ESTIMATING THE NUMBER AND LOCATION OF KNOTS IN SPLINE REGRESSIONS

by

Lawrence C. Marsh

The purpose of this article is to demonstrate a simple method of estimating the number and location of knots (join points) in spline regressions.

Spline regression models offer a convenient alternative to dummy (binary) variable models. Using a dummy variable to alter the intercept or slope of a model generally results in a break in the regression line. For example, a substantial drop in a firms's production may occur if it suddenly closes a major plant.

However, more subtle changes may occur at a given point in time that fundamentally alter the underlying structure without causing a break in the regression line. For example, the passage (or removal) by Congress of an energy tax credit could mark a structural turning point in the demand for home heating oil but cannot be expected to result in an instantaneous drop (or increase) in demand.

Splines piece together the line segments generated by dummy variables to eliminate artificial and inappropriate jumps in the regression line. In higher order polynomials, splines allow for even smoother transitions and more subtle structural shifts by piecing together different polynomial line segments. Sometimes the number and location of these structural shifts is not known and must be estimated.

1. INTRODUCTION

Suits et al. (1978) and Smith (1979) have presented clear and quite useful approaches to estimating spline regression functions with known knot locations. Their work was at least in part based upon the development of this method by such authors as Fuller (1969), Poirier (1973, 1975, 1976) and Buse and Lim (1977).

This paper presents a spline polynomial regression procedure for estimating the number and location of spline knots. This procedure is designed specifically for Hudson's (1966) Type Two splines. Gallant and Fuller (1973) have developed a more general, but more complicated, nonlinear procedure which is more appropriate for Hudson's Type Three splines.

In his dissertation and subsequent journal article Robison

An earlier version of this work was presented under a different title at the SUGI 8 conference in New Orleans. The helpful comments of conference attendees are gratefully acknowledged.

(1964) discussed estimating the point of intersection of two polynomial regressions. He outlined maximum likelihood methods for estimating the regression coefficients and the location of points of intersection of the two regression equations.

Hudson (1966) provided further insights into this problem. He examined four types of join points for joining two polynomial regressions depending primarily upon whether each join point was at an abscissa data point or between such points, and whether the regressions had equal or unequal slopes at the join points. Hudson suggests that in some cases a constrained least squares regression search routine could be used to estimate the location of the join points.

Gallant and Fuller (1973) made an important contribution in dealing with some of the issues raised by Hudson. Using continuity and differentiability conditions, they reparameterized a spline regression model to form a nonlinear regression model. This nonlinear model could then be estimated using Hartley's modified Gauss-Newton method of minimization to obtain least squares estimates of the regression parameters. They also pointed out some asymptotic properties and hypothesis testing implications of this approach.

The estimation technique to be presented herein will not be as sophisticated as the Gallant and Fuller approach but, for Hudson's Type Two models, will provide estimates of the number and location of spline knots using a simple extension of the Suits and Smith methods.

2. ESTIMATING SPLINE REGRESSIONS

The purpose of this paper is to demonstrate a simple method of estimating the number and location of knots in spline regressions. Unlike most previous work, discontinuities are not restricted to the highest order nonzero derivatives, but may occur for one or more derivatives at any knot location. Linear adjustments are shown to generate discontinuities only for the first derivatives. A quadratic adjustment generates discontinuities only for the second derivatives, while cubic adjustments only alter the third derivatives. Any combination of adjustments is permitted at any of the potential knot locations. However, this freedom can be restricted if one so desires.

There are many alternative ways of formulating spline regression models. Smith (1979) offers an approach to specifying spline models that is convenient as a basis for estimating knot locations. In particular, her "+" functions readily lend themselves to this stepwise method for estimating the number and location of spline knots.

Smith initially provides a general model for k knots in an n degree polynomial regression:

$$y = \sum_{j=0}^{n} B_{0,j} X^j + \sum_{i=1}^{k} \sum_{j=0}^{n} B_{i,j}(X - t_i)^j_+ + \varepsilon \qquad (2.1)$$

where y is any continuous dependent variable and X represents an

explanatory variable that has several special values that identify knot locations. There are k of these special X values, and they are designated t_i where i = 1, ...,k. The B's are the regression coefficients and ε is the usual, well-behaved regression error term. This model with no continuity restrictions could be called an unrestricted dummy variable model. Each segment has its unrestricted constant term and slope values. For example, a second degree polynomial with two knots provides for three quadratic sections as follows:

$$y = B_{0,0} + B_{0,1}X + B_{0,2}X^2 + B_{1,0}D1 + B_{1,1}(X - t_1)\,D1 + B_{1,2}(X - t_1)^2\,D1 + B_{2,0}D2 + B_{2,1}(X - t_2)\,D2 + B_{2,2}(X - t_2)^2\,D2 + \varepsilon \qquad (2.2)$$

