalbfleisch and Prentice [1973]. These approaches have the advantage of not specifying the hazard function in advance and, as such, are more robust than the above parametric methods. Their major drawback, however, is the computational difficulties in the presence of tied response times. In many practical situations the data are recorded in such a way as to make this a very real problem and serious enough to imply that an alternative procedure may be desirable.

This logistic regression model was also used by Myers *et al.* [1973] in conjunction with the assumption of a constant hazard. The model they considered incorporated concomitant information by assuming that the probability of responding within a unit time period followed a logistic regression function, while the actual time to response followed a particular distributional form. They chose a form which assumed a time-independent risk of responding—the exponential for a continuous time process or geometric for discrete time. This approach was extended by Hankey and Mantel [1974] by the addition of a time function to the logistic regression function. This time function was approximated by a low order polynomial.

Inherent in these exponential and logistic regression models is the assumption that the effects of the covariates are independent of time. The exponential model of Feigl and Zelen relates the expected survival time to the concomitant information and, since the exponential distribution is "without memory," the expected remaining survival time given survival up to some time point $T$ has the same relationship to the concomitant information no matter what the value of $T$. The logistic regression methods that have been proposed allow the underlying hazard to be a function of time but the relative effects of the covariates upon the probability of response have been assumed to be independent of time. The purpose of this paper is to develop a method for the study of this assumption of time-independence of the covariates.

## 2. THE LOGISTIC-HAZARD MODEL

The model that I propose to use for the study of the time-dependence of covariate information is a variation of the logistic-exponential model proposed by Myers *et al.* In the development of this model it will be assumed that the time dimension is divided into equal-width intervals and that interval within which the response occurs is observable rather than the exact time of response. This assumption is not too restrictive since the width of the time intervals may be made arbitrarily small and the assumption does correspond to most real-world applications. Following Myers *et al*, for the $i$'th individual within the $t$'th time interval, I define a response variable $Y_{ti}$ which takes on two values, 0 for non-response and 1 for response. Letting $X_i$ represent the univariate or multivariate concomitant information for this individual, the logistic regression method relates the covariates to the conditional probability of response within the $t$'th time interval given non-response up to this point by,

$$\log \frac{P(Y_{ti} = 1 \mid X_i , E_t)}{P(Y_{ti} = 0 \mid X_i , E_t)} = g(X_i , t, \beta), \tag{1}$$

where $E_t$ is the event $Y_{1i} = 0, \cdots , Y_{t-1,i} = 0$ and $(X_i , t, \beta)$ is some specified real function of $X_i$, $t$, and an unknown parameter vector $\beta$. It should be noted that $P(Y_{ti} = 1 \mid X_i , E_t)$ is simply the hazard function for this discrete formulation and can be written as

$$P(Y_{ti} = 1 \mid X_i , E_t) = P_{i,t} = \exp\, (g(X_i , t, \beta))[1 + \exp\, (g(X_i , t, \beta))]^{-1}. \tag{2}$$

The likelihood of the $i$'th individual responding during interval $t$ is $P_{i,t} \prod_{s<t} Q_{i,s}$ while the likelihood of not responding through interval $t$ is $\prod_{s \le t} Q_{i,s}$ where $Q_{i,s} = 1 - P_{i,s}$. For a sample of $n$ individuals, the $i$'th being observed for $t_i$ intervals at which point he either responds ($Z_i = 1$) or does not respond ($Z_i = 0$) and thereafter is lost to follow-up, the likelihood is given by,

$$L = \prod_{i=1}^{n} (P_{i,t_i}/Q_{i,t_i})^{Z_i} \prod_{t \le t_i} Q_{i,t} . \tag{3}$$

