Paper 378-2008

Fitting Generalized Additive Models with the GAM Procedure in SAS 9.2

Weijie Cai, SAS Institute Inc., Cary NC

ABSTRACT

Generalized additive models are useful in finding predictor-response relationships in many kinds of data without using a specific model. They combine the ability to explore many nonparametric relationships simultaneously with the distributional flexibility of generalized linear models. The approach often brings to light nonlinear dependency structures in your data. This paper discusses an example of fitting generalized additive models with the GAM procedure, which provides multiple types of smoothers with automatic selection of smoothing parameters. This paper uses the ODS Statistical Graphics to produce plots of integrated additive and smoothing components.

INTRODUCTION

PROC GAM is a powerful tool for nonparametric regression modeling. PROC GAM provides great flexibility in modeling predictor-response relationships, as do other nonparametric SAS/STAT procedures such as the TPSPLINE and the LOESS procedures. However, unlike PROC TPSPLINE and PROC LOESS, PROC GAM scales well with the increasing dimensionality and it yields interpretable models. You often benefit from this exploratory modeling with PROC GAM because it can inspire parsimonious parametric models.

Hastie and Tibshirani (1986, 1990) proposed the underlying methodology for the generalized additive models. Their models combine the ability to model data from distributions in the exponential family as generalized linear models (Nelder and Wedderburn 1972) with the ability to approximate multivariate regression functions by using additive models (Stone 1985).

Additive models assume nonparametric smoothing splines for predictors in regression models. Generalized linear models assume the dependency of the dependent variable on additive predictors through a monotonic nonlinear link function specified by a distribution member in the exponential family. By combining these two assumptions, generalized additive models can be used in a wide range of modeling scenarios.

Features of PROC GAM include:

- support of univariate smoothing splines, local regression smoothers, and bivariate thin-plate smoothing splines
- the ability to fit both nonparametric and semiparametric models
- support of multiple SCORE statements
- support of user-specified smoothing parameters or automatic smoothing parameter selection by optimizing the GCV (generalized cross validation) criterion

In addition to these features, PROC GAM in SAS 9.2 adds the following new functionalities:

- fast approximate analysis of deviance
- graphical display support via the ODS Statistical Graphics system
- character response support and sort order options for binary models
- sort order options for categorical variables

The next section describes the methodology and the fitting procedure that underlie generalized additive models.

COMPUTATION

Let Y be a response random variable and $X_1, \ldots, X_p$ be p predictor variables. A standard linear regression model assumes

$$E(Y) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

Given a sample of observations, the coefficients $\beta_i, i = 1, \ldots, p$ can be estimated by the least squares method, assuming that Y follows a normal distribution.

Generalized linear models extend the distribution of the response variable of the linear model into the exponential family. Let $E(Y) = \mu$ and $\eta = g(\mu)$, where g is a monotonic link function. The response-predictor relationship is defined by

$$\eta = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

The exponential family distribution has the following form:

$$f_Y(y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right)$$

where a, b, and c are arbitrary functions and ϕ is a scale parameter. Then the coefficients $\beta_i, i = 1, \ldots, p$ are estimated by a form of the iterative reweighted least squares method called the Fisher scoring procedure.

Additive models extend the parametric form of predictors in the linear model to nonparametric forms. Assuming that Y is normally distributed, an additive model is defined as

$$E(Y) = s_0 + \sum_{i=1}^{p} s_i(X_i)$$

If the nonparametric smoothers $s_i, i = 1, \ldots, p$ are smoothing splines, s_i can be estimated by the penalized least squares method.

Combining additive models and generalized linear models, generalized additive models have the following form:

$$\eta = s_0 + \sum_{i=1}^{p} s_i(X_i)$$

where the response variable has a probability density in the exponential family. Under the canonical link used by PROC GAM, $\eta = \theta$. Backfitting and local scoring described in the following two subsections form a doubly iterative method to estimate the smoothers $s_i, i = 1, \ldots, p$. Each smoother s_i is controlled by a single smoothing parameter. You can specify a desired smoothing parameter or choose it automatically by the generalized cross validation method (Wahba 1990).

THE BACKFITTING ALGORITHM

The backfitting algorithm is a general algorithm that can fit an additive model with any regression-type fitting mechanisms. It can be used with different smoothers such as smoothing splines and local regression smoothers. The basic idea is to find the jth smoother of the $(k+1)$th iteration by smoothing the partial residual defined by

$$R_j^{(k)} = Y - s_0 - \sum_{i<j} s_i^{(k+1)}(X_i) - \sum_{i>j} s_i^{(k)}(X_i)$$

The additivity assumption plays an important role here: the effects of predictors can be separated into additive terms. An iteration is done by cycling through all p smoothers in turn. When the change in each effect is sufficiently small, the iterations stop.

It can be shown that for many smoothers (including linear regression, univariate and bivariate splines, and combinations of these), the sum of squared residuals of the additive model never increases at any step. This implies that the algorithm always converges (Hastie and Tibshirani 1986). However, for distributions other than Gaussian, numerical instabilities can cause convergence problems.

THE LOCAL SCORING ALGORITHM

The local scoring algorithm is also an iterative algorithm. PROC GAM fits generalized additive models by using a modified form of adjusted dependent variable regression, as described for generalized linear models in McCullagh and Nelder (1989), with the additive predictor taking the role of the linear predictor. The Fisher scoring procedure is thus replaced by the local scoring algorithm because the predictions for the adjusted dependent variable are localized by nonparametric smoothers.

