REVIEW

Regression Models in Clinical Studies: Determining Relationships Between Predictors and Response[1,2]

Frank E. Harrell Jr,[3] *Kerry L. Lee,*[3] *Barbara G. Pollock*[3]

Multiple regression models are increasingly being applied to clinical studies. Such models are powerful analytic tools that yield valid statistical inferences and make reliable predictions if various assumptions are satisfied. Two types of assumptions made by regression models concern the distribution of the response variable and the nature or shape of the relationship between the predictors and the response. This paper addresses the latter assumption by applying a direct and flexible approach, cubic spline functions, to two widely used models: the logistic regression model for binary responses and the Cox proportional hazards regression model for survival time data. [J Natl Cancer Inst 1988;80:1198-1202]

Regression models are being applied to clinical and epidemiologic studies with increasing frequency to assess therapeutic efficacy, study risk factors, explore prognostic patterns, and derive predictions for individual patients, among other uses. Two models have come to the forefront because they accommodate the types of responses that commonly occur in clinical studies: binary responses, for example in-hospital death or presence/absence of a certain condition, and censored continuous responses such as the time until death or therapeutic response in a sample of patients not all of whom may have died or responded.

For a given individual, let $X_1, X_2, \ldots, X_p$ denote a set of predictor or descriptor variables. For a binary response variable Y with values 0 or 1, the logistic regression model (*1,2*) is stated in terms of the probability that the event $Y = 1$ occurs given the descriptor values $X = \{X_1, \ldots, X_p\}$:

$$\text{Prob}\{Y = 1 \mid X\} = [1 + \exp\{-(B_0 + B_1X_1 + B_2X_2 + \ldots + B_pX_p)\}]^{-1},$$

where $B_1, \ldots, B_p$ are weights or regression coefficients for the descriptors, B_0 is an "intercept," and exp(u) is the natural antilogarithm of u. The logistic model has a major advantage over older methods such as discriminant analysis in that it is a direct probability model with no assumptions about distributions of the variables.

Other versions of the logistic model are available for ordinal or polytomous responses. The logistic models include as special cases the Pearson Chi-square test, the Mantel-Haenszel Chi-square test for case-control studies (*3*), and the Wilcoxon two-sample rank test (*4*).

The logistic model can be restated as a linear model in the logit of the probability p that $Y = 1$, where logit denotes $\log[p/(1-p)]$ or the natural log of the odds that $Y = 1$ versus $Y = 0$ given the value of X:

$$\text{logit}\{Y = 1 \mid X\} = \text{logit}[\text{Prob}\{Y = 1 \mid X\}] = B_0 + XB,$$

where XB denotes the weighted sum of X's, $B_1X_1 + \ldots + B_pX_p$.

The Cox proportional hazards model (*5*) is the most widely used method for analyzing survival data. Let T denote a response variable representing the time until a clinical end point. The Cox model can be stated in terms of the survival function or the probability that the event will not occur before time t (i.e., that it will occur after time t):

$$S(t \mid X) = \text{Prob}(T > t \mid X) = S_0(t)^{\exp(XB)},$$

where $S_0(t)$ is the "underlying" survival function or the survival for a "standard" individual. A significant advantage of the Cox model is that it is "semiparametric"; the underlying survival curve $S_0(t)$ is arbitrary and unspecified. The model

[1]Received April 27, 1988; accepted June 8, 1988.

[2]Supported by contracts CM-67949 and CM-67945 from the Division of Cancer Treatment (National Cancer Institute) and research grant HL-17670 (National Heart, Lung, and Blood Institute), National Institutes of Health, Department of Health and Human Services; by research grants HS-04873 and HS-05635 from the National Center for Health Services Research and Health Care Technology; and by a grant from the Robert Wood Johnson Foundation.