In order to use such a model, the function $g(X_i , t, \beta)$ must first be specified and then the unknown parameter vector $\beta$ may be estimated by maximum likelihood (ML). Myers *et al* chose the function $g(X_i , t, \beta) = \sum_{j=0}^{r} \beta_j X_{ji}$ in which the $X_{ji}$ ($j = 0, \cdots , r$) are the regressor variable values for the $i$'th individual. The value of $X_{0i}$ is set equal to one so that $\beta_0$ represents an intercept parameter in regression terminology. It should be noted that the regression function is independent of time giving rise to a constant hazard. Hankey and Mantel incorporated time into their model by specifying the function $g(X_i , t, \beta) = \sum_{j=0}^{r} \beta_j X_{ji} + \sum_{k=1}^{s} \beta_{r+k} t^k$. This formulation allows the underlying hazard to be a specified polynomial function of time but the effects of the covariates remain time-independent. The above likelihood, which is based on a sample of $n$ individuals, the $i$'th having $t_i$ intervals of follow-up, can be considered as being based on a sample of $N = \sum_{i=1}^{n} t_i$ individuals, each having one interval of observation. In the case of a time-independent regression function, the $i$'th individual is treated as if he was $t_i$ individuals, each having the same concomitant information, while in the case of a time-dependent regression, these $t_i$ "individuals" will have differing concomitant information depending upon the time interval to which they relate. One advantage of viewing the problem in this manner is that it allows for a general computer program to handle all possible situations.

One modification can be made to the Hankey-Mantel model so as to allow the underlying hazard to be a general function of time. This consists of allowing the intercept parameter $\beta_0$ to change at each time interval in an arbitrary manner. For example, suppose that each individual is followed through a maximum of $T$ time intervals; then one can define a $(T - 1)$ dimensional auxiliary indicator regressor variable $X_i^*$ where $X_{ji}^* = 1$ for the $(j + 1)^{st}$ time interval and zero otherwise. Then the regressor function becomes

$$\begin{aligned} g(X_i\,, t, \beta) &= \sum_{j=0}^{r} \beta_j X_{ji} + \sum_{k=1}^{T-1} \beta_{r+k} X_{ki}^* \\ &= \beta_t^{\,0} + \sum_{j=1}^{r} \beta_j X_{ji}\,, \end{aligned} \tag{4}$$

where $\beta_1^{\,0} = \beta_0\,, \beta_2^{\,0} = \beta_0 + \beta_{r+1}\,, \cdots, \beta_T^{\,0} = \beta_0 + \beta_{r+T-1}$ . This results in the hazard being an arbitrary function of time and can be considered as a discrete approximation to the continuous time process—the continuous hazard being approximated by a step function with steps at the beginning of each time interval. There are, however, limitations on the number of different intercept parameters that may be included. The maximization of the likelihood over a large number of variables, besides being computationally infeasible, may lead to nonsensical results. In the extreme case where each interval contains at most one response, the pattern of responses will be completely explained by the intercept parameters and the covariate information will add nothing. Even in less extreme situations, the inclusion of many intercept parameters will lead to less precise estimates of the covariate parameters which are of prime interest. A coarser approximation may be made by grouping the time intervals into contiguous sets and allowing the hazard to vary between sets while remaining constant within each set. For the reasons outlined above, this coarse approximation was used for the example in the next section.

The $\beta_{r+k}$ $(k = 1, \cdots, T - 1)$ shown in (4) represent differences in the intercept parameters between the first and $(k + 1)^{st}$ time intervals. The model proposed by Myers *et al.* is a special case of this model with $\beta_{r+k} = 0$ for $k = 1, \cdots, T - 1$, and the extension by Hankey and Mantel to polynomial functions of time is also a special case with appropriate restrictions on the $\beta_{r+k}$, e.g. the assumption that the hazard contains a linear term of time implies the set of restrictions $\beta_{r+j} = j\beta_{r+1}$ $(j = 2, \cdots, T - 1)$.