The estimating procedure for generalized additive models consists of two loops. Inside each step of the local scoring algorithm (outer loop), a weighted backfitting algorithm (inner loop) is used until the convergence criterion for backfitting is satisfied. Then, based on the estimates from this weighted backfitting algorithm, a new set of weights is calculated, and the next iteration of the scoring algorithm starts. The scoring algorithm stops when the convergence criterion is satisfied or the change in deviance of the estimates is below a threshold.

SELECTION OF SMOOTHING PARAMETERS

Each smoother specified in PROC GAM has a single smoothing parameter. Cross validation can be used to choose the smoothing parameter. By leaving out one observation at a time, a model is built based on the remaining observations. A residual value is thus obtained for the omitted data point. After cycling through all data points, a sample of residuals is obtained. The average squared residual is used as an estimate for the prediction error. It is often computationally expensive to evaluate the cross validation errors. Proposed by Craven and Wahba (1979), the generalized cross validation method provides a convenient approximation to the leave-one-out cross validation that reduces computation cost. It can be viewed as a weighted cross validation. Minimizing the generalized cross validation criterion often yields a similar smoothing parameter to that obtained by the leave-one-out cross validation. You can specify the METHOD=GCV option in the MODEL statement in order to request that the generalized cross validation method be used to choose the smoothing parameters.

EXAMPLE

The Pima Indians Diabetes data set is obtained from the UCI Machine Learning Repository (Asuncion and Newman 2007). This data set is extracted from a larger database originally owned by the National Institute of Diabetes and Digestive and Kidney Diseases. All records here are for female patients who are at least 21 years old, of Pima Indian heritage, and live near Phoenix, Arizona, USA. The purpose of the study is to investigate the relationship between the diabetes diagnostic result and a list of variables that represent physiological measurements and medical attributes. The data set in the UCI repository contains 768 observations and 9 variables with no missing values reported. However, as some researchers point out, there are a number of impossible values, such as 0 body mass index and 0 plasma glucose. Furthermore, one attribute (2-hour serum insulin) contains almost 50% impossible values. To keep the sample size reasonably large, this attribute is removed from analysis. There are 236 observations that have at least one impossible value of glucose, blood pressure, triceps skin thickness, and body mass index. A reduced data set with 532 records is thus obtained by removing these observations. This reduced data set is used for statistical analysis in this paper.

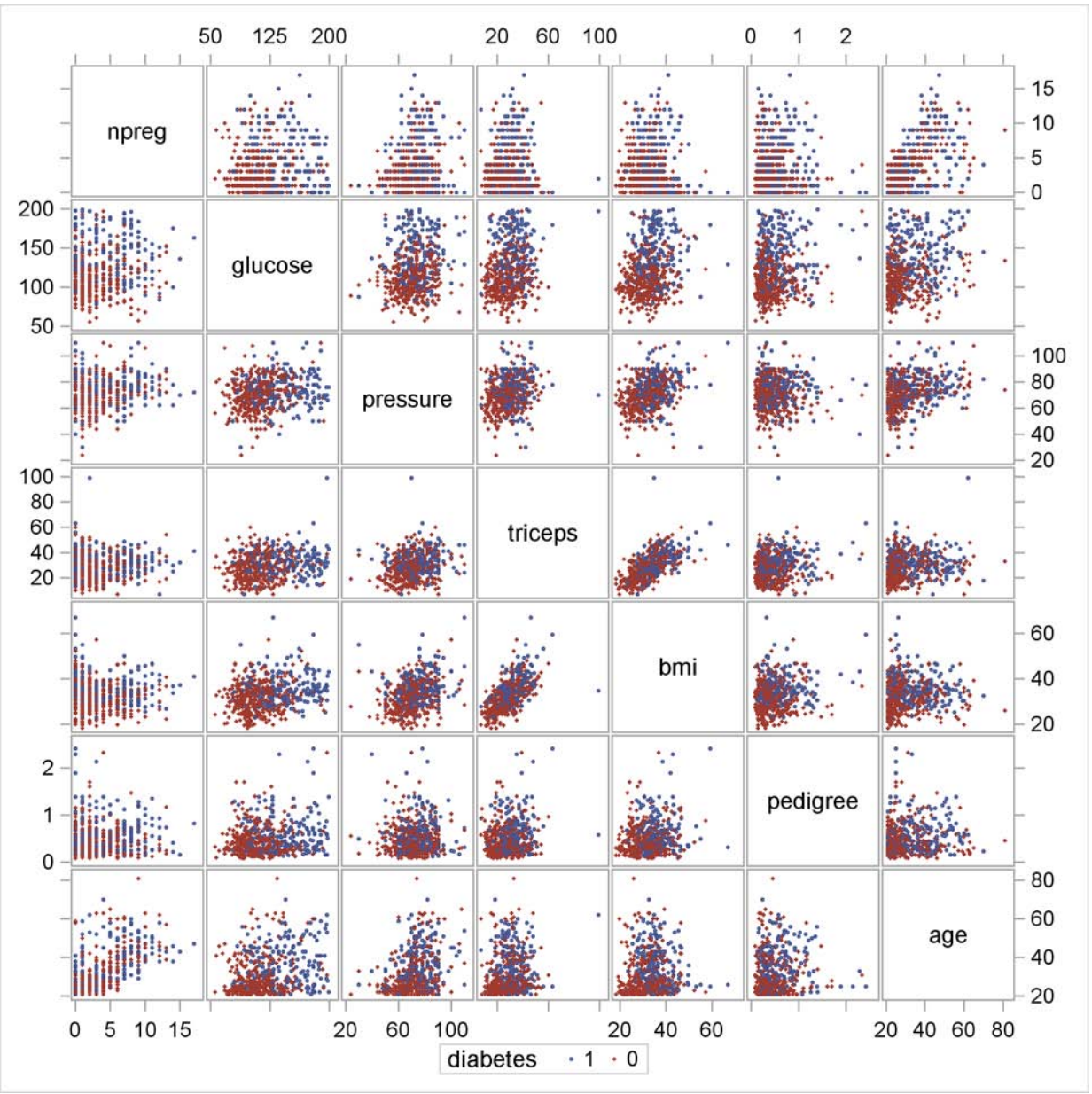
There are eight variables, including the binary response variable, in this data set; the variable meanings are listed in the following table:

Variable Name	Variable Explanation
npreg	number of times pregnant
glucose	2-hour plasma glucose concentration in an oral glucose tolerance test
pressure	diastolic blood pressure (mm Hg)
triceps	triceps skin fold thickness (mm)
bmi	body mass index (weight in kg/(height in m)^2)
pedigree	diabetes pedigree function
age	age in years
diabetes	classification variable, 0 for testing negative, 1 for testing positive

The following statements produce a scatter plot matrix for all the predictors in the Pima Indians Diabetes data set. Blue circles and red plus symbols distinguish patients with positive and negative diabetes tests, respectively.

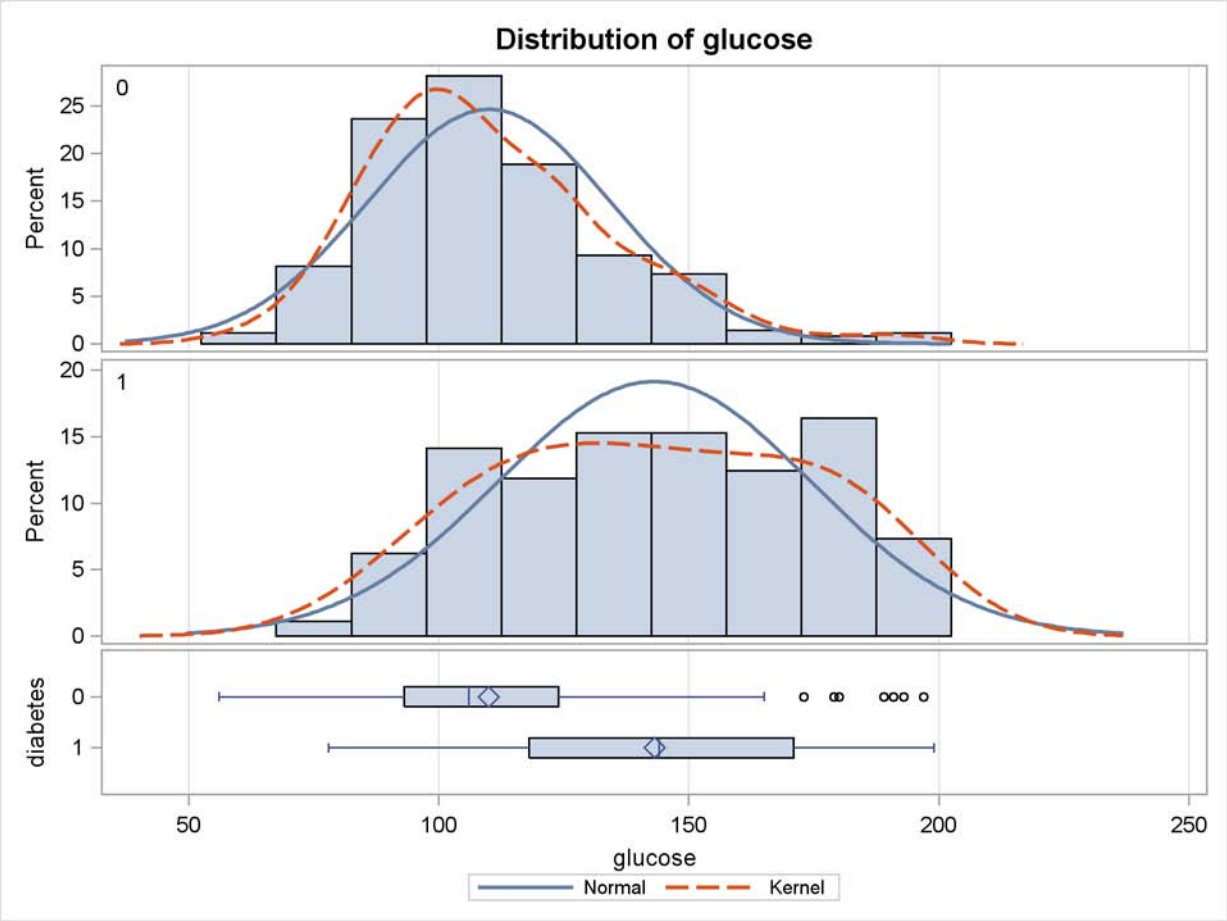
```
ods graphics on;
proc sgscatter data=pima;
   matrix npreg glucose pressure triceps bmi pedigree age /
      group=diabetes markerattrs=(size=2);
run;
```

Figure 1 Scatter Plot Matrix for Predictors in the Pima Indians Diabetes Data Set



You can see that bmi and triceps are positively associated. The variable glucose exhibits an approximate classification of diabetes, which you can see more clearly in the comparative histograms and box plots produced by PROC TTEST (Figure 2). The histograms with overlaid density estimates show that the distribution of glucose for the patients with and without diabetes is different. The box plots indicate that those patients who are not diagnosed with diabetes tend to have lower glucose levels than those diagnosed with diabetes.

