IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN:2319-765X. Volume 10, Issue 1 Ver. II. (Feb. 2014), PP 42-52
www.iosrjournals.org

Optimal Choice of Splines and Knots in TPSPLINE and TRANSREG Procedures

Rashmi Aggarwal[1], Suresh Kumar Sharma[1], Kanchan Jain[1]
[1](*Department of Statistics, Panjab University, Chandigarh, India)*

Abstract: *Multivariate functions observed with noise can be approximated by thin-plate smoothing splines. It does not depend on the assumptions of the parametric model. The amount of smoothing can be judged by generalized cross validation. If there is no prior knowledge about the model and the data is unable to represent a model with fixed number of parameters then TPSPLINE procedure is appropriate. With the increase in sample size, the model space also increases but this situation can be handled by thin-plate smoothing spline i.e. it is suitable for complicated situations. We worked on optimal choice of knots in TPSPLINE procedure. The procedure has been demonstrated by taking real life data set. The TRANSREG procedure which is applicable to many models including ordinary, multiple, multivariate regression with variable transformations is also discussed. It can also fit regression functions with smooth, spline or penalized B-splines. This procedure uses the method of alternating least-squares, that is, finding least-square estimates of the model parameters given the current scoring of the data, and then finding least-square estimates of the scoring parameters based on the current set of model parameters. In this paper, the fitting of the model through penalized B-splines using splines for AICC, AIC, CV, SBC and GCV criterion have been discussed and compared by taking a real life data set.*
Keywords*: Alternating least squares, Knots, Penalized B-splines, Piecewise polynomials, Spline*

I. Introduction

The penalized least square method, used by TPSPLINE procedure is appropriate to fit a non-parametric regression model. This procedure is more flexible in terms of regression surface i.e. it fits the data, particularly, where number of parameters can be as large as unique data points. It makes no assumption of a parametric form of the model. The generalized cross Validation (GCV) [1] or AIC function is generally used to select the amount of smoothing. Many regression procedures available in SAS such as REG, GLM and NLIN can be complemented with this procedure. The main features of TPSPLINE procedure are that it fits both semi parametric and non-parametric models, supports use of multidimensional data, provides penalized least square estimates, provides options for handling large datasets, supports multiple dependent variables and enables to choose a model by specifying the appropriate degrees of freedom or the optimal smoothing parameter. Penalized least squares regression and generalizations have been studied extensively over the years. Initiated by Kimeldrof and Wabha [5, 6, 7], it has also been carried forward by Wabha [10], Green and Silverman [13], and Gu [14] for comprehensive studies.

Another procedure that we worked on in this paper is TRANSREG procedure. This procedure makes transformations on dependent and independent variables and then fits linear models, optionally with spline, smooth and other non-linear transformations. The procedure may be used to fit a curve though a scatterplot or fit multiple curves, one for each level of a classification variable. It can also be used to code experimental designs and classification of variables prior to their use in the analysis. The SAS codes have been used to demonstrate these procedures. A spline is a smooth function derived using piecewise polynomials. The B-spline is also a piecewise polynomial function of degree k in a variable x. It is defined over a range $t_1 \le x \le t_m$, $m = k + 2$. The points where $x = t_j$ are known as break-points or knots. B-spline is unique, for any given set of knots. The basic idea of using B-splines is that any spline function of degree k on a given set of knots can be expressed as a linear combination of B-splines. PB-spline stands for "penalized B-spline", where the coefficients are determined partly by the data, and partly by an additional penalty function that imposes smoothness to avoid over fitting. The basic GCV and splines definitions required to implement these procedures are given below:

1.1 Generalized Cross Validation (GCV)

Cross Validation works by leaving points (x_i, y_i) estimating the smooth at x_i based on remaining (n-1) points. The cross-validation sum of squares is

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_{\lambda}^{-i}(x_i))^2$$

where $\hat{f}_{\lambda}^{-i}(x_i)$ indicates the fit at x_i computed by leaving out the i[th] data point. There is simple way to define $\hat{f}_{\lambda}^{-i}(x_i)$ given only smoother matrix S_λ. The generalized cross-validation (GCV) can be written as

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_\lambda(x_i)}{1 - tr(S_\lambda)/n}\right\}^2 , where \sum_{\substack{j=1 \\ j\neq i}}^{n} S_{ij}(\lambda)y_j = \hat{f}_\lambda(x_i).$$

1.2 Splines

A spline is a smooth function, basically a piecewise polynomial and it possesses a high degree of smoothness at the places where the polynomial pieces connect. It is a piecewise polynomial real function
S : [a, b] → R
On an interval [a, b] composed of k subintervals $[t_{i-1}, t_i]$ with $a = t_0 < t_1 < \ldots < t_{k-1} < t_k = b$. The restriction of S to an interval *i* is a polynomial $P_i: [t_{i=1}, t_i] \rightarrow R$ so that $S(t) = P_1(t), t_0 \leq t \leq t_1$,
$S(t) = P_2(t), t_1 \leq t \leq t_2$,
:
$S(t) = P_k(t), t_{k-1} \leq t \leq t_k$.
The order of spline S is the highest order of the polynomials $P_i(t)$.

In this paper, we discussed the optimal choice of splines and knots to use TPSPLINE and TRANSREG procedures. In section II, we discussed TPSPLINE model, the estimation of the model parameters, confidence intervals and choice of smoothing parameter. Section III describes the use of TRANSREG procedure, optimal choice of splines and knots and an example to demonstrate the procedure by taking a real life data. In section IV, the PBSPLINE transformations and their usefulness to real life data have been explored. Conclusions are given in section V.