The addition of another set of auxiliary regressor variables would allow for a time-dependent effect of any of the covariates. If it is desired to formulate a model in which the effect of the $m$'th covariate varies with time, then an additional set of regressor variables may be defined as $X_{ki}^{(m)} = X_{mi} \cdot X_{ki}^*$ for $k = 1, \cdots, T - 1$. The regression function in equation (4) would become,

$$\begin{aligned} g(X_i, t, \beta) &= \beta_t^{\,0} + \sum_{j=1}^{r} \beta_j X_{ji} + \sum_{k=1}^{T-1} \beta_{r+k} X_{ki}^{(m)}, \\ &= \beta_t^{\,0} + \beta_t^{\,m} X_{mi} + \sum_{j \neq m}^{r} \beta_j X_{ji}\,, \end{aligned}$$

where $\beta_1^{\,m} = \beta_m\,, \beta_2^{\,m} = \beta_m + \beta_{r+1}\,, \cdots, \beta_T^{\,m} = \beta_m + \beta_{r+T-1}$ .
A statistical test for the time-dependent effect of the $m$'th covariate can be ascertained by ML fitting of the models that do and do not include these additional parameters.

As an example of the contribution to the observation matrix from the $i$'th individual, consider the case of two covariates $X_{1i}$ and $X_{2i}$ taking the values $x_{1i}$ and $x_{2i}$, respectively, and let $T = 3$ and $t_i = 3$. The model which contains a time-dependent intercept and a time-dependent effect for covariate $X_{2i}$ would generate the following set of observation vectors,

| interval | $X_{0i}$ | $X_{1i}$ | $X_{2i}$ | $X_{1i}{}^*$ | $X_{2i}{}^*$ | $X_{1i}{}^{(2)}$ | $X_{2i}{}^{(2)}$ | $Z_i$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | $x_{1i}$ | $x_{2i}$ | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | $x_{1i}$ | $x_{2i}$ | 1 | 0 | $x_{2i}$ | 0 | 0 |
| 3 | 1 | $x_{1i}$ | $x_{2i}$ | 0 | 1 | 0 | $x_{2i}$ | 0 or 1 |

ML estimates of the unknown parameters and likelihood ratio tests of the significance of these parameters can be obtained through routine use of Newton-Raphson iterative procedures. The required derivatives of the log likelihood are as follows. It is assumed that $g(X_i, t, \beta) = \sum_{j=0}^{r} \beta_j X_{ji}$ where the $X_{ji}$ may represent covariate values, auxiliary variables or both. The logarithm of the likelihood in (3) is given by

$$\begin{aligned}\log L &= \sum_{i=1}^{n} \left[ Z_i \log \left( \frac{P_{it_i}}{Q_{it_i}} \right) + \sum_{t=1}^{t_i} \log Q_{it} \right] \\ &= \sum_{i=1}^{n} \left\{ Z_i \sum_{j=0}^{r} \beta_j X_{ji} - \sum_{t=1}^{t_i} \log \left[ 1 + \exp \left( \sum_{j=0}^{r} \beta_j X_{ji} \right) \right] \right\}. \qquad (5)\end{aligned}$$

Then,

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n} \left[ Z_i X_{ji} - \sum_{t=1}^{t_i} X_{ji} P_{it} \right] \qquad (6)$$

$$\frac{\partial^2 \log L}{\partial \beta_j \, \partial \beta_k} = - \sum_{i=1}^{n} \sum_{t=1}^{t_i} X_{ji} X_{ki} Q_{it} P_{it} \, .$$

## 3. AN EXAMPLE

To illustrate the use of this model for studying the time-dependence of covariate in formation upon survival, data on 1,484 cases of stomach cancer gathered from hospitals participating in the End Results Program of the National Cancer Institute will be used. These data consist of cases for which no evidence of direct extension or distant involvement of the primary tumor was found and which were treated by surgery. The complete analysis of assessing all covariates for their significance upon survival will not be given but rather an example showing the technique for studying the time dependence of one such covariate. The complete analysis showed the following five variables to be significantly associated with survival: (1) extent of primary tumor penetration, (2) size of primary tumor, (3) involvement of regional nodes, (4) sex of patient, and (5) age of patient. A description of the categories for each variable, the number of cases within these categories and the values of the indicator variables used in the logistic regression model are given in Table 1. The table shows that information on the extent of tumor penetration and the size of the primary tumor were not recorded for a large number of patients. These patients were included in the analysis, however, by use of an unknown category in order to reduce the error in estimating the effects of the other variables. The use of indicator variables for unknown information can be quite advantageous, since, for these data, only 210 of the 1,484 cases had complete information on the five variables above.

