An Evaluation of Splines in Linear Regression

Deborah Hurley, MSPH, University of South Carolina, Columbia, SC
James Hussey, PhD, University of South Carolina, Columbia, SC
Robert McKeown, PhD, University of South Carolina, Columbia, SC
Cheryl Addy, PhD, University of South Carolina, Columbia, SC

ABSTRACT

Linear regression is an analytic approach commonly used in public health to examine the relationship between continuous dependent (e.g., blood pressure) and independent (e.g., body mass index (BMI)) variables. For any regression procedure it is desirable to use models that closely fit the data. Transformations of the response variable can improve the fit and may correct violations of model assumptions such as constant error variance. Predictor variables may be separated into logical categories (e.g., age categories), or we may add additional terms that are functions of the existing predictors such as quadratic or cubic terms. Other approaches, such as *spline* modeling, may provide a better fit, taking into consideration the variation in the relationship between the predictor variable and the response variable, both within and between levels of the predictor variable. There is no one best approach, however, as some modeling methods may produce better results for predicted values (e.g., smaller confidence intervals) than other methods, depending on the data. Analyses using splines is often cumbersome and interpretations are often complex. Given these challenges, this project undertakes a simulation study to examine and compare several "traditional" models with spline models, under varying conditions (e.g., different sample sizes and magnitude of variation), for different data structures (e.g., true quadratic, cubic, or other data patterns), in an effort to determine if spline regression models provide a significantly better fit under these conditions than a regression model employing a simple linear relationship, or one with power term(s). Each scenario was assessed for model "fit" verses model simplicity (i.e., to see whether or not the more complex spline regression models provide any real advantage over what can be obtained with SLR or power models). In general, the best model has the same structure as the data. For these data, the choice of the best modeling method should take into account preliminary plots and the estimated standard deviation. Results show that splines are most appropriate when the plots of the data clearly indicate that they are needed (i.e., when the standard deviation is small enough that we can detect knots and changes in structure). When the plots do not show much detail (i.e., when the standard deviation is large), a simpler model (e.g., polynomial) is recommended. Results also reinforce the need to look at a plot of the predicted values for the model, as some of the usual selection criteria (MSE, PRESS, R^2) can give similar results for various models, but the coverage for these models may be diverse.

INTRODUCTION/BACKGROUND

Linear regression is an analytic approach commonly used in public health when we would like to examine the relationship between numeric dependent (e.g., blood pressure) and independent (e.g., body mass index) variables. For any regression procedure, it is desirable to use models that closely fit the data. Transformations of the response variable can improve the fit and may correct violations of model assumptions such as constant error variance. We may also consider separating a predictor variable into logical categories (e.g., age categories), or adding additional terms that are functions of the existing predictors such as quadratic or cubic terms. Still other methods, such as *spline* modeling, may provide a better fit, taking into consideration the variation in the relationship between the predictor variable and the response variable, both within and between levels of the predictor variable. There is no one best approach, however, as some modeling methods may produce better results for predicted values (e.g., narrower confidence intervals) than other methods, depending on the data. Greenland (1995) suggests using spline regression (and fractional polynomial regression) as an alternative to categorical analysis for dose response and trend analysis, stating that categorical analysis does not make use of within category information and is based on an unrealistic model for dose-response and trends. Spline regression, he contends, is based on more realistic category-specific models that are especially worthwhile when nonlinearities are expected.

Splines are lines or curves, which are usually required to be continuous and smooth. *Univariate polynomial splines* are piecewise polynomials in one variable of some degree *d* with function values and first *d*-1 derivatives that agree at the points where they join. The *join points* (or abscissa values) that mark one transition to the next are referred to as *break points*, *interior knots,* or simply *knots* (Poirier, 1976, Eubank, 1999). Knots give the curve freedom to bend and more closely follow the data. Splines with few knots are generally smoother than splines with many knots; however, increasing the number of knots usually increases the fit of the spline function to the data. (Hansen and Kooperberg, 2002).

Spline functions can be applied to medical and epidemiological investigations. These studies frequently involve survival analysis, linear dose-response problems, latency patterns, and data smoothing (to detect trends) as well as other studies. For example, to assess mortality in colon cancer using survival analysis methods, Bolard et al (2002) used restricted cubic splines to model time-by-covariate interactions. In another study to look at linear dose-response, Thurston et al (2002) applied penalized spline methodology to a cohort study of autoworkers exposed to metalworking fluids to examine the linearity assumption for prostate and brain cancer mortality. Hauptmann et al (2001) used spline function models to investigate latency patterns for radon progeny exposure and lung cancer in a cohort of uranium miners in Colorado. In a study estimating longitudinal immunological and virological markers in HIV patients with individual antiretroviral treatment strategies, Brown et al (2001) proposed univariate and bivariate cubic smoothing splines to fit CD4+ count and plasma viral load.

There are many types of splines and estimation procedures (Gu, 2002, Eubank, 1999). The analyses presented in this paper focus on univariate splines in ordinary least squares regression. Knot selection (number and location of knots) can be accomplished by various methods. One can use predetermined knots, natural division points, or visually inspect the data. There are also other (more complex) methods, such as nonlinear least squares methods, for knot selection (Eubank, 1999). Predetermined knots are used in this paper.