[3]Clinical Biostatistics, Duke University Medical Center, Durham, NC 27710; and Takima West Corporation, Chapel Hill, NC 27514.

includes as special cases the Mantel-Haenszel log-rank test (*6*), Kaplan-Meier survival estimation (*6–8*), and the conditional (stratified) logistic model (*9*). The Cox model can alternatively be stated in terms of the hazard function at time t (also called the force of mortality or instantaneous failure rate):

$$h(t \mid X) = h_0(t) \exp(XB),$$

where $h_0(t)$ is the hazard function for a "standard" individual. The two equivalent formulations of the model can be transformed to yield models that are linear in X:

$$\log\{-\log[S(t \mid X)]\} = \log\{-\log[S_0(t)]\} + XB, \text{ or}$$
$$\log[h(t \mid X)] = \log[h_0(t)] + XB.$$

Therefore the Cox regression model is linear in X with respect to $\log\{-\log[S(t|X)]\}$ or the log of the hazard function.

Regression coefficients in Cox and logistic models are estimated using the method of maximum likelihood.

Types of Assumptions

There are four kinds of regression model assumptions:

1. The study subjects are a random sample from the population about which inference is to be drawn, with independent observations.
2. The distribution of the response variable has certain properties. [The logistic model has no such assumption, and the Cox model assumes that the hazard functions for two individuals are proportional over time or equivalently that the log[−log survival] or log hazard curves for two individuals are equidistant over time.]
3. The function relating a predictor to the response has a certain shape. For example, the logistic model in its simplest form assumes that a continuous predictor is linearly related to the log odds of the outcome.
4. Predictors act in an additive fashion unless "interaction" terms are included in the model.

The remainder of the discussion deals only with assumption No. 3.

Determining the Shape of the Regression Function

Regression models stated in their simplest form assume that some property of the response variable is linearly related to the predictors, but there is no a priori justification for this assumption. There are four major statistical philosophies for estimating the true shape of the regression function or for assessing whether a postulated shape is correct. The first method involves adding terms to the model that are powers of the basic predictor variables. For example, one might add age^2 to a model containing age as a predictor. Although polynomials fit some nonlinear relationships well, there are many relationships (logarithmic or threshold effects, for example) that are not well described by polynomials (*10*). Additionally, data points in one small region can have undue influence on the global shape of the fitted polynomial, and high-order polynomials have undesirable peaks and valleys.

The second method for modeling relationships is to fit a nonparametric regression function with no prespecified shape (*11,12*). Although this method has the advantage of eliminating the need for specifying a functional form, it allows only for limited statistical inference, does not provide confidence limits, and produces only tabulated or graphic estimates (i.e., there is no predictive equation).

The third approach involves fitting a standard (e.g., linear) relationship and then graphically assessing failures in the fit of the model using a kind of "residual" or deviation of observed from estimated responses (*13*). Although this method can be used to identify departures from a hypothesized model, it does not allow formal statistical testing nor does it lead directly to correcting the model.

The final method is based on piecewise polynomials or splines to represent the relationship between a predictor and the response (*14,15*). Splines are smooth functions that can take on virtually any shape. The type of spline that is generally most useful is the cubic spline function that is restricted to be smooth at the junction of each cubic polynomial. As an example, consider a cubic spline in the predictor X = age in years with join points or knots at 6, 21, and 65 years. A logistic model with age as a predictor, assuming almost nothing about the shape of the relationship, is as follows:

$$\text{logit}\{Y = 1 \mid X\} = B_0 + B_1X + B_2X^2 + B_3X^3 + B_4(X-6)_+^3 + B_5(X-21)_+^3 + B_6(X-65)_+^3,$$

where $(u)_+^3 = u^3$ if $u > 0$, 0 otherwise, e.g., ignore the term $(X-6)^3$ if $X \leq 6$.

Ordinary cubic spline functions do have one undesirable property, namely instability in the tails of the fit, i.e., before the first knot or after the last knot. Stone and Koo (*16*) and Devlin and Weeks (*10*) advocate placing an additional restriction that the function be linear in the tails. Although one usually employs more than three knots, as an example consider the previous spline function specified with linear tail restrictions yielding:

$$\text{logit}\{Y = 1 \mid X\} = B_0 + B_1X + B_2X',$$

where

$$X' = (X-6)_+^3 - 59(X-21)_+^3/44 + 15(X-65)_+^3/44.$$

The constants 15, 44, and 59 come from the spacing between knots. Aside from the intercept B_0, a restricted cubic spline function in k knots requires estimating k−1 regression coefficients, as opposed to one coefficient if linearity is assumed or k+3 coefficients if an unrestricted cubic spline is fitted. Figure 1 displays some of the variety of shapes of restricted cubic spline functions with only four knots.