TABLE 1

DESCRIPTION OF DATA AND VARIABLES USED IN LOGISTIC REGRESSION MODELS

| Variable and Category Description | Number and Percent of Cases | | Indicator Covariates used in Logistic Regression | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ |
| Extent of Primary Tumor Penetration | | | | | | | | | | | | | | |
| Confined to mucosa | 106 | 7.1% | 0 | 0 | 0 | 0 | | | | | | | | |
| Invasion of muscularis | 148 | 10.0 | 1 | 0 | 0 | 0 | | | | | | | | |
| Involvement of serosa | 224 | 15.1 | 0 | 1 | 0 | 0 | | | | | | | | |
| Diffuse involvement | 90 | 6.1 | 0 | 0 | 1 | 0 | | | | | | | | |
| Unknown | 916 | 61.7 | 0 | 0 | 0 | 1 | | | | | | | | |
| Size of Primary Tumor | | | | | | | | | | | | | | |
| 2 cm. or smaller | 33 | 2.2 | | | | | 0 | 0 | 0 | | | | | |
| 2.1 - 3.9 cm. | 58 | 3.9 | | | | | 1 | 0 | 0 | | | | | |
| 4 cm. or larger | 165 | 11.1 | | | | | 0 | 1 | 0 | | | | | |
| Unknown | 1228 | 82.8 | | | | | 0 | 0 | 1 | | | | | |
| Involvement of Regional Nodes | | | | | | | | | | | | | | |
| No involvement | 580 | 39.1 | | | | | | | | 0 | | | | |
| Involvement | 904 | 60.9 | | | | | | | | 1 | | | | |
| Sex of Patient | | | | | | | | | | | | | | |
| Male | 892 | 60.1 | | | | | | | | | 0 | | | |
| Female | 592 | 39.9 | | | | | | | | | 1 | | | |
| Age at Treatment of Patient | | | | | | | | | | | | | | |
| 54 years or younger | 323 | 21.8 | | | | | | | | | | 0 | 0 | 0 |
| 55 - 64 years | 362 | 24.4 | | | | | | | | | | 1 | 0 | 0 |
| 65 - 74 years | 503 | 33.9 | | | | | | | | | | 0 | 1 | 0 |
| 75 years or older | 296 | 19.9 | | | | | | | | | | 0 | 0 | 1 |

The time until response, defined as death, or censorship is measured in units of one month, and the maximum time that each patient was followed, for purposes of the present analysis, is five years or 60 time intervals. Any patients that responded later than five years were treated as non-responders during the entire five-year period. The first model (Model 1) to be fitted to these data is one that includes no time dependencies, and as such can be considered as a discrete analog of the logistic-exponential model of Myers *et al.* for which the hazard function is constant. Thirteen covariates are employed, the 12 as shown in Table 1 with the addition of an intercept covariate $X_0 = 1$ for all patients. The values of the estimated parameters $\hat{\beta}_j (j = 0, \cdots, 12)$, along with their estimated standard errors, are shown in the first column of Table 2. A positive $\hat{\beta}_j$ implies an increased probability of response relative to that category having covariate values all zero. A graph of the estimated survival curve under Model 1 for the entire population of 1,484 patients is shown in Figure 1, along with the observed survival curve as derived by the actuarial method. The estimated survival curve is simply an average of the estimated survival curves for each patient. Model 1 can be considered as a model in which the response distributions of the 320 ($5 \times 4 \times 2 \times 2 \times 4$) groups of patients can be approximated by separate geometric distributions each having a constant hazard given by proper combinations of the $\beta_j$. Each of the 320 estimated survival curves, when plotted on a log scale, would be straight lines under the geometric distribution assumption, but a convex combination of such curves