Because analyses using splines is often cumbersome and interpretations complex, it is necessary to compare the tradeoff between model complexity and model fit in order to assess whether a much more complex model provides a significantly better fit. This project uses five criteria for the comparison of models. The first two are based on confidence intervals (CI) for the mean. These were examined to determine the proportion of times the true mean of the distribution was contained in the CI (i.e., coverage). The widths of these CI were also compared between models. Coverage proportions were cross-referenced with confidence interval widths to assess any relationship (e.g., wider confidence intervals associated with greater coverage proportions, or deviation from what would be expected).

The other three criteria are mean square error (MSE), the PRESS statistic (prediction sum of squares), and the R-square statistic (R^2). Under the assumption that the correct model should give an unbiased estimate of the variance, values for MSE were compared to the variance used to generate the data to see if the variance was being over or under-estimated. The PRESS statistic can be a good indication of the predictive power of a model. The PRESS residual is the difference between the observed value and the predicted value, when the model was fit without that point. The PRESS statistic is the sum of the squared PRESS residuals. Since this is a sum of squared errors, a good model has a small PRESS statistic. Finally, recall that R^2 is the proportion of variation in the dependent variable explained by the model fit using the independent variable(s); we can examine this proportion to get an idea of how well the predicted equation fits the data. Together, these measures were used to assess whether or not the more complex spline regression models provide any real advantage over what can be obtained with SLR or power models.

METHODS

To address the tradeoff between model complexity and model fit, we conducted a simulation study to compare "traditional" regression models with spline models under varying conditions (e.g., different sample sizes and magnitude of variation), for different data structures (e.g., true quadratic, cubic, or other data patterns). The goal was to determine if the added complexity of the spline regression models is justified by a significantly better fit (under certain conditions) than a regression model employing a simple linear relationship, or one with power term(s).

Data were simulated for five different structures (patterns), using one dependent variable and one independent variable. Each of these five structures was generated with three different sample sizes (n), and two different standard deviations, for a total of 6 scenarios for each structure. For each scenario, 2000 simulated data sets were generated. Six different regression models were evaluated for each of the 30 scenarios. These models were simple linear regression (SLR), polynomial regression (quadratic and cubic), and spline regression (linear, quadratic and cubic). All data generation and analyses were completed using SAS version 8.2e.

DATA STRUCTURES

The data from these data sets were designed to follow somewhat realistic patterns and to encompass the various scenarios and restrictions. Data for the first three structures were generated to obtain a general quadratic-type pattern, while data for the last two structures were generated to obtain a general cubic-type pattern. Because the support for independent variable (x) is uniformly distributed from 0 to 100 ($0 \le x \le 100$), the different sample sizes ($n = 51, 201, 1001$) also correspond to the density of the points in terms of the independent variable.

For each of the five data structures, two knots were chosen to allow the function to vary on up to three segments per data structure. The placement of the knots was at $x = 32$ and $x = 68$, which creates nearly equal intervals between 0 and 100. This was done in order to keep the number of observations within each segment approximately the same. The values for determining mean y were calculated using the quadratic, cubic, or piecewise equations necessary to produce the shape of the segment for the data structure of interest. Normally distributed random error was added to

each mean value to produce the data point for the dependent variable, using: $y_i = f(x_i) + \varepsilon_i$, where y_i is the dependent variable, $f(x_i)$ is the function used to generate the mean value of y, x_i is a value between 0 and 100 (generated in increments of 100/(n-1), depending on the sample size, n), and ε_i is the random error produced by the standard deviation (SD) multiplied by a random standard normal value. Different values of the SD correspond to the variation of the points (y) from the true, underlying relationship. Visual inspection of randomly selected plots of the data was used to determine the desired amount of variation. The larger of the two values chosen for the SD was 0.1. This value produced enough variability that the original pattern was discernable only at close inspection. The smaller value chosen for the SD was 0.03. This value produced data with some variability but where the original pattern was clearly evident when plotted.

Structures one through three are similar in that all take on a quadratic or quadratic-like form (see Figure 1). Structure one consists of three linear segments. Structure two consists of a middle linear segment sandwiched between two quadratic segments. Both structures are decreasing, constant, and increasing on the same intervals, with changes in structure occurring at the predetermined knots. Structure three is a purely quadratic structure over the same sample space. All structures pass through y= 0.9 at x=0, have a minimum y=0.1, and pass though y=0.5 at x=100.

Structures four and five have a general cubic form (see Figure 1). Structure four consists of three linear segments (increasing, constant, then increasing), with changes in structure occurring at the predetermined knots. Structure five is purely cubic (increasing, decreasing, then increasing) and has a local maximum at x = 32, and a local minimum at x = 68 (the knots). Both structures pass through y= 0.1 at x=0, and y=0.9 at x=100.

REGRESSION MODELS

The six regression models were analyzed for each simulated data set, using the PROC REG procedure in SAS. The SLR and power models were of the usual form (e.g., cubic model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$), including the highest order term and all lower order terms. The three spline models were a linear spline, a quadratic spline, and a cubic spline. For example, the cubic spline model is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$, where x_4 and x_5 are used to model a piecewise cubic independent variable. Here, x_4 represents the change in x^3 when x is greater than 32, and is 0 otherwise, and x_5 represents the change in x^3 when x is greater than 68, and is 0 otherwise. Thus,