Cubic spline regression models have many advantages over the other three approaches listed earlier:

1. Since they are piecewise polynomials, splines can be fitted using any existing regression program once certain derived predictors are constructed. Thus, flexible forms for the relationship between predictor and response can be specified with equal

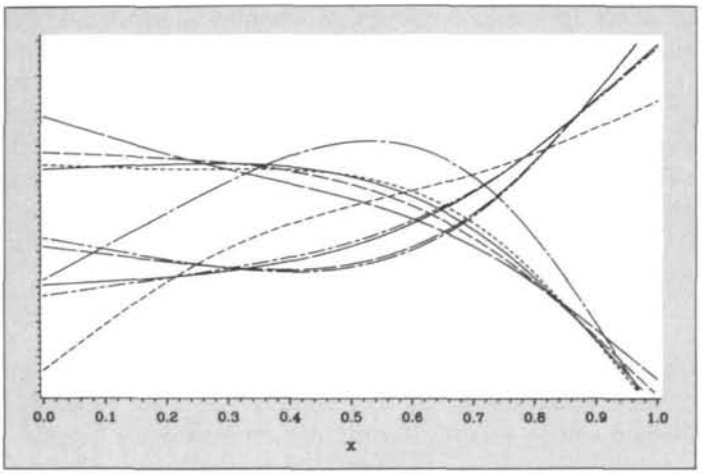

Figure 1. Some restricted cubic spline functions with 4 knots at x = .05, .25, .75, .95.

ease in ordinary multiple linear regression, logistic models, Cox survival models, parametric survival models, and all other multiple regression models.

2. Estimates of the coefficients of the spline function are derived using standard techniques, so statistical inferences can readily be drawn. For example, if age is modeled with four knots, the null hypothesis that age is related to response (without assuming linearity) can be tested with 3 degrees of freedom. The test for whether age affects the response (or log odds, log hazard) linearly has 2 degrees of freedom. In addition, confidence limits for the estimated spline function are readily computed.
3. The fitted spline function is a direct estimate of the transformation that a predictor requires to yield linearity. The graph of the fitted spline function frequently suggests a simple transformation (e.g., logarithm). The predictor can then be replaced with this transformation and fitted as a single term in the model. [Adequacy of this transformation can be tested by fitting a spline function to the transformed variable and testing its linearity.]
4. When the graph of the fitted spline function does not suggest a simple predictor transformation, the spline function itself can be used to represent the predictor in the overall model, and a predictive equation is still available.

Splines do assume that the regression function is smooth unless knots are closely spaced (*15*). Splines do have the disadvantage that the number and location of knots must be specified. However, four or five knots are usually adequate, and the fit is not greatly affected by altering knot placement (*17*). For the vast majority of cases, knots can be placed automatically at fixed percentiles of the predictor. If five knots are used, they can often be placed at the 5th, 25th, 50th, 75th, and 95th percentiles.

Examples

In this section we present examples of how restricted cubic splines can be used to model the relationship between predictors and the response in the context of logistic and Cox proportional hazards models. *See* references *10* and *16* for more examples of spline logistic models.

The first example comes from the Duke University Cardiovascular Disease Databank, which is comprised of data from patients undergoing cardiac catheterization for chest pain from 1969 to the present. A sample was drawn from the databank consisting of 1,518 patients found to have significant coronary disease (≥75% diameter narrowing of at least one major coronary artery). The response variable TVDLM (defined as three-vessel or left main coronary disease) is the presence or absence of severe coronary disease. TVDLM is coded as 1 or 0 for presence and absence of disease, respectively. There were 767 patients with TVDLM = 1. The predictor is the duration of symptoms of coronary disease. Figure 2 displays the fitted restricted cubic spline function with 5 knots (at the percentiles listed above and depicted with vertical lines) along with estimates of the logit {TVDLM = 1 | symptom duration} obtained by dividing the sample into 15 groups of equal numbers of patients and computing the logit of the proportion of TVDLM = 1 within each group. The estimated spline function suggests the following logistic model:

$$\text{logit}\{\text{TVDLM} = 1 \mid X\} = B_0 + B_1 \log(X),$$

which when fitted provided a near-perfect fit to the data.