Figure 2 Histogram for Glucose



As an initial step in analyzing the predictor-response relationship, you might try fitting a linear logistic regression model. In some cases, an examination of partial residual plots that you obtain might suggest that additional nonlinear relationships need to be modeled. Without prior knowledge, the process of determining the appropriate variable transformations is time consuming and subject to your perception. In such cases, you can use PROC GAM as a tool to aid in the identification of appropriate nonlinear transformations. This example explores this approach.

The first step is to fit a GAM model with all seven predictors included as follows:

```
proc gam data=pima;
   model diabetes(event=last)=spline(npreg)     spline(glucose)
                              spline(pressure)  spline(triceps)
                              spline(bmi)       spline(pedigree)
                              spline(age) /
                                 dist=binomial anodev=norefit;
   output out=pimaout p;
run;
```

The preceding statements request that PROC GAM fit a logistic additive model with a binary dependent variable diabetes against the seven continuous predictors. By default, the smoothing parameter for each B-spline term is chosen to yield four degrees of freedom. Of the four degrees of freedom, one is taken up by the linear portion of the fit and three remain for the nonlinear spline portion.

The function of the three additional options used in the MODEL statements is as follows:

- **`event=last`** This response option chooses diabetes=1 (positive) as the event, so that the probability of a positive diabetes test is modeled. If this option is not specified, PROC GAM designates the first ordered response category (diabetes=0 (negative) in this case) as the event to be modeled.
- **`dist=binomial`** This option specifies a logistic additive model. You can specify other distributions in the exponential family to model different types of data.
- **`anodev=norefit`** This option requests that PROC GAM perform approximate analysis of deviance for smoothing effects. With the default option **`anodev=refit`**, PROC GAM performs analysis of deviance by fitting seven additional GAM fits. In each of these fits, a nested GAM model is fit by omitting one smoothing term from the model. By contrast, the **`anodev=norefit`** option keeps the other six nonparametric terms fixed and does a weighted least squares fit for the parametric portion of the model only. This approximate analysis of deviance method is much faster for models with many smoothing variables, yet it usually yields similar conclusions to the more computationally demanding full analysis of deviance method.

The OUTPUT statement saves in the SAS data set pimaout the predicted probabilities of a positive diabetes test and nonparametric predictions for each individual predictors.

The output from PROC GAM is listed in Figure 3 and Figure 4. Figure 3 show three tables: the first table summarizes the input data set and the distributional family used in the GAM model; the second table summarizes the frequency counts for the binary response variable; and the third table provides a summary for the backfitting and local scoring algorithms.

Figure 3 Summary Statistics

```
                          The GAM Procedure
                      Dependent Variable: diabetes
 Smoothing Model Component(s): spline(npreg) spline(glucose) spline(pressure)
           spline(triceps) spline(bmi) spline(pedigree) spline(age)

                        Summary of Input Data Set

                  Number of Observations                 532
                  Number of Missing Observations           0
                  Distribution                      Binomial
                  Link Function                        Logit

                              Response Profile

                    Ordered                          Total
                      Value        diabetes       Frequency

                          1               0             355
                          2               1             177

NOTE: PROC GAM is modeling the probability that diabetes=1. One way to change
      this to model the probability that diabetes=0 is to specify the response
      variable option EVENT='0'.


                  Iteration Summary and Fit Statistics

         Number of local scoring iterations                        7
         Local scoring convergence criterion          5.581043E-11
         Final Number of Backfitting Iterations                    1
         Final Backfitting Criterion                  1.188347E-10
         The Deviance of the Final Estimate           420.76348675
```

The second part of the output (Figure 4) provides analytical information about the fitted model, including parameter estimates for the linear portion of the model, fit summary for smoothing components, and the "Approximate Analysis of Deviance" table. The chi-square values in the "Approximate Analysis of Deviance" table indicate that the three smoothing terms of glucose, pressure, and triceps are not statistically significant, npreg and pedigree are close to significant, bmi is significant, and age is highly significant.

Figure 4 Model Fit Statistics

Regression Model Analysis
Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	-8.80351	1.04780	-8.40	<.0001
Linear(npreg)	0.05983	0.04502	1.33	0.1845
Linear(glucose)	0.03603	0.00440	8.19	<.0001
Linear(pressure)	-0.01331	0.01048	-1.27	0.2046
Linear(triceps)	0.00288	0.01513	0.19	0.8492
Linear(bmi)	0.07014	0.02418	2.90	0.0039
Linear(pedigree)	1.04715	0.33770	3.10	0.0020
Linear(age)	0.04254	0.01517	2.80	0.0052

Smoothing Model Analysis
Fit Summary for Smoothing Components

Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline(npreg)	0.994765	3.000000	13.931251	17
Spline(glucose)	0.999999	3.000000	95.671627	126
Spline(pressure)	0.999531	3.000000	4.419230	42
Spline(triceps)	0.998848	3.000000	2.781179	50
Spline(bmi)	1.000000	3.000000	134.570168	222
Spline(pedigree)	1.000000	3.000000	97.613479	413
Spline(age)	0.999284	3.000000	5.151763	46

Smoothing Model Analysis
Approximate Analysis of Deviance

Source	DF	Chi-Square	Pr > ChiSq
Spline(npreg)	3.00000	7.3079	0.0627
Spline(glucose)	3.00000	3.2774	0.3508
Spline(pressure)	3.00000	1.1470	0.7657
Spline(triceps)	3.00000	1.5148	0.6789
Spline(bmi)	3.00000	9.6516	0.0218
Spline(pedigree)	3.00000	7.3015	0.0629
Spline(age)	3.00000	18.5773	0.0003