ANALYSES

Predicted means and their corresponding 95 percent CIs were calculated at seven "test" points of interest (x = 0, 16, 32, 50, 68, 84, 100). These points include the knots (x = 32 and 68), the two endpoints (x = 0 and 100), the midpoint (x = 50), and two points midway between each endpoint and the knot closest to it (x = 16 and 84). The observed proportion, $\hat{p}$, of the 2000 replications in which the true mean values were contained in the corresponding CI was recorded, and a Wald confidence interval was then calculated for this proportion, using the formula:

$$\hat{p} \pm (1.96) \sqrt{\hat{p}(1-\hat{p})/2000}$$

An interval that does not contain .95 indicates that the actual coverage is not the nominal 95%. Values of $\hat{p}$ in the range (0.9396, 0.9587) produce CIs that include 0.95. To get an idea of overall coverage, and for comparative purposes, this interval was also calculated for the set of *all* points in each data set, not just the set of individual "test" points. (Note that the derivation for the CI is for 2000 repetitions on a single point and is not relevant for "*all x*".)
Additionally, means and SDs of the widths of the CIs were calculated.
Finally, univariate analyses were done for the mean square error (MSE), the PRESS statistic and R^2 statistic. Results for MSE were compared by model within each of the 6 scenarios to see if the empirical values were similar to the true variance. The results for the PRESS statistic were compared by model within each scenario (smaller is better), as were the R^2 statistics (larger is better). Results for all three statistics were also graphed as histograms. The histograms were visually inspected for anything that might be interesting or unusual.

RESULTS

COVERAGE

For each test point, the proportion of coverage is obtained for the various sample size, variance, and model combinations. Proportions underlined and in **bold** in Table 1 have CIs that include the nominal value of 0.95. Results presented are only for the moderate sample size (n = 201) and the larger standard deviation ($\sigma = 0.1$), as they are representative of overall results. In general, the SLR model does not provide the nominal 95% coverage for majority of the points under the 6 scenarios. This is not surprising, and underscores the inadequacy of assuming a linear relationship when the true relationship is more complex. For the piecewise structures (1, 2, and 4), we would expect the "correct" model (the linear spline, quadratic spline, and linear spline models, respectively) to perform better than the polynomial models, as these structures are all continuous piecewise polynomials with known knot locations.

Results in Table 1 indicate that this is indeed the case. For the piecewise linear "quadratic" structure (1), the linear spline provides a superior fit, as it is the only model to have acceptable coverage at all seven test points. In fact, when looking at all test points for all 6 sample size/variability combinations, in only three cases did the coverage proportion yield a confidence interval that did not contain the nominal 95% value (results not shown). For the piecewise linear and quadratic structure (2), the quadratic spline provides a superior fit for all individual points. Again looking at all 6 sample size/variability combinations with this structure, there were only two proportions that led to confidence intervals that did not contain 95% (results not shown). For the piecewise linear "cubic" structure (4), the linear spline provides a superior fit for all individual points, as was the case for the first structure. Here, only two proportions led to confidence intervals that did not contain 95% (results not shown).

For the purely polynomial structures (3 and 5), we would expect the "correct" model (quadratic and cubic polynomial models, respectively) to perform best. However, we would also expect the corresponding spline models (quadratic and cubic, respectively) to also perform well, since the polynomial models are special cases of the spline models. Again, results in Table 1, show this is the case. For the quadratic structure (3), the quadratic, cubic, quadratic spline, and cubic spline models all fit well. For the cubic structure (5), the cubic, quadratic spline, and cubic spline models all fit well.

CONFIDENCE INTERVAL WIDTH

To see how the modeling method affects the width of the confidence interval, the mean and standard deviation of the widths (for the 2000 simulations) at the various test points are compared among the models for each structure (see Tables 2-6). Results are again presented only for the moderate sample size (n = 201), as they are representative of all sample sizes. Results are presented for both standard deviations ($\sigma = 0.1$ and $\sigma = 0.03$), however, as these results did differ by the amount of variability in the data.

For the three quadratic-type structures (1 – 3), the SLR model provides a very poor fit, as it has the widest average interval widths at all but a few test points over the 6 scenarios within each structure. Results are shown in Tables 2 - 4 for n=201 only. This is as one might expect, since we are modeling a quadratic relationship with a straight line. The results for the other five models are similar for all three structures when the standard deviation is relatively large ($\sigma = 0.1$). The quadratic polynomial model generally produces the narrowest intervals, and the cubic spline model produces the widest. For the piecewise linear "quadratic" structure (1), the cubic spline intervals range from 15 to 78 percent wider than those produced by the quadratic model. For the piecewise quadratic-linear-quadratic structure (2), the cubic spline intervals are 5 to 62 percent wider, and for the purely quadratic structure (3), the cubic spline intervals are 20 to 86 percent wider.

When the standard deviation is relatively small ($\sigma = 0.03$), the results differ for each structure. For structure 1, the linear spline model generally has the narrowest intervals, while the cubic spline model has the widest (see Table 2). The intervals from the cubic spline are 13 to 114 percent wider than those produced by the linear spline. For structure 2, the quadratic spline model generally has the narrowest intervals, and the linear spline model has the widest (see Table 3). The intervals from the linear spline are 41 to 167 percent wider than those from the quadratic spline. For structure 3, the results are very similar to those when $\sigma = 0.1$, with the quadratic polynomial model yielding the narrowest intervals, and the cubic spline model having the largest. The intervals from the cubic spline are 20 to 86 percent wider than those from the quadratic model.

The results for the two cubic-type structures (4 and 5, piecewise linear "cubic" and purely cubic, respectively), differ from one another. In general, for structure 4 and $\sigma = 0.1$, the SLR model produces the narrowest intervals, while the cubic spline model produces the widest, having intervals 36 to 136 percent wider than those for the SLR model (see Table 5). In fact, each spline model generally yields wider intervals than its corresponding polynomial model. When the standard deviation was small ($\sigma = 0.03$), the linear spline produced the narrowest intervals for this structure, and the quadratic polynomial produced the widest, having intervals 32 to 154 percent wider than the linear spline model. In contrast, the SLR and quadratic polynomial models generally yield larger intervals than the corresponding spline model, while the cubic polynomial and spline models yield similar intervals.