II. TPSPLINE Model

The Penalized least squares method provides versatile and effective nonparametric models for regression with Gaussian responses. The estimates provided by this method usually balance the fitting of the data and avoid excessive roughness or rapid variation. Generally, the estimate based on penalized least square satisfies the regularity conditions. Model can be constructed by considering a d-dimensional covariate vector z_i, and a p-dimensional covariate vector x_i. The dependent variable y_i's are associated with (z_i, x_i). The relation between z_i and y_i is unknown; however, we assume that the relationship between x_i and y_i is linear. Following semi-parametric model can be used to fit the data:

$$y_i = f(z_i) + x_i\beta + \epsilon_i$$

where *f* is an unknown function assumed to be reasonable smooth and $f \in H_m$. Here H_m is a space of functions whose partial derivatives of total order m are in $J_2(E^d)$, where E^d is the domain of z. The error ϵ_i, $i = 1, \ldots, n$ are independent with mean zero and β is a p-dimensional unknown parametric vector. The above model consists of two parts, i.e. $x_i\beta$, the parametric part where x_i are the regression variables and the other part $f(z_i)$ which is non-parametric and z_i are the smoothing variables.

2.1 Estimation of the Model Parameters

One may minimize the following quantity by using least squares method to estimate $f(z_i)$ and β

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(z_i) + x_i\beta)^2.$$

However, the functional space of *f(z)* is large but one can always find a function *f* that interpolates the data points. Penalized least square method helps us to obtain an estimate that fits the data well and has some degree of smoothness.

The penalized least square function is defined as

$$S_\lambda(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(z_i) - x_i\beta)^2 + \lambda L_m(f) \qquad (1)$$

where the penalty on the roughness of *f* is defined in terms *of* $\lambda L_m(f)$. The first term measures the goodness of fit and second measures the smoothness associated with f in (1). The trade-off between smoothness and goodness of fit is governed by smoothing parameter λ. When λ is small, it puts more emphasis on the goodness of fit and it heavily penalizes estimates with large second derivatives, when λ is large.

The penalty $L_m(f)$ for thin-plate smoothing spline with Z of dimension d is defined as

$$L_m(f) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \sum \frac{m!}{\alpha_1! \ldots \alpha_d!}\left(\frac{\partial^m f}{\partial {z_1}^{\alpha_1} \ldots \partial {z_d}^{\alpha_d}}\right)^2 dz_1 \ldots dz_d$$

where $\sum_i \alpha_i = m$. It may be noted that $L_m(f)$ may give zero penalty to some functions. The space that is spanned by the set of polynomials contributing zero penalty is generally called the polynomial space. The dimension of

the polynomial space *M* is a function of dimension d and order m of the smoothing penalty. For d=2 and m=2, $L_m(f)$ becomes:

$$L_2(f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\left(\frac{\partial^2 f}{\partial {z_1}^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial z_1 \partial z_2}\right)^2 + \left(\frac{\partial^2 f}{\partial {z_2}^2}\right)^2\right) dz_1 dz_2$$

and *M* = cardinality of $(\{1,\ z_1,\ z_2\}) = 3$.

Note that the order m of smoothing penalty and dimension d must satisfy the condition 2m-d >0. For the sake of simplicity, the formulas and equations generally assumes m = 2.

f_λ can be represented as

$$f_\lambda = \theta_0 + \sum_{j=1}^{d}\theta_j z_{ij} + \sum_{j=1}^{n}\delta_j E_2(z_i - z_j)$$

where $E_2(s) = \frac{1}{2^3\pi}\|s\|^2 log\|(s)\|$ for d = 2.

If we define K with elements $K_{ij} = E_2(z_i - z_j)$ and T with elements $T_{ij} = (Z_{ij})$, the goal is to find coefficients $\beta, \theta\ and\ \delta$ that minimize

$$S_\lambda(\beta, \theta, \delta) = \frac{1}{n}\|y - T\theta - K\delta - X\beta\|^2 + \lambda\delta^T K\delta.$$

A unique solution of the above equation is feasible if the matrix T is of full rank and $\delta^T K\delta \geq 0$.

If $\alpha = \begin{pmatrix}\theta\\ \beta\end{pmatrix}$ and $Z = (T\ X)$, the expression for S_λ can be written as:

$$\frac{1}{n}\|y - Z\alpha - K\delta\|^2 + \lambda\delta^T K\delta.$$

The unknown coefficients α and β can be obtained by solving

$$(K + n\lambda I_n)\delta + Z\alpha = y \qquad (2)$$
$$and \qquad Z^T\delta = 0.$$

In order to compute α and β, we partition Z as

$$Z = (Q_1 Q_2)\begin{pmatrix}R\\ 0\end{pmatrix}$$

Note that $(Q_1\ Q_2)$ is an orthogonal and R is an upper triangular matrix satisfying the condition $Z^T Q_2 = 0$.