The dummy variables D1 and D2 are turned on as X passes knots t_1 and t_2 respectively (i.e. $X < t_1$ implies D1 = 0, $X \geq t_1$ implies D1 = 1, $X < t_2$ implies D2 =0 , and $X \geq t_2$ implies D2 =1).

This unrestricted dummy variable model could be estimated as is, or a more refined model could be developed by imposing some continuity restrictions on the model. For example, the polynomial line segments can be made to touch by eliminating the $B_{1,0}$ and $B_{2,0}$ terms. This provides a quadratic spline model which is joined at the knots but may have sharp corners at the join points due to the first derivatives being unequal. This can be smoothed out by setting $B_{1,1}$ and $B_{2,1}$ both equal to zero. This reduces the sharpness of the turning points at the knots by forcing the first derivatives to be equal at each knot:

$$y = B_{0,0} + B_{0,1}X + B_{0,2}X^2 + B_{1,2}(X - t_1)^2\,D1 + B_{2,2}(X - t_2)^2\,D2 + \varepsilon \qquad (2.3)$$

This formulation makes the values of the functions equal at the knots as well as the values of the first derivatives. The second derivatives are left unequal. Of course, if we made the second derivatives equal at the knots as well, we would end up with just one big quadratic equation covering the entire range of data.

By leaving only the highest order, nonzero derivatives unequal, we have what Smith calls the smoothest possible spline, which she expresses in general terms as:

$$y = \sum_{j=0}^{n} B_{0,j}X^j + \sum_{i=1}^{k} B_{i,n}(X - t_i)_+^n + \varepsilon \qquad (2.4)$$

This is a spline model that is as restrictive as it can be without loosing its spline character. It offers some flexibility but is close to the single polynomial equation model.

If the location of the spline knots were known in advance, then we could try estimating various continuity restrictions for the different polynomial segments. For example, if salary is considered to be related to years of education, then one might

assume a spline knot at twelve years for the high school diploma, sixteen years for the BA degree, and eighteen years for the MBA. The degree of the polynomial and continuity restrictions could then be determined in reference to these three knots.

But what if the number and location of knots are not known in advance? The traditional approach to such a problem is to present it as a maximum likelihood estimation problem. This is often done because of the desirable consistency and asymptotic normality properties that are often forthcoming for maximum likelihood estimators.

However, an alternative approach using stepwise regression methods can be used profitably in many cases. Suppose that instead of viewing salary as a function of years of education, we wish to view it as a function of years of experience. Experience may not offer us well defined knot locations the way education did. Instead we may want to estimate the number and location of knots for years of experience.

Say that our data set has five thousand cases but less than one hundred different values for years of experience. We could search over all of the integers from one to one hundred for knot locations or we may wish to search only over the actual values of experience in the data set. In either case, we create a "+" function type dummy variable for each possible knot location. The X variable, experience, might take on values one through seventy-three with seventy-three corresponding knot locations $t_1 = 1$ through $t_{73} = 73$. A corresponding set of dummy variables (D1 through D73) can then be set up such that $Di = 0$ if $X < t_i$ and $Di = 1$ if $X \geq t_i$.

Now assume that a cubic spline model is desired. This means that three sets of spline variables will be needed. These are the linear spline variables: $(X - t_i)\,Di$, the quadratic spline variables: $(X - t_i)^2\,Di$, and the cubic spline variable: $(X - t_i)^3 Di$. Altogether for a cubic spline model with seventy-three possible knot locations, the unrestricted dummy variable model would be:

$$y = B_{0,0} + B_{0,1}X + B_{0,2}X^2 + B_{0,3}X^3 + \sum_{i=1}^{73}\sum_{j=0}^{3} B_{i,j}(X - t_i)^j\,Di + \varepsilon \qquad (2.5)$$

This certainly covers a lot of ground. There are two-hundred and ninety-six coefficients that potentially might be estimated here. Fortunately, a stepwise procedure can be devised to select out the ones that are statistically significant.