Since some spline terms with fixed four degrees of freedom (DF) do not appear significant, the fixed four degrees of freedom might not be suitable for all predictors. Without any prior knowledge, it is hard to specify appropriate values for the degrees of freedom. An alternative is to use the **method=gcv** option to choose smoothing parameters, as shown in the following statements:

```
proc gam data=pima plots=components(commonaxes);
   model diabetes(event=last)=spline(npreg)     spline(glucose)
                              spline(pressure)  spline(triceps)
                              spline(bmi)       spline(pedigree)
                              spline(age)/
                                 dist=binomial method=gcv
                                 anodev=norefit epsscore=1e-6;
   output out=pimaout2 p;
run;
```

Provided that you have enabled ODS Graphics, the option **plots=components(commonaxes)** requests that panels of smoothing component plots be produced with a common vertical axis. Using a common vertical axis helps you visually compare the importance of fitted terms in additive models. In this example, relaxing the convergence criterion by specifying **epsscore=1e-6** yields good results with a moderate amount of computation.

Figure 5 and Figure 6 show two panels of smoothing component plots produced by the second GAM model with the `method=gcv` option specified.

Figure 5 Smoothing Component Plots

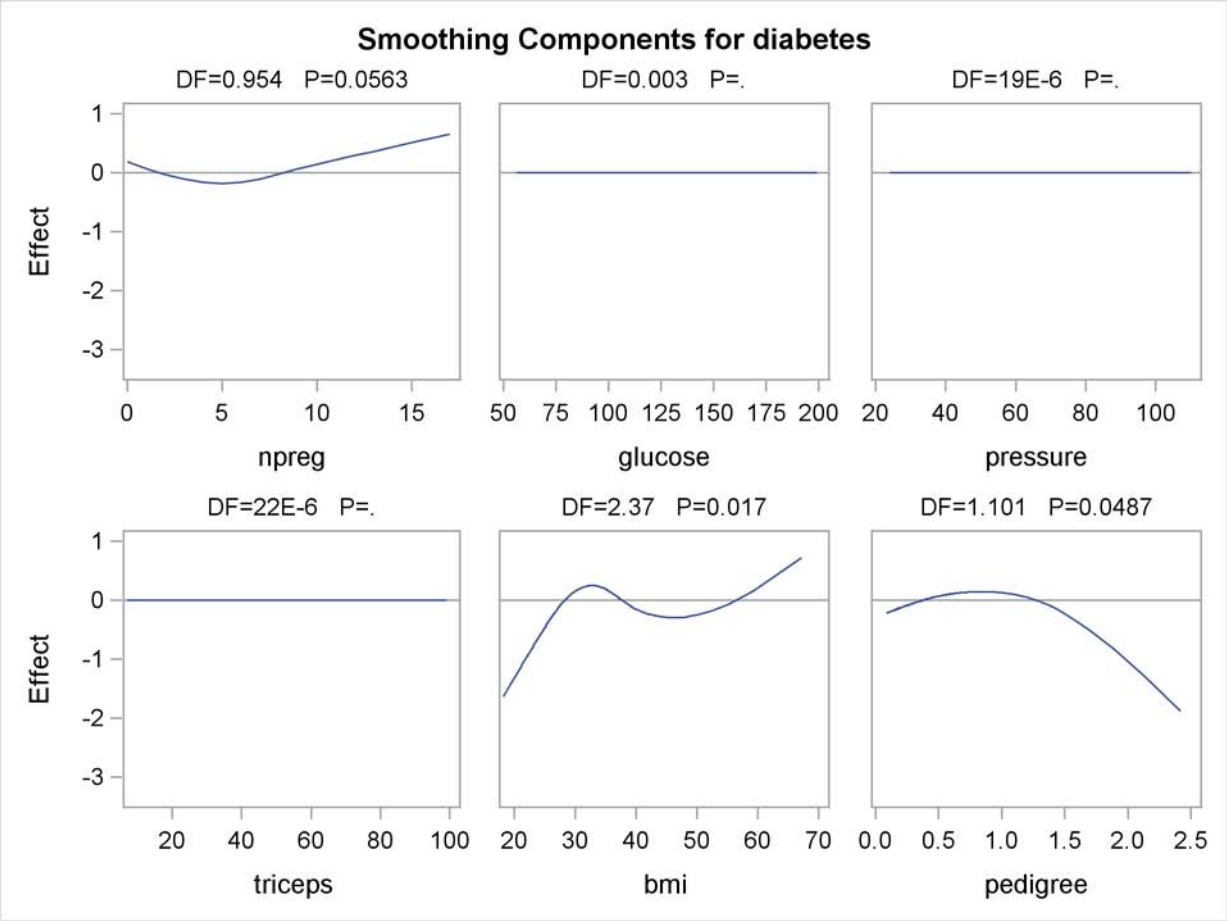
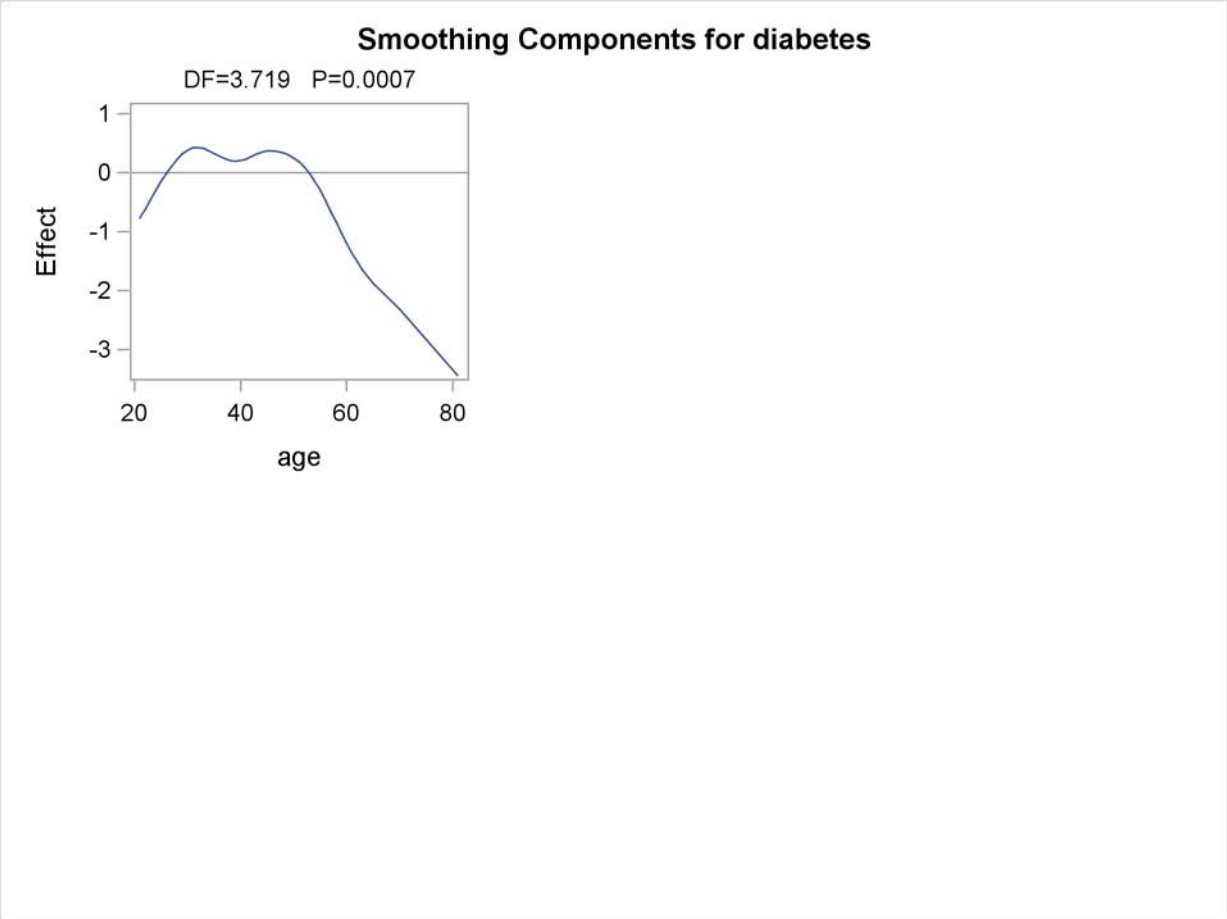