Table 6 shows the results for the purely cubic structure (5). In general, for structure 5 and $\sigma = 0.1$, the cubic model generally yields the narrowest intervals for both standard deviations, while the quadratic polynomial and cubic spline models yield the widest (20 to 44 percent wider). When the standard deviation was small ($\sigma = 0.03$), the quadratic polynomial model yields the widest intervals (up to 250 percent wider). However, the intervals from the cubic spline model remain 20 to 44 percent wider than those produced by the cubic model, regardless the standard deviation.

MSE, PRESS STATISTIC, and R^2 STATISTIC

Results for the model selection criteria (MSE, PRESS, and R^2) are not shown. In general, results for all of these statistics show that the model that is considered to be the "correct" model is in fact the best model. That is, for these models the empirical mean and standard deviation for the given statistic is best estimate compared to the other models. The "best" estimates would be a MSE that is closest to σ^2, the PRESS statistic that is the smallest, and an R^2

value that is the largest. Note that in some instance, there are one or more other models that perform just as well, or nearly as well. For example, the mean MSE for structure 3 (purely quadratic) is identical for the quadratic and cubic polynomial models, and the quadratic and cubic spline models. For structure 5 (purely cubic), the both the cubic polynomial and cubic spline models results for the MSE, PRESS statistic, and the R^2 are nearly identical.

CONCLUSIONS/DISCUSSION

Recall that the data structures were designed so that certain models would be "correct". In general, based on all criteria, the "correct" model for each structure is in fact the "best" overall model, as expected. Of particular interest are the simpler models that perform as well as the "correct" model, and the more complex models that perform unexpectedly poorly.

Reviewing each structure independently, it is possible to draw some general conclusions about the quadratic-type structures and the cubic-type structures, and to determine when to use the more complicated spline models. If the data between all knot locations and endpoints (all partitions) has a straight-line linear data structure, are continuous and have different slopes (such as structures 1 and 4), then the linear spline (the "correct" model) is the most appropriate model for prediction. If there is curvature on any of the partitions, then a spline that is representative of the overall structure (e.g., quadratic, as in data structure 2) is the best model. Finally, when a data structure that follows a more purely polynomial shape (e.g., strictly linear, or purely quadratic or cubic on all partitions), then a polynomial that is representative of the overall structure (e.g., quadratic or cubic, as in data structures 3 and 5, respectively) is the best model. For the purely polynomial structures, both the polynomial model and the spline model are "correct" models for this type of structure. However, the simpler polynomial models provide more precise estimates and do not require knowledge of knot location and variation in the data.

Results indicate that, in general, the sample size is not particularly important in determining which modeling method works best. All models generally did better with an increase in sample size, as we might expect, but the ordering of models from best to worst remained constant as the sample size changed.

There is a greater disparity in model performance, in general, when the standard deviation was relatively small. For example, when $\sigma = 0.1$, the confidence intervals were nearly twice as wide for the "worst" model as compared to the "best" model. When $\sigma = 0.03$, however, the ratios of the widths for the "worst" and "best" models were between two and five. Similar relationships exist when looking at MSE and the PRESS statistic. Fortunately, less variability improves the ability to discriminate between models, a key element in selecting the correct model. Thus, preliminary plots should guide the choice of a modeling method. These results show that splines are most appropriate when variability in the data is small enough that plots of the data reveal knots and changes in structure). This is especially true of the linear spline, which is clearly most suited for piecewise linear data structures. When the plots do not show much detail (i.e., when the variability is large enough to hide any clear changes in structure), it is better to use a simpler model (e.g., polynomial), as the simpler models perform as well as or even better than the more complex models in some situations. These results also reinforce the need to look at a plot of the predicted values and residuals for your model, as some of the usual selection criteria (MSE, PRESS, R^2) can give similar results for various models, but the coverage for these models may be very different (e.g., see results for Structure 4, the piecewise linear "cubic" structure). Conversely, coverage can be similar for different models, but the selection criteria results may vary (e.g., see results for Structure 3, the purely quadratic structure).
Overall results are generally consistent with the literature. Greenland (1995) suggests spline regression is based on more realistic category-specific models that are especially worthwhile when structure changes are expected. However, our data suggest that even when structure changes are expected, a simpler polynomial model may perform as well as or better than a spline model when knots and structure definition are not clear due to variability in the data.

Therefore, complex spline models, though useful for some defined data structures, are not appropriate for all situations. While they perform well when used to smooth a series of data, they may not perform well if used in a regression setting where prediction is of interest. This agrees with the well-known phenomenon of decreased precision when adding any unnecessary term to a regression model. Thus, users should exercise caution in the application of such models, and not assume that they will be appropriate in any modeling situation.

Our study was limited, in that our structures were the result of simulation and not based on real epidemiological or other data where we might want to apply these techniques. It is not difficult, however, to conceptualize these structures as representative of changes in data structure across age group, dose, or time, for example. We also were limited in that we only examined one regression type (linear) with only one independent variable. However, we believe that the general results found here would obtain in more complicated models.

Future research should include a more in-depth look at the effect of the standard deviation on the these modeling methods, the effect of more complicated data structures, and evaluation using multivariable models, other types of regression (e.g., logistic), and the use of real data.