δ is the column space of Q_2, since $Z^T\delta = 0$. Therefore, for a vector γ, δ can be expressed as $\delta = Q_2\gamma$. Substituting $\delta = Q_2\gamma$ into (2) and multiplying throughout by Q_2^T gives

$$Q_2^T(K + n\lambda I_n)Q_2\gamma = Q_2^T y$$
$$\text{or } \delta = Q_2\gamma = Q_2[Q_2^T(K + n\lambda I)Q_2]^{-1}Q_2^T y.$$

The unknown constant α can be obtained by solving

$$R\alpha = Q_1^T[y - (K + n\lambda I)]\delta$$

and therefore the influence matrix $A(\lambda)$ can be defined as

$$\hat{y} = A(\lambda)y$$

where $$A(\lambda) = I - n\lambda Q_2[Q_2^T(K + n\lambda I)Q_2]^{-1}Q_2^T. \qquad (3)$$

If we consider the trace of $A(\lambda)$ as the degrees of freedom for the model and the trace of $(1 - A(\lambda))$ as the degrees of freedom for the error, then the estimate σ^2 is given by

$$\hat{\sigma}^2 = \frac{RSS(\lambda)}{tr(I - A(\lambda))} \qquad (4)$$

where $RSS(\lambda)$denote the residual sum of squares.

2.2 Confidence Intervals for spline

The confidence intervals for smoothing spline estimates can be obtained from:

$$\widehat{f}_\lambda(z_i) \pm z_{\alpha/2}\sqrt{\hat{\sigma}^2 a_{ii}(\lambda)} \qquad (5)$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution and$a_{ii}(\lambda)$is the ith diagonal element of $A(\lambda)$ matrix.

2.3 Choice of Smoothing Parameter λ

The smoothing parameter λ governs the trade-off between the goodness of fit and the smoothness of the estimate. A small λ places less of a penalty on the rapid change in $f^{(m)}$, resulting in estimate that tends to interpolate the data points. A large value of λ heavily penalizes the mth derivative of the function, thus forcing $f^{(m)}$ close to zero. One way to tackle the situation is to perform several analyses with different values of λ and compare the estimates. The more effective and sensitive way to select the smoothing parameter is to use cross validation function, as this is an approximation to predicted mean squared error. The Generalized Cross Validation (GCV) function is defined as:

$$GCV(\lambda) = \frac{(1/n) \parallel \mathrm{I} - \mathrm{A}(\lambda)\mathrm{y} \parallel^2}{[(1/n)tr(\mathrm{I} - \mathrm{A}(\lambda)\mathrm{y}]^2}$$

where $\mathrm{A}(\lambda)$ is defined in (3).

The choice of λ in GCV function is based on asymptotic theory. Usually, good results are not expected from small sample sizes (as there is less information in the data to separate model from error component). The simulation studies suggest that for independent and identically distributed Gaussian noise; one can obtain reliable estimates of λ for n greater than 25. This is the reason that GCV is fairly robust under non-homogeneity of variances and non-Gaussian errors. In practice, calculation with λ=0 or λ close to 0 results into unsatisfactory solution. Moreover, simulation studies reveals that a λ with $\log_{10}(n\lambda)$> -8 is considered to be small enough that the final estimate based on this λ almost interpolates the data points. Thus a GCV value based on a $\lambda \leq 10^{-8}$ might not be accurate.

2.4 Example

Suspended particulate matter (SPM) in air is generally considered to be all airborne solid and low vapour pressure liquid particles. Suspended particulate matter consists of a spectrum of aerodynamic particles and its size ranges from below 0.01 μm to 100 μm and larger. The usual standard for TSP is in terms of PM_{10} standard that includes particles with an aerodynamic diameter of 10 μm or less. PM_{10} is usually measured in the units of $10\mu g/m^3$. In order to model concentration of PM_{10}, we considered three variables x_1, x_2 and y. Our aim is to fit a surface by using the variables x_1, x_2 and to model the response variable y i.e. PM_{10} concentration. On [-1 x 1] x [-1 x 1] square the variables x_1 and x_2 are evenly spaced, and the response variable y is generated by adding a random error to a function $f(x_1, x_2)$. We considered 50 values of PM_{10} along with equally spaced x_1, x_2. In order to visualize replicates, half of the data points are shifted a little bit by adding a small value (0.001) to x_2 values. The raw data is plotted using G3D procedure (Fig. 1) in SAS as follows:

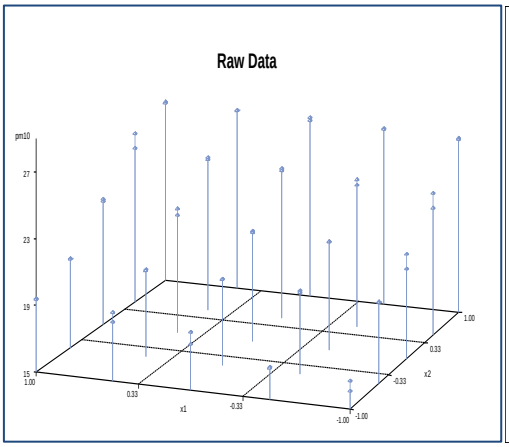

Figure 1

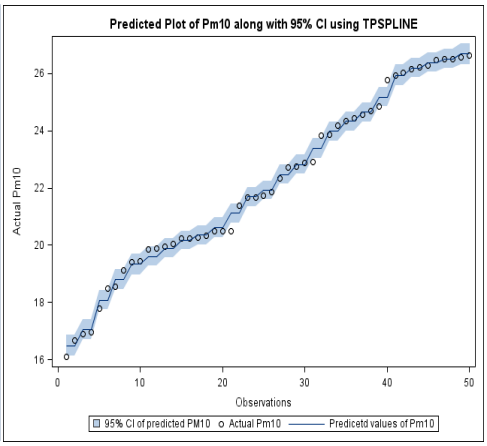

Figure 2

The TPSPLINE procedure was applied to the above datasets with x_1 and x_2 as smoothing variables. The LOGNLAMBDA option provides a list of GCV values with $\log_{10}(n\lambda)$ ranging from -3 to -1. The above graph consists of 50 observations with 25 unique design points. The final model contains no parametric regression terms but two smoothing variables. The ordered derivative in the penalty is 2 (by default) and the dimension of polynomial space is 3. It is evident that GCV function is minimized at 0.127468 for which $\log_{10}(n\lambda)$ = -2.1 (round-off). The final thin-plate smoothing spline estimate is based on LOGNLAMBDA =-2.0688. The residual sum of squares (RSS) is 2.4933 and the degrees of freedom are 18.7204. The standard deviation, computed from equation (4) is 0.2823. The predicted values along with 95% confidence limits (CL), computed using (5), are given in Fig. 2.

Table 1: The $\log_{10}(n\lambda)$ and GCV

GCV Function

log10(n*Lambda)	GCV
-3.000000	0.149338
-2.900000	0.146840
-2.800000	0.144076
-2.700000	0.141108
-2.600000	0.138044
-2.500000	0.135030
-2.400000	0.132254
-2.300000	0.129929
-2.200000	0.128269
-2.100000	0.127468*
-2.000000	0.127667
-1.900000	0.128932
-1.800000	0.131244
-1.700000	0.134491
-1.600000	0.138486
-1.500000	0.142988
-1.400000	0.147733
-1.300000	0.152471
-1.200000	0.156993
-1.100000	0.161154
-1.000000	0.164881

Note: * indicates minimum GCV value.

Summary Statistics of Final Estimation

log10(n*Lambda)	-2.0688
Smoothing Penalty	124.7272
Residual SS	2.4933
Tr(I-A)	31.2796
Model DF	18.7204
Standard Deviation	0.2823

III. TRANSREG Procedure

PROC TRANSREG performs transformation regression in which both the outcome and predictors(s) can be transformed and splines can be fitted. Splines are basically piecewise polynomials that can be used to estimate relationships that are difficult to fit with a single function. The applications of TRANSREG procedure can be extended to following types of linear models:

- Ordinary regression and Analysis of Variance
- Metric and non-metric conjoint analysis [4, 2]
- Linear models with Box-Cox transformations of the dependent variables.
- Regression with a smooth [8], spline[9, 12], monotone spline [11], or penalized B-Spline [3] fit function
- Simple and multivariate regression with variable transformations
- Redundancy analysis with transformations
- Canonical correlation analysis
- Response surface regression analysis

PROC TRANSREG extends the simple general linear regression model by using optimal transformations (on dependent and/or independent variables) that are iteratively derived. It works on the principal of alternating least squares i.e. first finding least-squares estimates of the model parameters given the current scoring of the data, and then finding least-squares estimates of the scoring parameters given the current set of model parameters. The Output Delivery System (ODS) in SAS helps us to draw fitting, residual and other plots.

3.1 Splines and Knots

The choice of splines and knots is not easy to make in TRANSREG procedure. Splines are generally smooth curves assumed to be continuous. Splines are derived using piecewise polynomials of degree n and the first (n-1) derivatives that agree at the points where they join. *Knots* are the joining points on the X-axis values. The term "spline" is also used for polynomials (splines with no knots) and piecewise polynomials with more than one discontinuous derivative. Splines with more knots are generally less smooth as compared to splines with few knots. Knots give the freedom to bend the curve more closely to follow the data. However, specifying fewer (more) knots than required may results into under-fitting (over-fitting) of the model and thus a proper balance has to be maintained while justifying the number of knots. Generally, the number of knots is kept small (less than 10, although one can specify more). Most of the studies reveals that a degree three spline with nine knots (one placed at each decile), can easily fit into a large variety of curves. Fitting of the model involves (p+q) parameters where p is the degree of each spline transformation with q knots. TRANSREG procedure assumes that the total number of model parameters must be less than the number of observations. It is recommended that

each parameter should have at least five to ten observations to get more reliable results. For example, when spline transformations of degree 3 and with 9 knots are used for six variables then the required number of observations must be at least 5 or 10 times of 72 (because the total number of parameter is 6 x (3+9)). For n observations, if we specify q knots, then each of the (q+1) segments of the spline must contain n/(q+1) observations on an average. It is important to have sufficient number of observations in each interval.

3.2 Fitting a Curve through a Scatter Plot

We have several options of transformations available as in variable expansions, non-optimal transformations, nonlinear fit transformations, optimal transformations and some other transformations. Here, we will walk through an example of PROC TRANSREG with the spline option and explore its defaults.

Serum Glutamic Oxaloacetic transaminase (SGOT) is an enzyme that is normally present in liver and heart cells. It is subjected to biochemical changes after death. A sample of 214 subjects was taken, in which the SGOT levels and Time since Death (TSD) was available during post mortem period. The sample consists of three modes of death which includes accidental cases (n=101), burn (n=63) and poison (n=50). The time of death ranges from 2 to 22 hours and SGOT from 80 to 1090 U/L. In this example, we have explored different transformations to estimate TSD using TRANSREG procedure.

TRANSREG procedure fit curves through data and detect nonlinear relationships among variables. The dependent variable TSD was specified with an IDENTITY transformation, which means that it will not be transformed as in ordinary regression. The independent variable SGOT is transformed by using a cubic spline with 4 knots (roughly, decided based on the shape of the curve). The results are given in Table 3.