The programming requirements for this type of model are fairly well defined. The program should be able to handle an unknown number of cases with an unknown number of unique knot locations where some maximum value is specified for the degree of the polynomial. A stepwise procedure can then be used to select the statistically significant knot locations and the degree of the polynomial appropriate within each spline segment defined by these knot locations. In other words, any combination of polynomials of various degrees may be found to fit the data. No <u>a priori</u> restriction is made on the degree of the polynomial

within each segment except that it not exceed the overall maximum set in advance. Since cubic splines seem fashionable the maximum degree could be set at three. However, there is nothing to prevent that maximum from being set at four or five or even nine or ten if it were desired. Of course, one eventually reaches the limits of reasonable CPU core and time usage.

The next section presents the proposed spline regression estimation procedure using the SAS computer language. Similar stepwise estimation procedures may be possible in other computer languages.

3. SPLINE REGRESSION PROGRAMMING

In this section a SAS spline regression program will be discussed for estimating the number and location of spline knots. The spline knots are statistically selected from a set of potential knot locations that could be specified a priori or could be defined as the set of unique values taken on by a particular observed variable. For example, the latter approach might be appropriate if one wished to restrict the search for knots to the actually observed values of years of education or experience for the individuals in the sample. The former approach might be better if time itself was being used as the variable of interest and one wished to check each and every year from, say, 1930 through 1980 for knots. Of course, if year is the variable of interest and there is one and only one observation per year, then these two approaches provide the same set of potential knot locations. In that case, the number of potential knot locations would be equal to the sample size. As noted above, if the X variable has repeated values, a subset of unique values can be obtained by sorting the data in SAS by using the PROC SORT; BY X; statements and the IF LAST.X THEN OUTPUT; statement.

Once the subset of unique, potential knot location values has been found, then the SAS procedure PROC TRANSPOSE can be used to create the corresponding number of dummy variables needed to identify each potential knot location for the stepwise regression procedure. Alternatively, the transpose function in PROC MATRIX could be used for this purpose.

The large number of dummy variables thus created can then be used to create the corresponding large number of linear, quadratic, and cubic spline terms (and higher order terms if desired). To do this efficiently, array statements must be used since hundreds of variables are to be created. Since the length of these arrays is not known in advance, a method must be devised to create arrays of unknown length. This must be done in such a way that the string of variables thus created can be referred to without knowing the number of variables involved. This may be demonstrated as follows:

SAS SPLINE REGRESSION PROGRAM

```
DATA XY;  INPUT X Y @@;  N+1;  CARDS;
***    data cards    ***
PROC MEANS NOPRINT;  VAR N;
OUTPUT OUT=NUMBER MAX=NCOUNT;
```

```
PROC SORT DATA=XY;  BY X;
DATA REDUCED;  SET;  BY X;
IF LAST.X THEN OUTPUT;  KEEP X;

PROC TRANSPOSE DATA=REDUCED PREFIX=KNOT;
DATA KNOTS;  SET;  KNOTEND=1;
PROC TRANSPOSE DATA=REDUCED PREFIX=D;
DATA DS;  SET;  DEND=1;
PROC TRANSPOSE DATA=REDUCED PREFIX=L;
DATA ARRAYL;  SET;  LEND=1;
PROC TRANSPOSE DATA=REDUCED PREFIX=Q;
DATA ARRAYQ;  SET;  QEND=1;
PROC TRANSPOSE DATA=REDUCED PREFIX=C;
DATA ARRAYC;  SET;  CEND=1;

DATA MATCH;  MERGE KNOTS DS ARRAYL ARRAYQ ARRAYC NUMBER;
DO I=1 TO NCOUNT;  OUTPUT;  END;
DROP  _NAME_  I  NCOUNT;

DATA;  MERGE XY MATCH;
ARRAY KNOT KNOT1--KNOTEND;
ARRAY D D1--DEND;
ARRAY L L1--LEND;
ARRAY Q Q1--QEND;
ARRAY C C1--CEND;
DO OVER KNOT;
IF X LT KNOT THEN D=0;
IF X GE KNOT THEN D=1;
L=D*(X-KNOT);
Q=D*(X-KNOT)**2;
C=D*(X-KNOT)**3;
END;  X2=X**2;  X3=X**3;

PROC STEPWISE;  MODEL Y=X X2 X3 L1--LEND Q1--QEND C1--CEND /
INCLUDE=1 SLE=1 SLS=1;
```