Table 1: Proportion of Coverage for True Mean: N= 201, σ= 0.1(All Structures)

Regression Model:		Polynomial			Spline		
		SLR	Quadratic	Cubic	Linear	Quadratic	Cubic
Structure	x						
Piecewise	0	0.0000	0.9370	0.4586	**0.9425**	0.6110	**0.9465**
Linear	16	0.0000	0.6530	0.1600	**0.9505**	0.3080	0.8770
"Quadratic"	32	0.0000	0.0000	0.0000	**0.9530**	0.0000	0.0065
(1)	50	0.0000	0.2380	0.2120	**0.9450**	0.4385	0.2120
	68	0.0000	0.9595	0.2585	**0.9410**	0.2970	0.1445
	84	0.0000	0.4430	0.9295	**0.9515**	**0.9420**	0.9035
	100	0.0000	0.0265	0.9355	**0.9520**	**0.9465**	0.9185
	all x	*0.0968*	*0.4421*	*0.6374*	*0.9477*	*0.6815*	*0.7311*
Piecewise	0	0.0000	0.0000	0.1025	0.0000	**0.9470**	0.8545
Linear &	16	0.0000	0.0000	0.0235	0.0000	**0.9460**	0.9365
Quadratic	32	0.0000	0.0010	0.8650	0.1945	**0.9490**	0.8360
(2)	50	0.0000	0.0010	0.0010	0.1695	**0.9485**	0.7850
	68	0.0000	0.3590	0.8200	0.9360	0.9375	0.8575
	84	0.5065	0.4535	0.0240	0.0245	**0.9535**	0.9330
	100	0.0000	0.9690	0.0800	0.0340	**0.9515**	0.9115
	all x	*0.1074*	*0.3492*	*0.3189*	*0.3465*	*0.9477*	*0.8664*
Quadratic	0	0.0000	**0.9480**	**0.9445**	0.6570	**0.9510**	**0.9555**
(3)	16	0.0020	**0.9495**	**0.9450**	0.6585	**0.9525**	**0.9540**
	32	0.0000	**0.9530**	**0.9485**	0.2160	**0.9470**	**0.9565**
	50	0.0000	**0.9475**	**0.9480**	0.1835	**0.9465**	**0.9490**
	68	0.0000	**0.9455**	**0.9415**	0.1910	**0.9410**	**0.9410**
	84	0.0025	**0.9485**	**0.9415**	0.6660	**0.9460**	**0.9530**
	100	0.0000	0.9395	**0.9505**	0.6945	**0.9500**	**0.9470**
	all x	*0.1174*	*0.9485*	*0.9469*	*0.7114*	*0.9467*	*0.9489*
Piecewise	0	0.0000	0.0085	0.6265	**0.9540**	0.7420	**0.9465**
Linear	16	0.9600	0.9675	0.3440	**0.9570**	0.4790	**0.9470**
"Cubic"	32	0.0000	0.0000	0.0500	**0.9530**	0.0350	0.7205
(4)	50	0.9820	0.9755	0.9595	**0.9515**	**0.9565**	**0.9470**
	68	0.0000	0.0000	0.0480	**0.9515**	0.0305	0.7290
	84	**0.9570**	0.9600	0.3345	**0.9545**	0.4645	**0.9425**
	100	0.0000	0.0065	0.5890	**0.9560**	0.7420	**0.9465**
	all x	*0.2133*	*0.2604*	*0.6417*	*0.9533*	*0.6386*	*0.9313*
Cubic	0	0.0000	0.0000	**0.9500**	0.0035	**0.9475**	**0.9500**
(5)	16	0.0015	0.0030	**0.9505**	0.0040	**0.9465**	**0.9515**
	32	0.0000	0.0000	**0.9470**	0.1285	**0.9425**	**0.9545**
	50	0.9980	0.9975	**0.9580**	0.9730	**0.9545**	**0.9570**
	68	0.0000	0.0000	**0.9575**	0.1255	**0.9460**	**0.9545**
	84	0.0005	0.0005	**0.9495**	0.0015	**0.9450**	**0.9465**
	100	0.0000	0.0000	**0.9425**	0.0020	**0.9445**	**0.9545**
	all x	*0.1336*	*0.1697*	*0.9516*	*0.4948*	*0.9182*	*0.9513*

Proportions in **bold** are those where the confidence intervals appear to achieve nominal coverage.

Table 2: Confidence Interval Widths: Mean and (Standard Deviation)
For Piecewise Linear "Quadratic" Structure (1), N = 201

Regression Model:		Polynomial			Spline		
		SLR	Quadratic	Cubic	Linear	Quadratic	Cubic
Sigma	x*						
0.1	0/100*	0.12644 (0.00375)	0.08859 (0.00449)	0.11295 (0.00574)	0.09022 (0.00461)	0.12817 (0.00653)	0.15753 (0.00803)
	16/84*	0.09776 (0.00290)	0.04760 (0.00241)	0.04943 (0.00251)	0.04538 (0.00232)	0.05951 (0.00303)	0.07088 (0.00361)
	32/68*	0.07468 (0.00221)	0.04065 (0.00206)	0.05071 (0.00258)	0.06524 (0.00333)	0.05100 (0.00260)	0.06022 (0.00307)
	50	0.06346 (0.00188)	0.04474 (0.00227)	0.04315 (0.00219)	0.03863 (0.00197)	0.05721 (0.00292)	0.05150 (0.00263)
0.03	0/100*	0.11473 (0.00116)	0.04043 (0.00158)	0.04348 (0.00199)	0.02706 (0.00137)	0.04971 (0.00226)	0.05786 (0.00274)
	16/84*	0.08870 (0.00089)	0.02173 (0.00085)	0.01903 (0.00087)	0.01361 (0.00069)	0.02308 (0.00105)	0.02604 (0.00123)
	32/68*	0.06776 (0.00068)	0.01855 (0.00073)	0.01952 (0.00089)	0.01957 (0.00099)	0.01978 (0.00090)	0.02212 (0.00105)
	50	0.05758 (0.00058)	0.02042 (0.00080)	0.01661 (0.00076)	0.01159 (0.00059)	0.02219 (0.00101)	0.01892 (0.00090)