The next two examples demonstrate restricted spline Cox survival models. Data were simulated from known population distributions so that spline estimates can be compared with true population estimates. First, a uniformly distributed random number X between 0 and 1 was generated for each of 3,000 hypothetical subjects. A failure time T was generated for each subject, from an exponential distribution with constant hazard function:

$$h(t \mid X) = h(X) = .02 \exp(3 \mid X - .5 \mid).$$

Followup time was censored according to a uniform distribution, resulting in 783 failures out of 3,000 subjects.

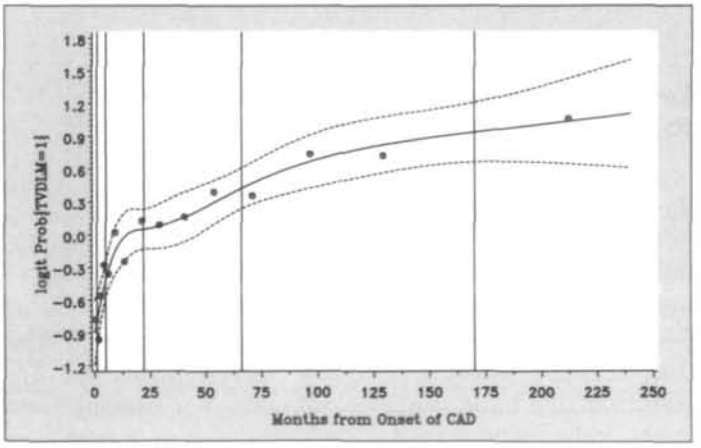

Figure 2. Fitted spline logistic regression model and 95% confidence limits. Predictor variable is duration of symptoms of coronary artery disease. Circles represent logit of proportions of patients with severe (left, main, or three-vessel) coronary disease divided into 15 groups with equal No. of patients.

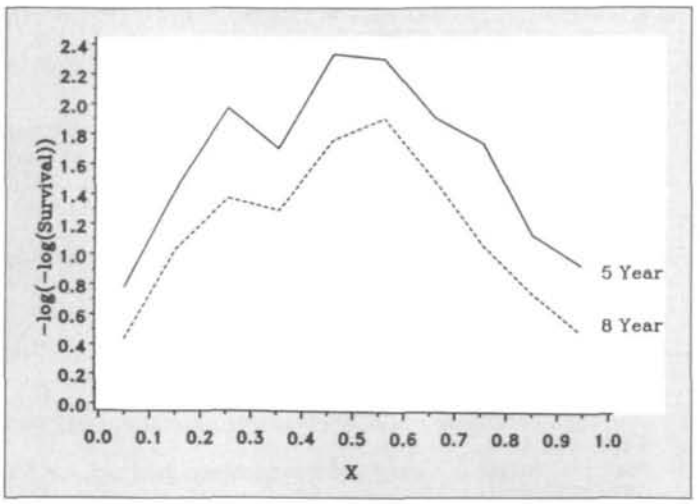

Figure 3. Transformed Kaplan-Meier 5- and 8-year survival estimates by deciles of X. Data were simulated from an exponential survival distribution with true (population) hazard function given by $h(x) = .02 \exp(3|x-.5|)$.