TABLE 2

Log likelihoods, estimated parameter values and their estimated standard errors for each alternative model

| Covariate | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Log Likelihood | -4043.548 | -3999.252 | -4017.740 | -3994.982 |
| Intercept | | | | |
| $\hat{\beta}_0$ | -5.095±0.343 | -4.885±0.346 | -5.094±0.344 | -4.803±0.348 |
| Degree of penetration | | | | |
| $\hat{\beta}_1$ | -0.053±0.183 | -0.060±0.183 | -0.039±0.183 | -0.055±0.183 |
| $\hat{\beta}_2$ | 0.167±0.169 | 0.133±0.170 | 0.140±0.170 | 0.136±0.170 |
| $\hat{\beta}_3$ | 0.485±0.194 | 0.462±0.194 | 0.488±0.194 | 0.459±0.194 |
| $\hat{\beta}_4$ | -0.191±0.151 | -0.157±0.152 | -0.169±0.151 | -0.153±0.152 |
| Size of tumor | | | | |
| $\hat{\beta}_5$ | 0.167±0.378 | 0.112±0.378 | 0.142±0.378 | 0.112±0.378 |
| $\hat{\beta}_6$ | 0.477±0.326 | 0.478±0.326 | 0.469±0.326 | 0.476±0.326 |
| $\hat{\beta}_7$ | 0.587±0.311 | 0.620±0.312 | 0.594±0.312 | 0.619±0.312 |
| Nodal involvement | | | | |
| $\hat{\beta}_8$ | 0.907±0.071 | 0.833±0.072 | 0.988±0.089 | 0.679±0.106 |
| Sex | | | | |
| $\hat{\beta}_9$ | -0.168±0.073 | -0.161±0.073 | -0.169±0.073 | -0.164±0.073 |
| Age | | | | |
| $\hat{\beta}_{10}$ | 0.176±0.110 | 0.151±0.110 | 0.152±0.111 | 0.151±0.111 |
| $\hat{\beta}_{11}$ | 0.471±0.101 | 0.446±0.101 | 0.449±0.101 | 0.447±0.101 |
| $\hat{\beta}_{12}$ | 0.965±0.111 | 0.897±0.112 | 0.933±0.112 | 0.900±0.112 |
| Time-dependent intercept | | | | |
| $\hat{\beta}_{13}$ | | -0.003±0.084 | | -0.207±0.123 |
| $\hat{\beta}_{14}$ | | -0.293±0.106 | | -0.381±0.143 |
| $\hat{\beta}_{15}$ | | -0.774±0.142 | | -0.990±0.193 |
| $\hat{\beta}_{16}$ | | -0.904±0.162 | | -0.845±0.194 |
| Time-dependent nodal involvement | | | | |
| $\hat{\beta}_{17}$ | | | 0.181±0.114 | 0.386±0.168 |
| $\hat{\beta}_{18}$ | | | -0.211±0.158 | 0.163±0.213 |
| $\hat{\beta}_{19}$ | | | -0.512±0.208 | 0.469±0.284 |
| $\hat{\beta}_{20}$ | | | -1.129±0.300 | -0.293±0.356 |

will be convex. Therefore it was felt that, even though the observed survival curve was convex, the combination of survival curves with constant hazard might well approximate this curve. Figure 1 shows that the estimated survival curve for the total population is only very slightly convex and clearly does not approximate the observed curve. This difference in the survival curves leads to the second model (Model 2) to be fitted.

In this model the intercept parameter $\beta_0$ is assumed to vary with time. Ideally, it would be desirable to have a different intercept parameter for each time interval, but, for the reasons discussed in the previous section, inclusion of these 59 additional parameters is not practical. A compromise between desirability and feasibility was reached in the following manner. Model 2 includes a time-dependent hazard by assuming that within each year the intercept parameter is constant but is allowed to vary from year to year.