Figure 6 Smoothing Component Plots



The use of a common vertical axis in Figure 5 and Figure 6 enables you to see that glucose, pressure, and triceps do not exhibit any nonlinear trends. Although you might view local structures of the nonparametric predictions for the three variables with the **commonaxes** option turned off, the small DF values indicate that the nonparametric predictions are almost linear and close to zero, which is also suggested by the "Approximate Analysis of Deviance" table (Figure 7).

Figure 7 Analysis of Deviance Table

Smoothing Model Analysis
Approximate Analysis of Deviance

Source	DF	Chi-Square	Pr > ChiSq
Spline(npreg)	0.95390	3.5307	0.0563
Spline(glucose)	0.00346	0.0011	.
Spline(pressure)	0.00001912	0.0000	.
Spline(triceps)	0.00002174	0.0000	.
Spline(bmi)	2.37042	8.9361	0.0170
Spline(pedigree)	1.10135	4.1322	0.0487
Spline(age)	3.71942	18.5095	0.0007

For the three variables (glucose, pressure, and triceps), the chi-square test statistics are very close to zero and the p-values are all missing because their corresponding DF values are too small. This numerical result and the visualization in the smoothing component plot suggest that no obvious nonlinearities remain for the three variables in this logistic additive model and that most nonparametric trends are explained by the four spline terms.

The preceding analysis suggests trying a simpler semiparametric model, as in the following statements:

```
proc gam data=pima;
   model diabetes(event=last)=spline(npreg)    param(glucose)
                              param(pressure)  param(triceps)
                              spline(bmi)      spline(pedigree)
                              spline(age)/
                                 dist=binomial method=gcv
                                 anodev=norefit epsscore=1e-6;
   output out=pimaout3 p;
run;
```

This GAM model contains three strictly linear terms specified by `param()`, in addition to four univariate smoothing spline terms whose degrees of freedom are chosen by the GCV method. The parameter estimates table in the model output is shown in Figure 8.

Figure 8 Parameter Estimates for Linear Portion

The GAM Procedure
Dependent Variable: diabetes
Regression Model Component(s): glucose pressure triceps
Smoothing Model Component(s): spline(npreg) spline(bmi) spline(pedigree) spline(age)

Regression Model Analysis
Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	-8.99184	1.03851	-8.66	<.0001
glucose	0.03570	0.00429	8.32	<.0001
pressure	-0.01201	0.01042	-1.15	0.2495
triceps	0.00256	0.01467	0.17	0.8618
Linear(npreg)	0.05751	0.04450	1.29	0.1967
Linear(bmi)	0.07302	0.02387	3.06	0.0023
Linear(pedigree)	1.15861	0.34884	3.32	0.0010
Linear(age)	0.04243	0.01497	2.83	0.0048

The linear term of triceps is not significant because it has a strong correlation with bmi; the correlation can be observed in the scatter plot matrix shown in Figure 1. The linear term of pressure is not significant either. This suggests that you further reduce the additive model by removing these two terms sequentially and observing no significant changes in tests for other terms. After analyzing the numerical results and visualizing the smoothing components, the variable npreg is also removed from the model. This leads to a simple four-variable logistic additive model, specified in the following statements:

```
proc gam data=pima plots=components(clm);
   model diabetes(event=last)=param(glucose)
                              spline(bmi) spline(pedigree)
                              spline(age)/
                                 dist=binomial method=gcv
                                 anodev=refit epsscore=1e-6;
   output out=pimaout4 p;
run;
```

The analytical information about the final additive model is listed in Figure 9.