Figures 1 and 2 were generated by replacing PROC STEPWISE with:

```
PROC REG;
MODEL Y = X C36 Q50 L63 L89;
OUTPUT OUT=B P=YFIT;
GOPTIONS DEVICE=TEK4010;
PROC GPLOT DATA=B;
PLOT Y*X YFIT*X;
SYMBOL2 I=SPLINE;
```

After using PROC TRANSPOSE to transpose the subset of potential knot location values, the single observation that results has an unknown number of newly created variables. That unknown number is the number of potential knot location values and corresponds to the number of unique values of X in the data set. If one wished to search over a prespecified set of values instead of using the observed X values, a variable Z containing those values could simply be loaded into the data set called REDUCED just prior to the first PROC TRANSPOSE.

In order to be able to refer to a string of variables in an ARRAY statement, we need to know the name of the first variable and the last variable. The first variable name is simple enough. It is the prefix with the number 1 attached, which is KNOT1 in this case. The last variable is the problem. There is an unknown number of variables here. We don't know what the name of the last variable will be, so we just throw in a variable whose name we do know. By creating a single additional variable at this point (e.g. KNOTEND=1), the string of variables KNOT1--

KNOTEND may be referred to without knowing the number of variables in the string.

In a similar manner the corresponding dummy variables and linear, quadratic and cubic variables can be referenced by array statements without knowing their length. The initial values loaded into these array data sets are needed only to create the variable names. The proper values are assigned to these variables in the final DO loop.

Once all of the appropriate spline regression variables have thus been created, the PROC STEPWISE regression procedure can be used to pick out the knots that are statistically significant. The line segments that are fitted in this manner may represent a combination of linear, quadratic and cubic equations that are joined at the knots. If one wishes to allow the functions to be unequal at the knots then the D1--DEND variables should be included in the MODEL statement for PROC STEPWISE. This will allow the functions to be unequal at the knots if such a break in the function is statistically significant. Any of the derivatives may be unequal at the knots as well. Thus, any combination is possible ranging from the completely unrestricted dummy variable model with many segments to the single polynomial function model with no knots.

4. AN INTEREST RATE APPLICATION

Estimating a spline model for interest rates on commercial bonds provides a convenient example of determining the number and location of spline knots. The New York City open market rates for four-to-six month commercial paper with Aa rating or equivalent will be used for the period 1890 through 1981.

The stepwise regression procedure selected terms for the spline regression model that attained a level of significance of at least .01 or better. The Y variable is the interest rate and the X variable is the year. Variables beginning with the letter C represent cubic adjustments, those with Q represents quadratic adjustments, and L stands for linear adjustments. The number following these letters indicates the knot location and corresponds to the "i" subscript in (2.1) and (2.5). Table 1 displays the spline regression model results.

Table 1. Interest Rate, Y, as a Function of Time in Years, X

Variable Name	Estimated Regression Coefficient	Student t Statistics	Prob Value
INTERCEPT	144.47767186	9.0990	.0001
X	-.07291068	-8.8004	.0001
C36	-.00068578	-5.7497	.0001
Q50	.08271411	5.7691	.0001
L63	-.64603088	-2.9252	.0044
L89	2.74175672	6.6394	.0001

N = 92 R^2 = .8574 $\bar{R}^2$ = .8491 F = 103.4 F-Prob < .0001

In general terms these results provide the following functional form:

$$y = B_{0,0} + B_{0,1}X + B_{36,3}(X - t_{36})^3\ D36 + B_{50,2}(X - t_{50})^2\ D50$$

$$+ B_{63,1}(X - t_{63})\ D63 + B_{89,1}(X - t_{89})\ D89 + \varepsilon \qquad (4.1)$$

where $B_{0,0}$ is the intercept term, $B_{0,1}$ is the coefficient of X, $B_{36,3}$ is the C36 coefficient, $B_{50,2}$ is the Q50 coefficient, $B_{63,1}$ is the L63 coefficient, and $B_{89,1}$ is the L89 coefficient.

Substituting in for the estimated coefficient values and knot locations results in the following fitted values for Y:

$$y = 144 - .073\ X - .0007\ (X - 1925)^3\ D36 + .083\ (X - 1939)^2\ D50$$

$$-.646\ (X - 1952)\ D63 + 2.74\ (X - 1978)\ D89 \qquad (4.2)$$

Note that observation 36 is 1925, observation 50 is 1939, observation 63 is 1952, and observation 89 is 1978. Five separate equations represent the five time period segments found by collecting terms on X in (4.2).