*Results for "grouped" test points were the same due to the layout of the data.

Table 3: Confidence Interval Widths: Mean and (Standard Deviation)
For Piecewise Linear & Quadratic Structure (2), N = 201

Regression Model:		Polynomial			Spline		
		SLR	Quadratic	Cubic	Linear	Quadratic	Cubic
Sigma	x*						
0.1	0/100*	0.11390 (0.00379)	0.09554 (0.00459)	0.11990 (0.00589)	0.10174 (0.00492)	0.12356 (0.00608)	0.15458 (0.00763)
	16/84*	0.08807 (0.00293)	0.05134 (0.00246)	0.05247 (0.00258)	0.05118 (0.00247)	0.05736 (0.00282)	0.06955 (0.00343)
	32/68*	0.06727 (0.00224)	0.04384 (0.00210)	0.05382 (0.00264)	0.07357 (0.00356)	0.04916 (0.00242)	0.05909 (0.00292)
	50	0.05716 (0.00190)	0.04825 (0.00232)	0.04580 (0.00225)	0.04356 (0.00211)	0.05515 (0.00271)	0.05054 (0.00249)
0.03	0/100*	0.10097 (0.00117)	0.05414 (0.00163)	0.05968 (0.00214)	0.05447 (0.00182)	0.03708 (0.00189)	0.05044 (0.00255)
	16/84*	0.07807 (0.00090)	0.02909 (0.00088)	0.02612 (0.00094)	0.02740 (0.00091)	0.01722 (0.00088)	0.02269 (0.00115)
	32/68*	0.05963 (0.00069)	0.02484 (0.00075)	0.02679 (0.00096)	0.03939 (0.00132)	0.01475 (0.00075)	0.01928 (0.00098)
	50	0.05067 (0.00059)	0.02734 (0.00083)	0.02280 (0.00082)	0.02332 (0.00078)	0.01655 (0.00084)	0.01649 (0.00083)

*Results for "grouped" test points were the same due to the layout of the data.

Table 4: Confidence Interval Widths: Mean and (Standard Deviation) For Quadratic Structure (3), N = 201

Regression Model:		Polynomial			Spline		
		SLR	Quadratic	Cubic	Linear	Quadratic	Cubic
Sigma	x*						
0.1	0/100*	0.11228 (0.00377)	0.08251 (0.00415)	0.10906 (0.00549)	0.09196 (0.00463)	0.12359 (0.00624)	0.15325 (0.00775)
	16/84*	0.08682 (0.00291)	0.04434 (0.00223)	0.04773 (0.00240)	0.04626 (0.00233)	0.05738 (0.00290)	0.06895 (0.00349)
	32/68*	0.06632 (0.00223)	0.03786 (0.00190)	0.04896 (0.00247)	0.06651 (0.00335)	0.04917 (0.00248)	0.05858 (0.00296)
	50	0.05635 (0.00189)	0.04167 (0.00210)	0.04166 (0.00210)	0.03937 (0.00198)	0.05517 (0.00279)	0.05010 (0.00253)
0.03	0/100*	0.09921 (0.00120)	0.02479 (0.00123)	0.03277 (0.00163)	0.03266 (0.00157)	0.03714 (0.00186)	0.04605 (0.00231)
	16/84*	0.07671 (0.00093)	0.01332 (0.00066)	0.01434 (0.00071)	0.01643 (0.00079)	0.01724 (0.00086)	0.02072 (0.00104)
	32/68*	0.05860 (0.00071)	0.01138 (0.00057)	0.01471 (0.00073)	0.02362 (0.00114)	0.01478 (0.00074)	0.01760 (0.00088)
	50	0.04979 (0.00060)	0.01252 (0.00062)	0.01252 (0.00062)	0.01398 (0.00067)	0.01658 (0.00083)	0.01506 (0.00076)

*Results for “grouped” test points were the same due to the layout of the data.

Table 5: Confidence Interval Widths: Mean and (Standard Deviation) For Piecewise Linear “Cubic” Structure (4), N = 201