Figure 9 Model Fit Statistics

The GAM Procedure
Dependent Variable: diabetes
Regression Model Component(s): glucose
Smoothing Model Component(s): spline(bmi) spline(pedigree) spline(age)

Regression Model Analysis
Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	-9.65852	0.95566	-10.11	<.0001
glucose	0.03481	0.00420	8.29	<.0001
Linear(bmi)	0.07189	0.01869	3.85	0.0001
Linear(pedigree)	1.15369	0.34582	3.34	0.0009
Linear(age)	0.04990	0.01134	4.40	<.0001

Smoothing Model Analysis
Fit Summary for Smoothing Components

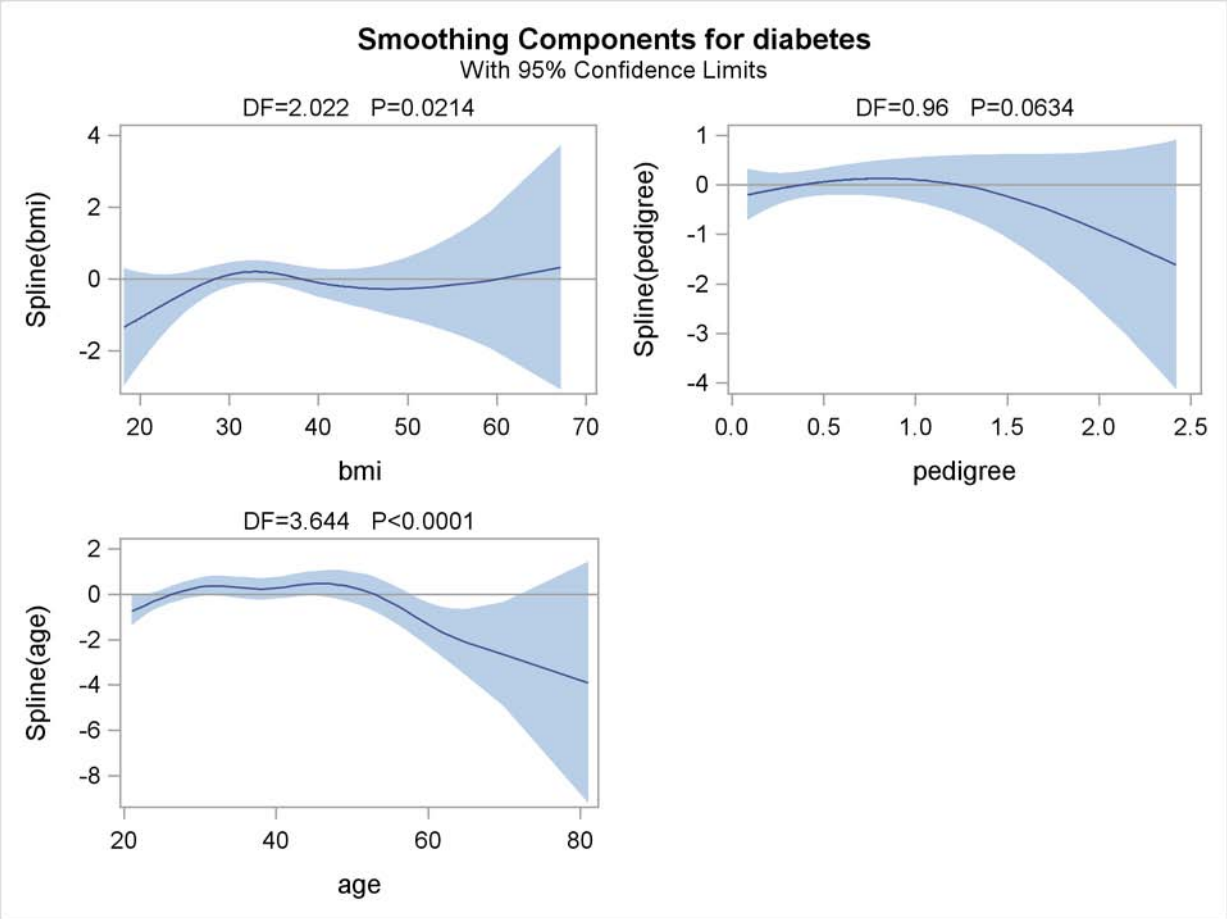
Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline(bmi)	1.000000	2.021787	45.792025	222
Spline(pedigree)	1.000000	0.959616	35.840708	413
Spline(age)	0.998614	3.643917	4.541968	46

Smoothing Model Analysis
Analysis of Deviance

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(bmi)	2.02179	7.734990	7.7350	0.0214
Spline(pedigree)	0.95962	3.350877	3.3509	0.0634
Spline(age)	3.64392	23.962670	23.9627	<.0001

The "Parameter Estimates" table in Figure 9 shows that all linear terms are highly significant. The "Analysis of Deviance" table in Figure 9 indicates that there are significant nonlinear contributions from the three variables bmi, pedigree, and age. You can clearly see these nonlinearities in the smoothing component plots (Figure 10) with the 95% Bayesian confidence limits specified by the option **`plots=components(clm)`**. You can even try to explore interaction patterns by fitting new GAM models with **`spline2`** terms—for example, **`spline2(bmi,pedigree)`**. It turns out there are no strong interactions between variables within this data. You can also visualize the combined effect of the linear and nonparametric contributions for each spline term by specifying the **`plots=components(additive)`** option.