1890 - 1924: $y = 144 - .073\ X$ (4.3)

1925 - 1938: $y = 4{,}892{,}038 - 7623.8\ X + 3.96\ X^2 - .0007\ X^3$ (4.4)

1939 - 1951: $y = 5{,}203{,}020 - 7944.6\ X + 4.04\ X^2 - .0007\ X^3$ (4.5)

1952 - 1977: $y = 5{,}204{,}281 - 7945.2\ X + 4.04\ X^2 - .0007\ X^3$ (4.6)

1978 - 1981: $y = 5{,}198{,}858 - 7942.5\ X + 4.04\ X^2 - .0007\ X^3$ (4.7)

The initial linear relationship becomes cubic in 1925. The intercept term and the coefficient for X adjust dramatically to compensate for the introduction of the cubic and quadratic terms. These relationships can be seen in Figures 1 and 2.

The first, second, and third derivatives can be derived from (4.1) as follows:

$$\frac{dy}{dX} = B_{0,1} + 3\ B_{36,3}(X - t_{36})^2\ D36 + 2\ B_{50,2}(X - t_{50})\ D50$$

$$+ B_{63,1}\ D63 + B_{89,1}\ D89 \qquad (4.8)$$

$$\frac{d^2y}{dX^2} = 6\ B_{36,3}(X - t_{36})\ D36 + 2\ B_{50,2}\ D50 \qquad (4.9)$$