Regression Model:		Polynomial			Spline		
		SLR	Quadratic	Cubic	Linear	Quadratic	Cubic
Sigma	x*						
0.1	0/100*	0.06492 (0.00305)	0.09685 (0.00455)	0.11200 (0.00548)	0.09001 (0.00441)	0.12751 (0.00625)	0.15344 (0.00755)
	16/84*	0.05020 (0.00235)	0.05204 (0.00245)	0.04901 (0.00240)	0.04528 (0.00222)	0.05920 (0.00290)	0.06904 (0.00340)
	32/68*	0.03834 (0.00180)	0.04444 (0.00209)	0.05028 (0.00246)	0.06509 (0.00319)	0.05073 (0.00249)	0.05865 (0.00289)
	50	0.03258 (0.00153)	0.04891 (0.00230)	0.04279 (0.00209)	0.03854 (0.00189)	0.05692 (0.00279)	0.05016 (0.00247)
0.03	0/100*	0.03793 (0.00109)	0.05666 (0.00163)	0.04172 (0.00196)	0.02709 (0.00141)	0.04895 (0.00225)	0.04713 (0.00245)
	16/84*	0.02933 (0.00084)	0.03045 (0.00087)	0.01826 (0.00086)	0.01363 (0.00071)	0.02273 (0.00105)	0.02121 (0.00110)
	32/68*	0.02240 (0.00064)	0.02600 (0.00075)	0.01873 (0.00088)	0.01959 (0.00102)	0.01948 (0.00090)	0.01802 (0.00094)
	50	0.01903 (0.00055)	0.02861 (0.00082)	0.01594 (0.00075)	0.01160 (0.00060)	0.02185 (0.00101)	0.01541 (0.00080)

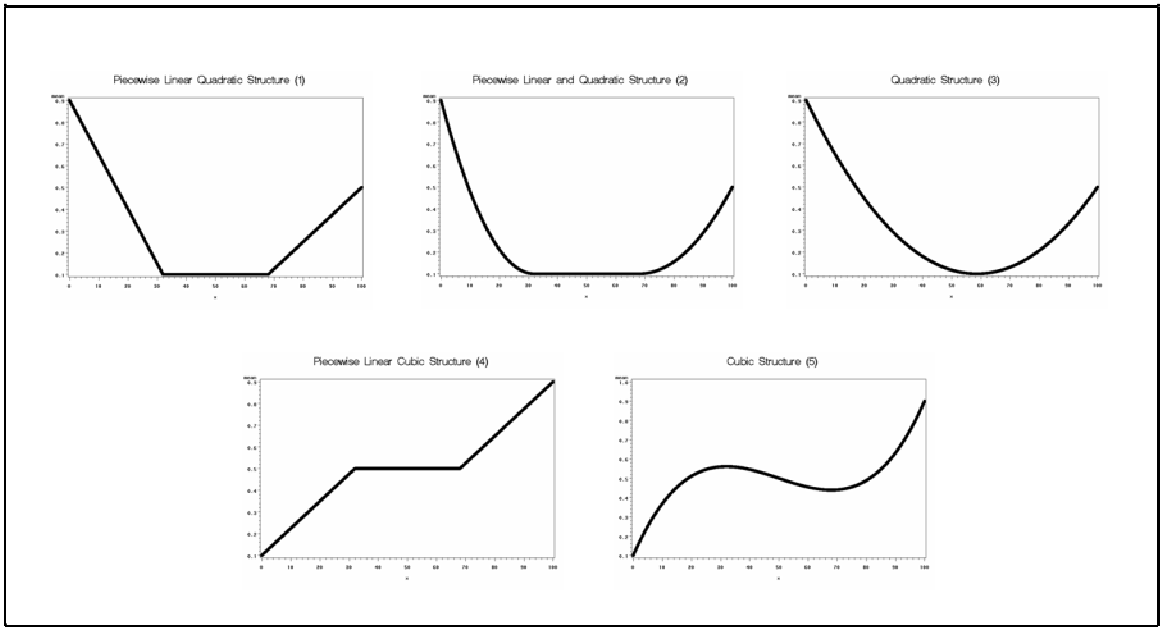
*Results for “grouped” test points were the same due to the layout of the data.

Table 6: Confidence Interval Widths: Mean and (Standard Deviation)
For Cubic Structure (5), N = 201

Regression Model:		Polynomial			Spline		
		SLR	Quadratic	Cubic	Linear	Quadratic	Cubic
Sigma	x*						
0.1	0/100*	0.07875 (0.00341)	0.11755 (0.00509)	0.10907 (0.00562)	0.09667 (0.00491)	0.12394 (0.00642)	0.15324 (0.00796)
	16/84*	0.06089 (0.00264)	0.06317 (0.00274)	0.04773 (0.00246)	0.04863 (0.00247)	0.05754 (0.00298)	0.06895 (0.00358)
	32/68*	0.04651 (0.00201)	0.05394 (0.00234)	0.04896 (0.00252)	0.06991 (0.00355)	0.04931 (0.00255)	0.05858 (0.00304)
	50	0.03952 (0.00171)	0.05936 (0.00257)	0.04167 (0.00215)	0.04139 (0.00210)	0.05533 (0.00286)	0.05010 (0.00260)
0.03	0/100*	0.05830 (0.00117)	0.08711 (0.00175)	0.03272 (0.00164)	0.04420 (0.00172)	0.03814 (0.00192)	0.04597 (0.00232)
	16/84*	0.04507 (0.00091)	0.04681 (0.00094)	0.01432 (0.00072)	0.02223 (0.00087)	0.01771 (0.00089)	0.02068 (0.00104)
	32/68*	0.03443 (0.00069)	0.03997 (0.00080)	0.01469 (0.00074)	0.03196 (0.00124)	0.01518 (0.00077)	0.01757 (0.00089)
	50	0.02926 (0.00059)	0.04399 (0.00088)	0.01250 (0.00063)	0.01892 (0.00074)	0.01703 (0.00086)	0.01503 (0.00076)

*Results for “grouped” test points were the same due to the layout of the data.