Figure 10 Smoothing Component Plots



The general trends in the smoothing component plots for pedigree and age suggest possible quadratic dependence on these variables, and the trend in bmi resembles a cubic polynomial. Both types of transformations make sense physically for these variables and can provide insights into the problem. The next step of the analysis is to assess these transformations in a parametric model by using PROC GENMOD. The following statements fit a logistic regression model with a cubic polynomial of bmi and quadratic polynomials of pedigree and age:

```
proc genmod data=pima descending;
   model diabetes=glucose bmi|bmi|bmi pedigree|pedigree age|age /
                    link=logit dist=binomial;
run;
```

Figure 11 Linear Model by PROC GENMOD

The GENMOD Procedure

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square
Intercept	1	-39.8950	10.9163	-61.2906	-18.4995	13.36
glucose	1	0.0351	0.0043	0.0267	0.0436	66.79
bmi	1	2.0302	0.8843	0.2969	3.7634	5.27
bmi*bmi	1	-0.0500	0.0237	-0.0966	-0.0035	4.44
bmi*bmi*bmi	1	0.0004	0.0002	0.0000	0.0008	3.90
pedigree	1	3.0901	0.9335	1.2604	4.9197	10.96
pedigree*pedigree	1	-1.1782	0.5185	-2.1945	-0.1620	5.16
age	1	0.3339	0.0740	0.1889	0.4789	20.37
age*age	1	-0.0036	0.0009	-0.0055	-0.0018	14.77
Scale	0	1.0000	0.0000	1.0000	1.0000	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	Pr > ChiSq
Intercept	0.0003
glucose	<.0001
bmi	0.0217
bmi*bmi	0.0350
bmi*bmi*bmi	0.0483
pedigree	0.0009
pedigree*pedigree	0.0231
age	<.0001
age*age	0.0001
Scale	

NOTE: The scale parameter was held fixed.

As you can see from the partial output of PROC GENMOD in Figure 11, all of the higher order terms such as `bmi*bmi` together with the linear terms such as `glucose` are statistically significant. This is consistent with the nonlinearities discovered by using PROC GAM.

Introducing these higher order terms also improves the model classification performance. Lim, Loh, and Shih (2000) conducted an extensive comparison study on 33 classification algorithms on various data sets including the reduced Pima Indian Diabetes data set used in this example. Their study reveals that the lowest error rate, estimated by 10-fold cross validation, that can be achieved for this data set is 0.22.

You can evaluate the obtained nonparametric and parametric models for the Pima Indians Diabetes data set by using a similar approach. To estimate the error rate, the same 10-fold cross validation scheme as in Lim, Loh, and Shih (2000) is used:

1. Split the full data into two sets: positive (diabetes=0) and negative (diabetes=1). Randomly divide both data sets into 10 disjoint subsets. Form 10 subsets by joining one positive subset and one negative subset.
2. Construct a classifier on the rest of the data by omitting each subset sequentially. Then use the classifier to classify the observations in the omitted subset. Dichotomize the predicted probabilities for the subset with the threshold of 0.5. Compute the misclassification rate for each subset.
3. Use the overall mean misclassification rate as the estimate for the error rate. Note that there will be missing predicted probabilities if new observations are outside the ranges of smoothing variables. Remove these missing predictions from averaging.

The following table summarizes the error rate estimates for the three models studied in the preceding analysis.

Model	Error Rate
glucose, s(bmi), s(pedigree), s(age)	0.212
glucose, bmi, pedigree, age	0.214
glucose, bmi, bmi^2, bmi^3, pedigree, pedigree^2, age, age^2	0.201

The error rates, estimated by the 10-fold cross validation scheme, show that the three reduced models are competitive with those by the best classification algorithms studied by Lim, Loh, and Shih (2000).

In this example, you can see that PROC GAM not only detects nonlinearities and indicates possible transformations, but also confirms the linearity of some variables by analysis of deviance and smoothing component visualizations. The overall exploratory study yields one simple nonparametric model and motivates two simple yet competitive parametric models.

CONCLUSION

In this paper, the underlying methodology of PROC GAM is introduced. PROC GAM provides power and flexibility in modeling a wide range of practical data. Its nonparametric nature does not require much prior information and can also shed light on underlying parametric relationships. The Pima Indians Diabetes data set is used to illustrate the usage of PROC GAM. Simple yet competitive models are derived.

REFERENCES

Asuncion, A. and Newman, D. J. (2007), "UCI Machine Learning Repository," http://www.ics.uci.edu/~mlearn/MLRepository.html.

Craven, P. and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions," *Numerical Mathematics*, 31, 377–403.

Hastie, T. and Tibshirani, R. (1986), "Generalized Additive Models," *Statistical Science*, 3, 297–318.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, New York: Chapman & Hall.

Lim, T. S., Loh, W. Y., and Shih, Y. S. (2000), "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms," *Machine Learning*, 40(3), 203–229.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Second Edition, London: Chapman & Hall.

Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *Annals of Statistics*, 13, 689–705.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: Society for Industrial and Applied Mathematics.

CONTACT INFORMATION

Weijie Cai,
SAS Institute Inc.
500 SAS Campus Drive,
Cary, NC 27513
Work Phone: (919) 531-0359
E-mail: Weijie.Cai@sas.com
Web: http://www.sas.com