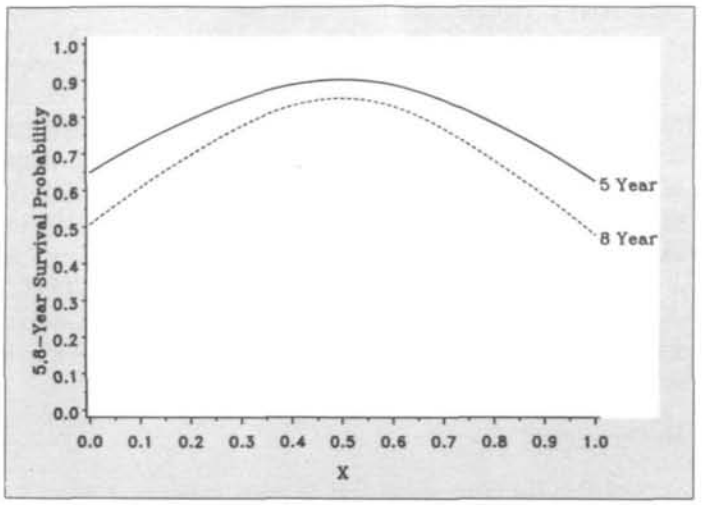

Figure 5. Spline Cox model estimates of 5- and 8-year survival as a function of X. Derived from estimates in figure 4.

The log{−log} transformation of the true population survival function is:

$$\log\{-\log[S(t \mid X)]\} = \log(.02\ t) + 3\ |X-.5|\ .$$

If a Cox model with linearity in X were fitted to these data, the resulting regression equation would be flat instead of the correct V-shaped function of X.

Figure 3 displays −log{−log} Kaplan-Meier (*7*) 5- and 8-year survival estimates obtained by stratifying the sample by deciles of X, i.e., into 10 groups each containing 300 subjects. Although the Kaplan-Meier estimates do not assume a shape for the regression, the estimates are "noisy" because subgrouping reduces the sample size. Figure 4 displays a restricted cubic spline fit to this relationship using five knots placed at percentiles listed above (here at X = .05, .25, .49, .75, and .95). Note the more precise fit, as compared with stratified Kaplan-Meier estimates, in this highly nonlinear example. Figure 5 shows the Cox estimated 5- and 8-year survival probability (*6*) as a function of X based on this fitted spline function.

In the final example, data were again simulated for 3,000 hypothetical subjects but with true population hazard and survival functions given by:

$$h(t \mid X) = h(X) = .02 \exp[\ .7 \log(X)\ ]$$

$$\log\{-\log[S(t \mid X)]\} = \log(.02t) + .7 \log(X)\ .$$

The five-knot restricted cubic spline fit is depicted in figure 6 for t = 8 years. Note the excellent agreement with the true logarithmic effect of the predictor.

Computer Programs

Many computer programs are available in the IBM mainframe version of the SAS (*18*) for performing the kinds of analyses discussed in this paper. *See* references *19, 10, 8,* and *4* for more information.

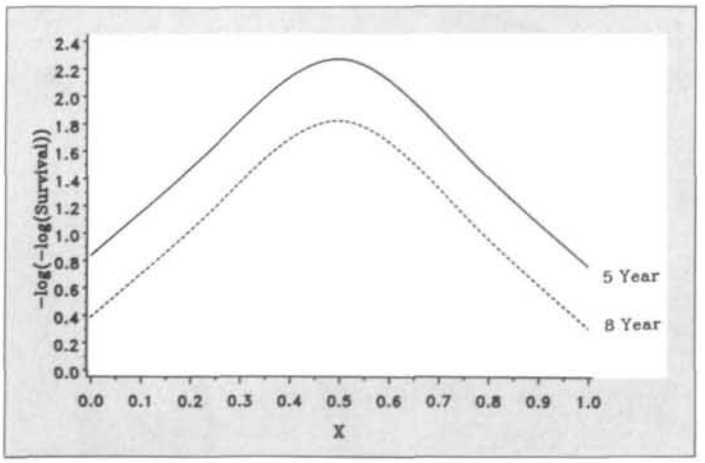

Figure 4. Restricted cubic spline estimate of the effect of X on −log{−log survival} corresponding to empirical estimates in figure 3.