Table 3: Iteration, ANOVA and Regression Results

Dependent Variable Identity(TSD)

TRANSREG MORALS Algorithm Iteration History for Identity(TSD)

Iteration Number	Average Change	Maximum Change	R-Square	Criterion Change	Note
0	0.97470	3.30079	0.06033		
1	0.00000	0.00000	0.89185	0.83152	Converged

The TRANSREG Procedure Hypothesis Tests for Identity(TSD)

Univariate ANOVA Table Based on the Usual Degrees of Freedom

Source	DF	Sum of Squares	Mean Square	F Value	Pr> F
Model	7	5896.018	842.2882	242.67	<.0001
Error	206	714.998	3.4709		
Corrected Total	213	6611.015			

Root MSE	1.86303	R-Square	0.8918
Dependent Mean	11.81210	Adj R-Sq	0.8882
Coeff Var	15.77217		

Univariate Regression Table Based on the Usual Degrees of Freedom

Variable	DF	Type II Sum of Coefficient	Mean Squares	Square	F Value	Pr> F	Label
Intercept	1	1.00000000	21190.6	21190.6	6105.28	<.0001	Intercept
Spline(SGOT)	7	1.00000000	5896.0	842.3	242.67	<.0001	SGOT

PROC TRANSREG increases the squared multiple correlation from the original value of 0.06033 to 0.89185. Iteration 0 shows the fit before the transformation and iteration 1 after the transformation. The ANOVA results show a model fitting with 7 model parameters (4 knots plus a degree 3 for cubic spline). The transformations are shown in Fig. 3.

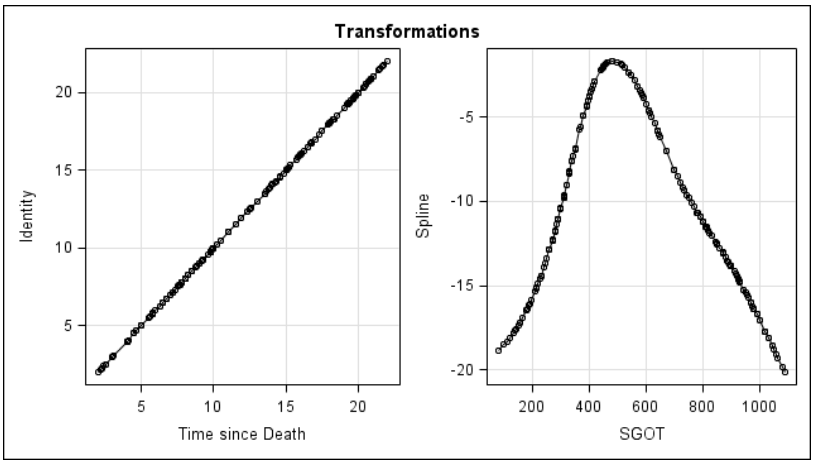

Figure 3: The transformation plots: Identity transformation for TSD and non-linear transformation for SGOT

Fig. 4 shows the residuals as a function of the transformed independent variable SGOT. The spline regression fit for TSD shown in Fig. 5 which displays non-linear regression function plotted through original data along with 95% confidence and prediction limits. The values of TSD are higher in the range of SGOT, 400 to 600 U/L and much lower for extreme values of SGOT.

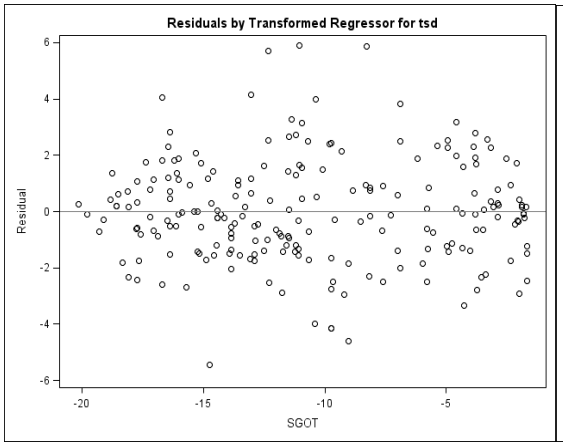

Figure 4

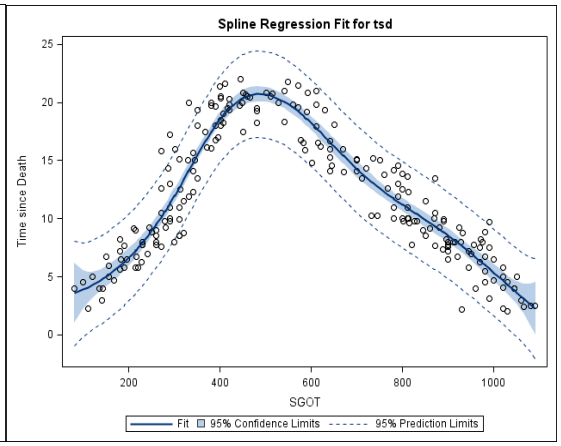

Figure 5

The observed by predicted plot for TSD is shown in Figure 6, where the dependent variable is plotted as a function of the regression predicted values with a linear regression line. The residual differences between the transformed data and the regression line show how well the non-linearly transformed data fit a linear regression model. The residuals are scattered mostly at random.