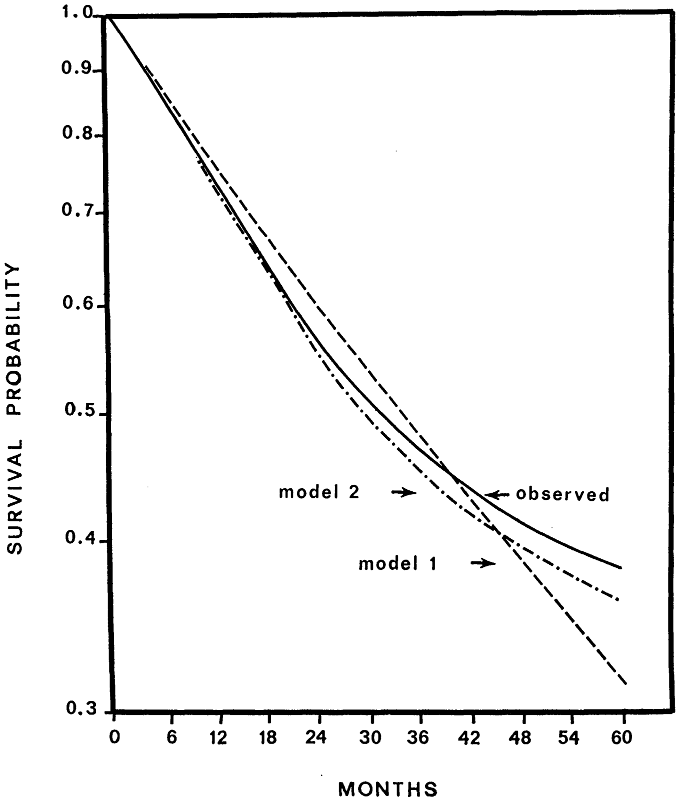

FIGURE 1
Comparison of observed survival curve with those estimated under model 1 and model 2

This means that the hazard function is to be approximated by a step function having jumps at 12-month intervals. This model contains the same covariates as Model 1 along with the inclusion of four additional covariates ($X_{13}, \cdots, X_{16}$) with corresponding parameters ($\beta_{13}, \cdots, \beta_{16}$). The indicator variable $X_{13}$ equals 1 when the time interval falls in the second year following diagnosis and 0 otherwise, $X_{14}$ relates to the third year, $X_{15}$ to the fourth and $X_{16}$ to the fifth year following diagnosis. Therefore ($\beta_{13}, \cdots, \beta_{16}$) represent changes (relative to the first year) in the intercept for the second through fifth years. A positive value of $\beta_j$ ($j = 13, \cdots, 16$) means an increased probability of response relative to the first year since diagnosis. The values of the estimated parameters $\hat{\beta}_j$ ($j = 0, \cdots, 16$) and their estimated standard errors are shown in the second column of Table 2. The estimates $\hat{\beta}_{13}, \cdots, \hat{\beta}_{16}$ imply that the underlying hazard is a decreasing function of time. A graph of the estimated survival curve under Model 2 for the entire population is shown in Figure 1. It can be seen that this model appears to fit the observed survival curve much better than Model 1. A chi-square test for the improvement in fit based on twice the difference in the maximum log-likelihood values between Models 1 and 2 yields a 4 degree-of-freedom (D.F.) chi-square value of 88.6 which is highly significant, so we would reject the hypothesis of equal intercepts. One interpretation of Model 2 is that the underlying hazard (that portion of the hazard which is not explained by the covariate information) changes with time—for these data the hazard diminishes with time—while the hazard associated with the covariates remain fixed over time. Although this model improves the fit to the data, further analyses that consider time dependencies for the covariates should be made.