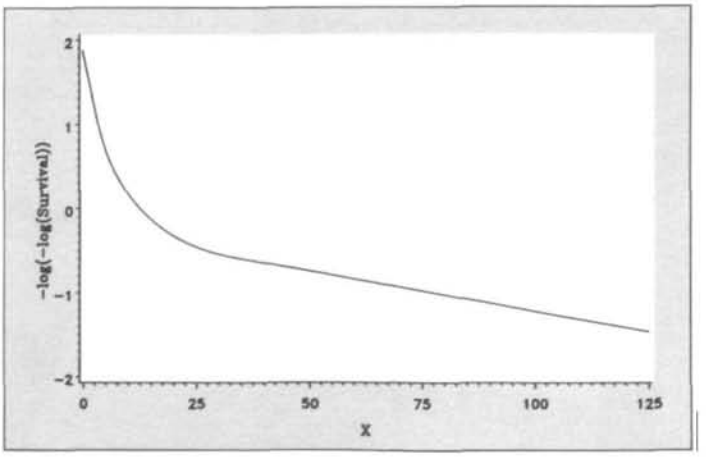

Figure 6. Restricted cubic spline estimate of the effect of X on transformed survival. Data were simulated from an exponential survival distribution with true hazard function given by $h(x) = .02 \exp(.7 \log(x))$.

Summary

Multiple regression models, if used properly, are powerful tools in the analysis of data from clinical studies. Of the four types of regression model assumptions, this paper has dealt with one, namely the assumption regarding the shape of the function relating a continuous predictor variable to the response variable. Restricted cubic spline functions are useful tools for correctly modeling and extracting available information from a continuous predictor. More accurate prediction, better control of confounding, and more powerful statistical tests will result from the application of methods such as these.

References

1. Cox DR. The regression analysis of binary sequences. J R Stat Soc 1958;B20:215-242.
2. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika 1967;54:167-179.
3. Day NE, Byar DP. Testing hypotheses in case-control studies—equivalence of Mantel-Haenszel statistics and logit score tests. Biometrics 1979;35:623-630.
4. Harrell FE, Lee KL. The practical value of logistic regression. In: Proceedings of the 10th annual SAS Users Group International Conference. Cary, NC: SAS Institute, Inc., 1985:1031-1036.
5. Cox DR. Regression models and life tables (with discussion). J R Stat Soc 1972;B34:187-220.
6. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. New York: Wiley, 1980.
7. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457-481.
8. Harrell FE, Lee KL. Verifying assumptions of the Cox proportional hazards model. In: Proceedings of the 11th annual SAS Users Group International Conference. Cary, NC: SAS Institute, Inc., 1986:823-828.
9. Breslow NE, Day NE, Halvorsen KT, et al. Estimation of multiple relative risk functions in matched case-control studies. Am J Epidemiol 1978;108:299-307.
10. Devlin TF, Weeks BJ. Spline functions for logistic regression modeling. In: Proceedings of the 11th annual SAS Users Group International Conference. Cary, NC: SAS Institute, Inc., 1986:646-651.
11. Copas JB. Plotting p against x. Appl Stat 1983;32:25-31.
12. Hastie T, Tibshirani R. Generalized additive models (with discussion). Stat Sci 1986;1:297-318.
13. Landwehr JM, Pregibon D, Shoemaker AC. Graphical methods for assessing logistic regression models (with discussion). J Am Stat Assoc 1984;79:61-83.
14. Smith PL. Splines as a useful and convenient statistical tool. Am Stat 1979;33:57-62.
15. de Boor C. A Practical Guide to Splines. New York: Springer-Verlag, 1978.
16. Stone CJ, Koo CY. Additive splines in statistics. Proc Statist Computing Sect ASA, 1985:45-48.
17. Stone CJ. Comment: Generalized additive models [*see* ref. *12*]. Stat Sci 1986;1:312-314.
18. SAS Institute, Inc. SAS User's Guide: Basics, Version 5 Edition. Cary NC: SAS Institute, Inc., 1985.
19. Harrell FE, Pollock BG, Lee KL. Graphical methods for the analysis of survival data. In: Proceedings of the 12th annual SAS Users Group International Conference. Cary, NC: SAS Institute, Inc., 1987:1107-1115.