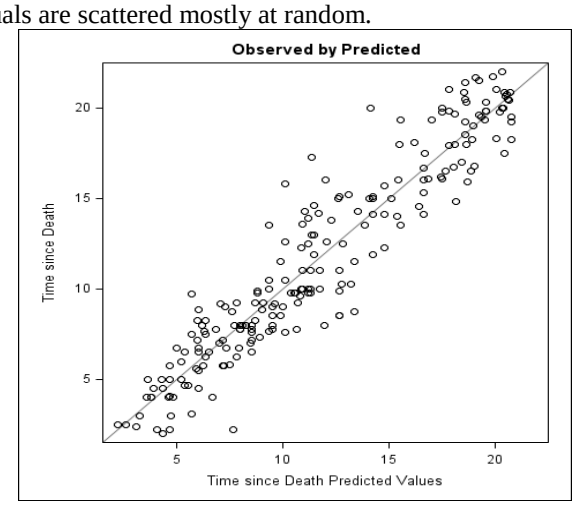

Figure 6

Separate functions were also plotted for each mode of casualty i.e. accident, burn and poison. Table 4 shows the iteration and regression results. Note that the PROC TRANSREG increases the squared multiple

correlation from original value of 0.07064 to 0.93679. The important point is that there is convergence for each category of mode.

Table 4: Iteration, ANOVA and Regression Results (For each mode of casualty)

```
                        The TRANSREG Procedure

                    Dependent Variable Identity(TSD)

      TRANSREG MORALS Algorithm Iteration History for Identity(TSD)

 Iteration    Average    Maximum                 Criterion
    Number     Change     Change    R-Square       Change    Note
         0    0.44880    3.79618     0.07064
         1    0.00000    0.00000     0.93679      0.86615    Converged
Algorithm converged.

      Hypothesis Test Iterations Excluding Spline(Mode Accident SGOT)
      TRANSREG MORALS Algorithm Iteration History for Identity(TSD)

 Iteration    Average    Maximum                 Criterion
    Number     Change     Change    R-Square       Change    Note
         0          0    0.00000     0.00000      0.49445
         1    0.00000    0.00000     0.49445      0.00000    Converged
Algorithm converged.

        Hypothesis Test Iterations Excluding Spline(Mode Burn SGOT)
      TRANSREG MORALS Algorithm Iteration History for Identity(TSD)

 Iteration    Average    Maximum                 Criterion
    Number     Change     Change    R-Square       Change    Note

         0    0.00000    0.00000     0.64490
         1    0.00000    0.00000     0.64490      -.00000    Converged
Algorithm converged.

       Hypothesis Test Iterations Excluding Spline(Mode Poison SGOT)
      TRANSREG MORALS Algorithm Iteration History for Identity(TSD)

 Iteration    Average    Maximum                 Criterion
    Number     Change     Change    R-Square       Change    Note
         0    0.00000    0.00000     0.73741
         1    0.00000    0.00000     0.73741      0.00000    Converged
Algorithm converged.
```

The transformations for all the three groups have been shown in Fig. 7.There is approximate quadratic fit to all the three categories of mode. Fig. 8 shows distinct functions of the data for Accident, Burn & Poison separately. The increase in fit over the previous model is because of individually fitting each group instead of providing an aggregate fit.

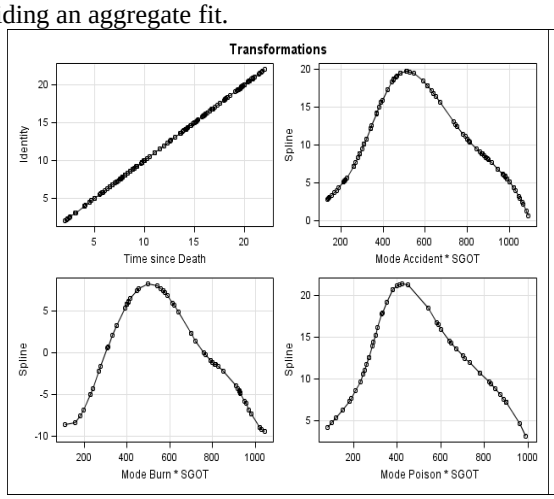

Figure 7

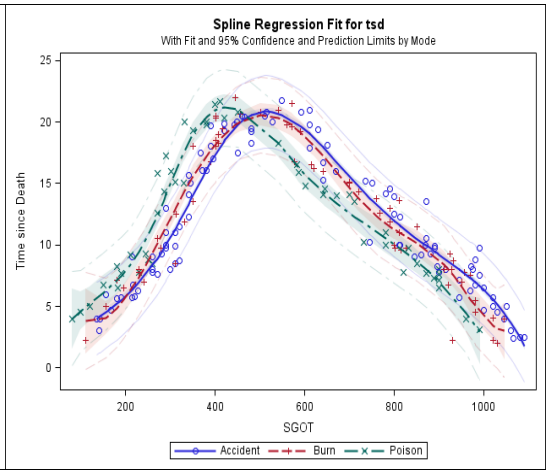

Figure 8

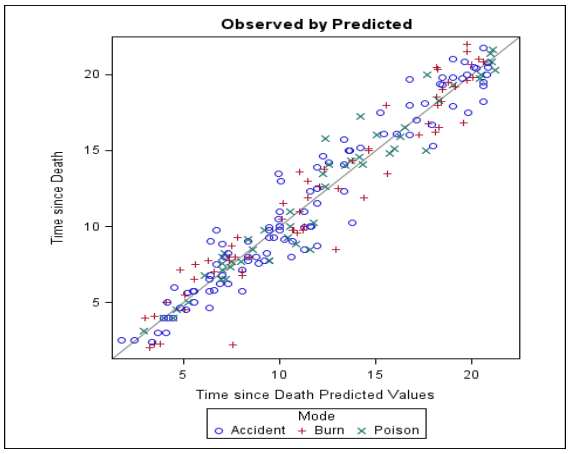

Figure 9

The residuals in the observed and predicted plot (Fig. 9) are much better when individual analysis is carried out for each mode.