Chi-square tests based on the $\hat{\beta}_j$ and their covariance matrix as estimated under Model 2 were calculated to test the prognostic significance of each of the five variables. The covariance matrix was estimated by the negative inverse of the matrix of second partial derivatives of the log-likelihood evaluated at the ML estimates $\hat{\beta}_j$. The results of these tests are shown below.

| *Variable* | *Chi-Square* | *D.F.* |
|---|---|---|
| extent of penetration | 23.1 | 4 |
| size of tumor | 9.2 | 3 |
| nodal involvement | 135.2 | 1 |
| sex | 4.2 | 1 |
| age | 78.8 | 3 |

The effects of all five variables are significant beyond the 0.05 level with involvement of regional nodes being the overwhelmingly most significant single variable. The decision to examine the possible time-dependency of this variable, involvement of regional nodes, leads to the formulation of Model 3. For this model it is assumed that the intercept is constant over time while the effect of nodal involvement is allowed to vary. This requires the addition of four indicator covariates ($X_{17}, \cdots, X_{20}$) along with their corresponding parameters to those of Model 1. The possible values of this set of covariates are similar to those of the set ($X_{13}, \cdots, X_{16}$) except that they are non-zero only for those cases having positive nodal involvement. The ($\beta_{17}, \cdots, \beta_{20}$) represent changes in the hazard for those individuals with positive nodal involvement and, as such, may be thought of as changes over time in the effect of nodal involvement, assuming a constant underlying hazard. The values of all the parameters and their standard errors as estimated under Model 3 are shown in the third column of Table 2. Estimates of the changes in the effect of nodal involvement indicate that the hazard increases slightly in the second year and then decreases steadily thereafter which, in conjunction with $\beta_8$, would imply a weakening effect of nodal involvement upon the conditional probability of response. Use of the likelihood-ratio test for the improvement in fit of Model 3 over Model 1 yields a chi-square value of 51.6 with 4 D.F. which, if Model 2 had not been fitted, could lead to the conclusion that the effect of positive nodes does indeed vary with time. However, Model 2 indicates that the underlying force of mortality varies with time while the effect of positive nodes remains fixed. In order to reconcile these two models, one more model must be fitted to the data.

This model, Model 4, includes both a time-dependent intercept and a time-dependent effect of regional nodes and, hence includes all the auxiliary variables ($X_{13}, \cdots, X_{20}$). The values of the estimated parameters $\hat{\beta}_j$ ($j = 0, \cdots, 20$) and their standard errors are shown in the fourth column of Table 2. A complete analysis of the time-dependent effect of nodal involvement can be summarized by the following 4 D.F. likelihood-ratio chi-square tests:

| | *Chi-Square* |
|---|---|
| Model 3 vs. Model 1 | 51.6 |
| Model 4 vs. Model 3 | 45.5 |
| Model 4 vs. Model 2 | 8.5 |

The comparison between Model 3 and Model 1 can be interpreted as a test of the null hypothesis that, for those cases with positive nodes, the force of mortality is constant in time, while the comparison between Model 4 and Model 3 can be interpreted as a test of the null hypothesis that, conditional on the previous hypothesis being false (i.e., for those cases with positive nodes the force of mortality does vary with time), the force of mortality does not vary for those *without* positive nodes. The first two chi-square tests above indicate that both null hypotheses would be rejected and we would conclude that, for both nodal involvement groups, the force of mortality does indeed vary over time. This leads to the last test. The comparison between Models 2 and 4 can be interpreted as a test of the null hypothesis that, given the force of mortality changes with time for both groups of patients, it changes no differently for those with and without positive nodes. The chi-square value of 8.5 leads to acceptance of this hypothesis at the 0.05 level of significance and we would conclude that the effect of nodal involvement is independent of time. It should be noted that for Model 4 the parameters $\beta_{17}, \cdots, \beta_{20}$ represent differences in the time-dependent hazard between those with and without positive nodes. The estimated differences are moderate in magnitude and show no consistent trend which strengthens the conclusion of no difference in the time-dependent hazards.