Figure 1. Five Data Structures



REFERENCES

Ahlberg, J.H., Nilson, E.N., and Walsh, J.L. (1967). The Theory of Splines and Their Applications. New York: Academic Press. Pp. 1-74, 109-152.

Bolard, P., Quantin, C., Abrahamowicz, M., Esteve, J., Giorgi. R., Chadha-Boreham, H., Binquet, C., Faivre, J. (2002). Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions. *J Cancer Epidemiol Prev* 7(3): 113-122.

Brown, E.R, Ma Whinney, S., Jones, R.H., Kafadar, K., Ypung, B. (2001). Improving the fit of bivariate smoothing splines when estimating longitudinal immunological and virological markers in HIV patients with individual antiretroviral treatment strategies. *Stat Med* 20(16): 2489-2504.

De Boor, Carl (2001). A Practical Guide to Splines. New York: Springer-Verlag New York, Inc. Pp. 17-37, 69-76, 79-86, 207-224.

Dierckx, Paul (1993). Curve and Surface Fitting with Splines. New York: Oxford University Press, Inc. Pp.3-5, 53-67.

Eubank, R. L. (1999). Nonparametric Regression and Spline Smoothing, 2nd Ed. New York: Marcel, Dekker, Inc. Pp. 1-23, 27-37, 119-144, 291-307.

Eubank, R. L. (1988). Spline Smoothing and Nonparametric Regression. New York: Marcel, Dekker, Inc. Pp. 1-40, 89-101, 189-195, 353-363.

Faraway, Julian J. (2002). Practical Regression and Anova using R. Copyright Julian J. Faraway. Pp. 95-105.

Green, P.J. and Silverman, B.W. (2000). Nonparametric Regression and Generalized Linear Models. Boca Raton, Fl.: Chapman & Hall/ CRC (reprint). Pp. 11-23.

Greenland, Sander. (1995). Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology* 6(4): 356-365.

Greville, T.N.E, Jerome, J.W., Loscalzo, Frank R., Schoenberg, I.J., Schumaker, L.L., and Varga, R.S. (1969). Theory and Applications of Spline Functions. New York: Academic Press. Pp. 1-65.

Gu, Chong (2002). Smoothing Spline ANOVA Models. New York: Springer-Verlag New York, Inc. Pp. 2-3, 30-52, 111-142.

Hansen M. and Kooperburg, C. "Spline Adaptation in Extended Linear Models." Statistic Science 17.1 (2002): 1-51.

Hauptmann, M., Berhane, K., Langholz, B., Lubin, J. (2001). Using splines to analyze latency in the Colorado Plateau uranium miners cohort. *J Epidemiol Biostat* 6(6): 417-424.

Kleinbaum, David G., Kupper, Lawrence L., Muller, Keith E., and Nizam, Azhar (1998). Applied Regression Analysis and Other Multivariable Methods, 3rd Ed. Pacific Grove, CA.: Duxbury Press, Brooks/Cole Publishing Co. Pp. 47-48, 118-119.

Lancaster, Peter and Salkauskas, Kestutis (1986). Curve and Surface Fitting An Introduction. New York: Academic Press. Pp. 1-107.

Meyling, Gmelig R.H.J. (1986). Polynomial Spline Approximation in Two Variables. Amsterdam: University of Amsterdam. Pp. i–xxi.

Milton, J. Susan (1999). Statistical Methods in the Biological and Health Sciences. New York: McGraw-Hill. Pp. 418-420.

Nurnberger, G. (1989). Approximation by Spline Functions. New York: Springer-Verlag Berlin Heidelberg. Pp. v-vii, 93-98, 107, 190-194.

Poirier, Dale J. (1976). The Econometrics of Structural Change with Special Emphasis on Spline Functions. New York: North-Holland Publishing Co. Pp. 1-7, 9-12, 21-52, 54-56.

Prenter, P.M. (1975). Splines and Variational Methods. New York: John Wiley and Sons, Inc. Pp. 77-106.

Rothman, Kenneth J., Greenland, Sander (1998). Modern Epidemiology. Philadelphia: Lippincott Williams and Wilkins Publishers. Pp. 392-394.

Ruppert, D. (2001). Review of ``Nonparametric regression and Spline Smoothing" by Randall Eubank, JASA 96, 1523-24.

Sard, Arthur (1971). A Book of Splines. New York: John Wiley and Sons, Inc.

Schumaker, Larry L. (1981). Spline Functions Basic Theory. New York: John Wiley and Sons, Inc. Pp. 1-11, 108-134, 309-316.

Thomas, George B., Finney, Ross L. (1996). Calculus and Analytic Geometry, 9th Edition. New York: Addison-Wesley Publishing Company, Inc. Pp. 17-18, 55, 87-93, 109, 394.

Thurston, Sally, Eisen, Ellen, Schwartz, Joel. (2002). Smoothing in survival methods: an application to workers exposed to metalworking fluids. *Epidemiology* 13(6): 685-692.

Wahba, Grace (1990). Spline Models for Observational Data. Philadelphia: Society for Industrial and Applied Mathematics. Pp. *vi-xii*.

Walpole, Ronald E., Myers, Raymond, H. (1993). Probability and Statistics for Engineers and Scientists. New York: Macmillan Publishing Company. Pp. 89-96, 405-406, 452-456.

CONTACT INFORMATION

Deborah Hurley, MSPH
South Carolina Central Cancer Registry
Dept. of Health and Environmental Control
2600 Bull Street
Columbia, SC 29201
Work Phone: (803) 898-3652
Fax: (803) 898-3599
Email: hurleydm@dhec.sc.gov