IV. Fitting Through Penalized B-Splines

Another way to fit curves is using penalized B-spline transformations. In this method, the automatic selection of smoothing parameter is made. With penalized B-splines one can find a transformation that minimizes any of the following criteria: CV, GCV, AIC, AICC and SBC. All these criteria's are functions of smoothing parameter λ. Although, most of the criteria produce nearly identical results, however, for some problems the choice of criteria can have large effect. In order to define Penalized B-spline criteria, we use following notations:

n	number of observations
y	dependent variable
W	diagonal matrix of observation weights
w_i	weight for ith observation
B	B-spline basis for the independent variable
λ	nonnegative smoothing parameter
D	difference matrix, penalizes lack of smoothness
$H = B(B'WB + \lambda D'D)^{-1}B'$	hat matrix
h_{ii}	*i*th diagonal element of H
$\hat{y} = Hy$	penalized B-spline transformations of y
$SSE = \sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$	error sum of squares
$t = \sum_{i=1}^{n} w_i h_{ii}$	weighted trace of H
$\sum_{i=1}^{n} w_i \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}}\right)^2$	CV - cross validation criterion
$\sum_{i=1}^{n} w_i \left(\frac{y_i - \hat{y}_i}{n - t}\right)^2$	GCV - generalized cross validation
2t + nlog(SSE/n)	AIC - Akaike's information criterion
$1 + \log(SSE/n) + \frac{2(t+1)}{n-t-2}$	AICC - corrected AIC (default)
N log(SSE/n) + t log(n)	SBC - Schwarz's Bayesian criterion

The weighted trace of the hat matrix $t = \sum_{i=1}^{n} w_i h_{ii}$, provides an estimate of the number of parameters needed to find the transformation and is used to find degrees of freedom (df). By default, PBSPLINE fits a cubic spline with 100 evenly spaced knots, three evenly spaced exterior knots and a difference matrix of order 3. This procedure choose smoothing parameter λ, by minimizing one of the information or cross validation criteria, described in section IV. The range of λ varies from λ=0 and λ=1, 10, 100, 1000, 10000, 100000, 1000000. If it finds the range that includes the minimum, it stops otherwise it takes larger λ values.

4.1 Example

Following data is related to air pollution study of Hong Kong for the year 1998. In this study, pollutants PM_{10}, SO_2, NO_2 and O_3 were measured. The dataset contains 365 measurements of each day for these pollutants. We have used this data to fit a curve using penalized B-spline function. Our main aim is to choose a criteria using above definitions which fits well to the given data based on maximum value of R^2 and minimum value of smoothing parameter λ. We carried out the analysis for all the pollutants; however the results in this paper are presented only for that criterion for which R^2 is high and minimum value of smoothing parameter λ.

The dependent variables PM_{10}, SO_2, NO_2 and $O_3$are specified with IDENTITY transformation i.e. with no transformation. The independent variable 'days' is specified with a PBSPLINE transformation i.e. a penalized B-spline model is fit. Based on the above criterions the values of R^2 and λ are given Tables 5 and 6.

Table 5: Values of coefficient of Determination R^2

Criterion	**Values of R-Square**			
	PM10	SO2	NO2	O3
AICC	0.7938	0.1772	0.7409	0.7936
AIC	**0.8138**	**0.4804**	**0.7646**	**0.8090**
GCV	0.8057	0.4372	0.7542	0.8023
SBC	0.7123	0.0652	0.5361	0.6688
CV	0.8062	0.4445	0.7556	0.8015

Table 6: Values of Smoothing Parameter λ

Criterion	**Values of λ**			
	PM10	SO2	NO2	O3
AICC	0.0191	109.40	0.0318	0.0238
AIC	**0.0042**	**0.0045**	**0.0076**	**0.0076**
GCV	0.0088	0.0173	0.0157	0.0136
SBC	0.4948	1000000	74.9418	2.9229
CV	0.0085	0.0142	0.0144	0.0143

It is evident from Tables 5 & 6 that R^2 value is high and λ is low for AIC criterion. The penalized B-spline fit for PM_{10}, SO_2, NO_2 and O_3 with AIC criterion along with 95% confidence and prediction limits are plotted in Fig. 10.