## 4. DISCUSSION

The general logistic regression model as described in the previous sections is not meant to be biologically valid for describing the processes underlying the survival experience of a group of cancer patients. However, the model should represent a reasonable approximation to the results of these processes and has the decided advantage of being computationally feasible for moderate to large sets of data. It should be noted that the assumption of different intercept parameters for each time interval is similar to a modification of Cox's approach as proposed by Breslow [1972], in which the underlying survival distribution is parameterized as continuous but having constant hazard between each pair of distinct response times. Ideally, it would be nice to estimate any time dependencies by inclusion of parameters for each time interval of observation but it has been shown that a coarse step function approximation to the hazard may lead to a reasonable fit to the data. One generalization to the step function approach could be to use segmented straight lines (Bellman and Roth [1969]) or the fitting of cubic splines (Poinier [1973]). Either technique would result in a smoothed estimate of a time-dependent hazard. Another generalization could be to employ the step function approach along with an additional time-scale parameter $W$ as done in Myers *et al.* This would generalize the model so that the regression for the log odds of responding within a unit of time becomes the log odds of responding within time $W$.

Since the model treats one individual having $t$ time intervals of observable data in the same manner as $t$ individuals having identical concomitant information, but each being observed for one time interval, it is particularly simple to allow the concomitant information to change over time. This idea was proposed by Mantel and Byar [1974] in their analysis of heart transplant data in which the subject's status could change from an untreated to a treated status. One further advantage to a model which allows the underlying hazard to vary over time is that it should lead to improved estimates and significance tests of the effect parameters. The assumption of a constant hazard when false would lead to biased results. As seen in Section 3, care must be exercised in the use of this technique for analyzing time dependencies. Fitting only a subset of the possible models could easily lead to erroneous conclusions. A variety of models should be fit to the data, if more proper interpretations are to be made.

## ACKNOWLEDGMENTS

The author wishes to thank Dr. David Byar for his encouragement and support of this research and two referees whose suggestions led to significant improvements.

## REFERENCES

Bellman, R. and Roth, R. [1969]. Curve fitting by segmented straight lines. *J. Amer. Stat. Assoc. 64,* 1079–84.

Breslow, N. [1972]. Contribution to the discussion on the paper of D. R. Cox cited below.

Cox, D. R. [1972]. Regression Models and Life Tables (with discussion). *J. Roy. Stat. Soc. (B) 34,* 187–220.

Feigl, P. and Zelen, M. [1965]. Estimation of exponential survival probabilities with concomitant information. *Biometrics 21,* 826–38.

Glasser, M. [1967]. Exponential survival with covariance. *J. Amer. Stat. Assoc. 62,* 561–8.

Hankey, B. F. and Mantel, N. [1974]. A logistic regression analysis of response-time data where the hazard function is time dependent. (Submitted for publication).

Kalbfleisch, J. D. and Prentice, R. L. [1973]. Marginal likelihoods based on Cox's regression and life table model. *Biometrika 60,* 267–78.

Mantel, N. and Byar, D. [1974]. Evaluation of response-time data involving transient states: An illustration using heart transplant data. *J. Amer. Stat. Assoc. 69,* 81–6.

Mantel, N. and Myers, M. H. [1971]. Problems of convergences of maximum likelihood iterative procedures in multiparameter situations. *J. Amer. Stat. Assoc. 66,* 484–91.

Myers, M. H., Hankey, B. F. and Mantel, N. [1973]. A logistic-exponential model for use with response-time data involving regressor variables. *Biometrics 29,* 257–69.

Poinier, D. J. [1973]. Piecewise regression using cubic splines. *J. Amer. Stat. Assoc. 68,* 515–24.

Zippin, C. and Armitage, P. [1966]. Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. *Biometrics 22,* 665–72.

*Received April* 1974; *Revised March* 1975

*Key Words*: Time-dependent parameters; Logistic survival model; multivariate logistic regression.