$$\frac{d^3y}{dX^3} = 6\ B_{36,3}\ D36 \qquad (4.10)$$

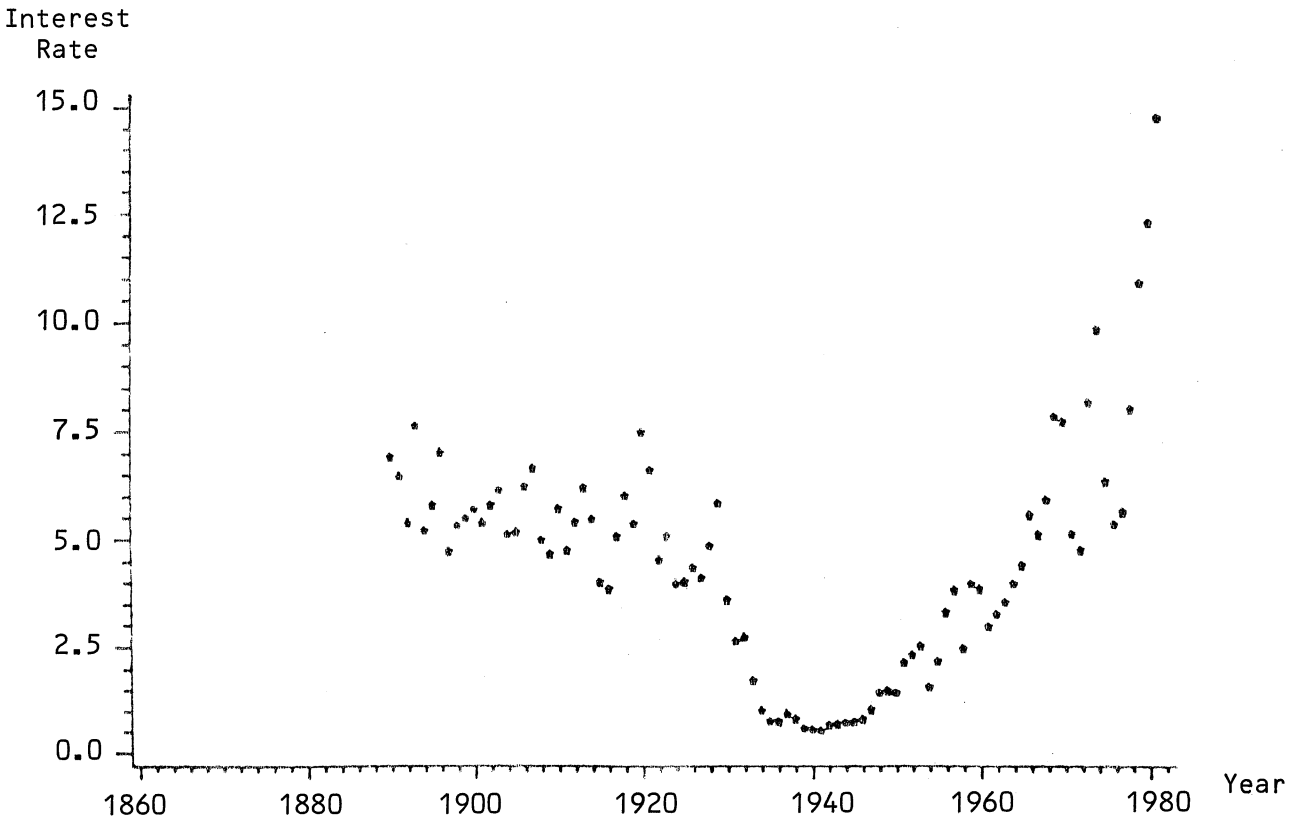

Figure 1. Interest Rate on Commercial Bonds: Original Data Points

Figure 2. Interest Rate on Commercial Bonds: Spline Regression Function

The functions themselves as well as their first and second derivatives are equal at the 1925 knot. The third derivatives are not equal at this first join point.

A quadratic adjustment takes place in 1939. At this second join point the first and third derivatives are equal but not the second derivatives. This allows for an interesting minimizing loop in the predicted interest rates around 1940. In 1952, a discontinuity takes place in the first derivative due to a linear adjustment. However, the functions and their corresponding second and third derivatives are equal at the 1952 knot.

Another final linear adjustment takes place in 1978 which substantially straightens out the fitted relationships. The general effect on the derivatives of this linear adjustment is by necessity the same as the 1952 linear adjustment. In other words, linear adjustments result in discontinuities only for the first derivatives. Similarly, quadratic adjustments cause discontinuities only for the second derivatives while cubic adjustments only generate discontinuities for the third derivatives.

In summary, this spline regression model shows that although the interest rate on commercial bonds had been gradually falling since at least the late 1890's, a dramatic downward slide in this interest rate occurred with the imposition of large import tariffs and the onset of the great depression. This lasted from about 1925 when interest rates took a downward turn with a cubic adjustment, C36, to around 1939 when interest rates bottomed out with the help of a quadratic adjustment, Q50.

Interest rates continued rising throughout World War II until modified by a linear adjustment, L63, at the conclusion of the Korean Conflict. A final linear adjustment, L89, took place in the late 1970's as the full impact of the energy crisis set in and interest rates rose dramatically.

5. CONCLUSIONS

This paper provided a method for estimating the number and location of spline knots (join points) for spline regression models. A SAS program for carrying out this estimation has been explained and demonstrated. The interest rate example has shown the power of this technique for fitting spline regressions. An analysis of the derivatives shows how tightly or loosely the polynomial line segments are connected at each join point. The interest rate for prime commercial paper provided some dramatic changes that demonstrated the need for the flexibility of the polynomial spline fitting regression technique.

REFERENCES

Buse, A. and Lim, L. (1977), "Cubic Splines as a Special Case of Restricted Least Squares," Journal of the American Statistical Association, 72, pp. 64-68.

Fuller, W.A. (1969), "Grafted Polynomials as Approximating Functions," Australian Journal of Agricultural Economics, pp. 35-46.

Gallant, A.R. and Fuller, W.A. (1973), "Fitting Segmented Polynomial Regression Models Whose Join Points Have to be Estimated," Journal of the American Statistical Association, 68, pp. 144-147.

Hudson, D.J. (1966), "Fitted Segmented Curves Whose Join Points Have to be Estimated," Journal of the American Statistical Association, 61, pp. 1097-1129.

Poirier, D.J. (1973), "Piecewise Regression Using Cubic Splines," Journal of the American Statistical Association, 68, pp. 515-524.

_____ (1975), "On the Use of Bilinear Splines in Economics," Journal of Econometrics, 3, pp. 23-24.

_____ (1976), The Econometrics of Structural Change with Special Emphasis on Spline Functions, Amsterdam: North Holland Publishing Company.

Robison, D.E. (1964), "Estimates for the Points of Intersection of Two Polynomial Regressions," Journal of the American Statistical Association, 59, pp. 214-224.

Smith, P.L. (1979), "Splines as a Useful and Convenient Statistical Tool," The American Statistician, 33, pp. 57-62.

Suits, D.B., Mason, A., and Chan, L. (1978), "Spline Functions Fitted by Standard Regression Methods," Review of Economics and Statistics, 60, pp. 132-139.