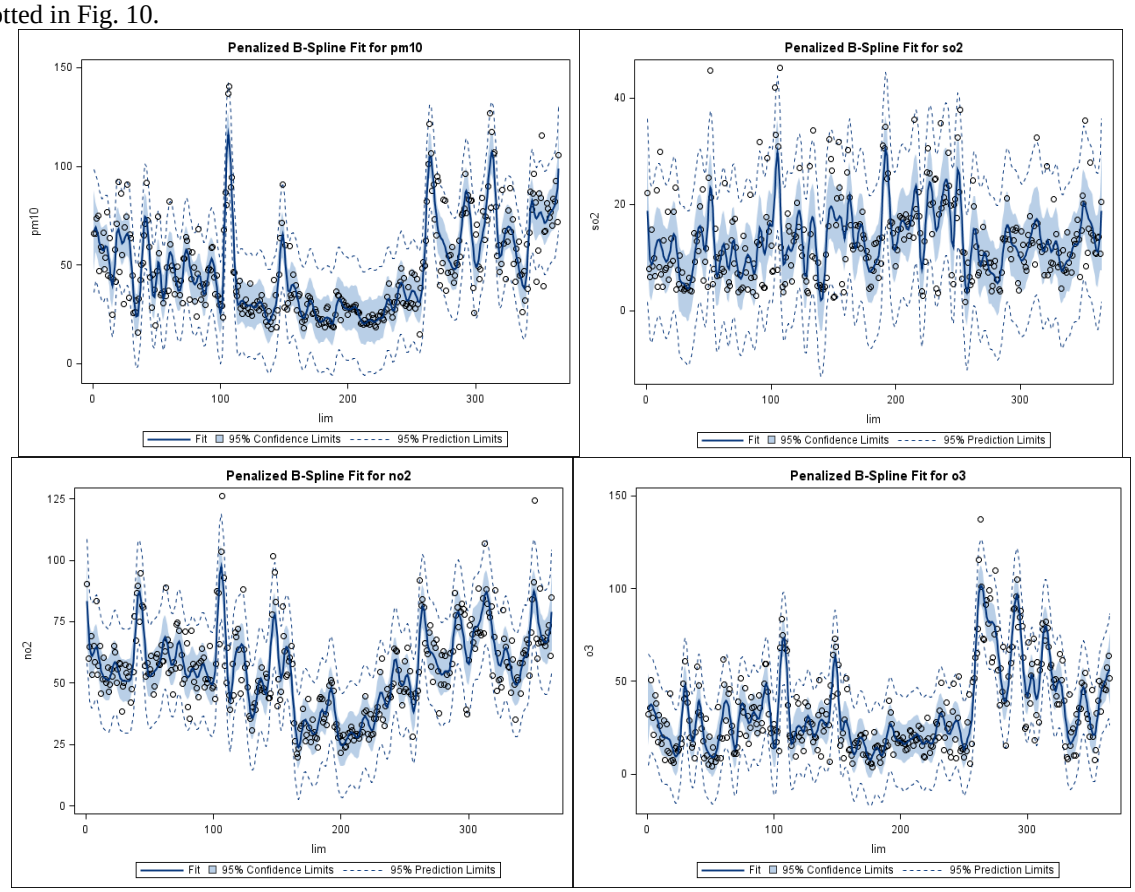

Figure10

The procedure choose a smoothing parameter λ= 0.0042 (for PM_{10}), 0.0045 (for SO_2), 0.0076 (for NO_2) and 0.0076 (for O_3). The maximum value of R^2 and smallest value of λ for AIC criterion results a better fit with close confidence and prediction intervals as compared to other criterions i.e. CV, GCV, AICC and SBC. Thus AIC criterion is even better than default AICC criterion for this data.

V. Conclusion

We discussed two methods of fitting a model i.e. TPSPLINE and TRANSREG. The TPSPLINE procedure is appropriate when there is no prior knowledge about the model or when the data cannot be represented by a model with fixed number of parameters. The procedure is demonstrated by taking a real life dataset with optimum choice of knots. TRANSREG procedure makes transformations of the dependent and independent variables and regression functions with smooth, spline or penalized B-splines can be fitted through this procedure. The fitting of the model through penalized B-splines by comparing AICC, AIC, CV, SBC and GCV criterion with practical example have been discussed. The model fitting through AIC criterion was found to be better than other criterions.

References

Journal Papers:

[1] Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math. 31, 377 - 403.

[2] De Leeuw, J., Young, F.W., and Takane, Y. (1976), "Additive Structure in Qualitative Data: An lternating Least Squares Approach with Optimal Scaling Features," Psychometrika, 41, 471–503.

[3] Eilers, P. H. C. and Marx, B. D. (1996), "Flexible Smoothing with B-splines and Penalities," Statistical Science, 11, 89 – 121, with discussion.

[4] Green, P.E., and Wind, Y. (1975), "New Way to Measure Consumers' Judgements," Harvard Business Review, July–August, 107–117.

[5] Kimeldorf, G. and G. Wahba (1970a), A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. Ann. Math. Statist. 41, 495 - 502.

[6] Kimeldorf, G. and G. Wahba (1970b).Spline functions and stochastic processes.Sankhya Ser. A 32, 173 - 180.

[7] Kimeldorf, G. and G. Wahba (1971). Some results on TchebycheÆan spline functions. J. Math. Anal.Applic. 33, 82 - 85.

[8] Reinsch, C. H. (1967), "Smoothing by spline Functions," Numerische Mahematik, 10, 177 – 183.

[9] Van Rijckeveorsel, J. (1982), "Canonical Analysis with B-Splines," in H. Caussinus, P. Ettinger, and R. Tomassone (ed.), COMPUSTAT 1982, Part I, Wein, PhysicaVerlag.

[10] Wahba, G. (1990). Spline Models for Observational Data, Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM.

[11] Winsberg, S., and Ramsay, J.O. (1980), "Monotonic Transformations to Additivity Using Splines," Biometrika, 67, 669–674.

Books:

[12] De Boor, C. (1978), A Practical Guide to Splines, New York: Springer Verlag.

[13] Green, P. J. and B. W. Silverman (1994), Nonparametric Regression and Generalized Linear Models. London: Chapman & Hall.

[14] Gu, C. (2002), Smoothing Spline ANOVA Models, New York: Springer-Verlag.