The Korean Communications
in Statistics Vol. 7, No. 2, 2000
pp. 393-401

A Procedure for Fitting Nonadditive Models[1]

Han Son Seo[2]

Abstract

Many graphical methods have been suggested for obtaining an impression of a curvature in regression problem in which some covariates enter nonlinearly. However when true model does not belong to the class of additive models, graphical methods may contain a serious bias. A method is suggested which can avoid such bias in the fitting of nonaddive models.

Keywords : Inverse response plot, CERES plot, Nonadditive model

1. Introduction

Standard linear regression is based on some assumptions such as consistency of variance and linearity of the regression function. Linear relationship between explanatory variables and response variable, however, is not so sure in many problems, Thus the following model is considered:

$$Y = \beta_0 + \beta_1^T X_1 + g(X_2) + \varepsilon \qquad (1.1)$$

where Y is the response, X_i is $p_i \times 1$ vector of covariates $i = 1, 2$ and g is unknown function with $E(\varepsilon \mid x_1, x_2) = 0$. For the specification of the curvature g in the model (1.1), many graphical methods are suggested including added variable plot(Chamber et al., 1983, p.272), partial residual plot (Larsen and McLeary, 1972; Weisberg, 1985), augmented partial residual plot (Mallows, 1986) and CERES plot (Cook 1993). Berk and Booth (1995) compared nine plots including these plots concerning with the ability to reveal a curve.

Partial residual plot is described as plot of $e + \hat{\phi} X_2$ versus X_2 where $\hat{\phi}$ and e are the LSE and the OLS residual respectively from linear regression model. This plot should work well if the conditional expectations $E(X_1 \mid X_2)$ are all linear. Augmented partial residual plot is a plot of $e + \widehat{\phi}_1 X_2 + \widehat{\phi}_2 X_2^2$ versus X_2 from the quadratic regression model

[1] This paper was supported by Konkuk University in 1999.

[2] Associate Professor, Department of Applied Statistics, Konkuk University, Seoul 143-701
E-mail : hsseo@kkucc.konkuk.ac.kr

$$Y = \rho_0 + X_1\rho_1 + \phi_1 X_2 + \phi_2 X_2^2 + \varepsilon.$$

Augmented partial residual plot can depict g better than partial residual plot if $\phi_1 X_2 + \phi_2 X_2^2$ provides a better approximation of $g(X_2)$ that provided by ϕX_2 alone. CERES plots include partial residual plots and augmented partial residual plots as special cases. CERES plots are useful for obtaining the visualization of g when the conditional expectations $E(X_1 \mid X_2)$ are neither linear nor quadratic. All these plots were developed under the assumption that X_1 enters the model linearly. Failure occurs when the model (1.1) is misspecified, when the true model is

$$Y = \beta_0 + f(X_1, X_2) + \varepsilon \qquad (1.2)$$

where f is an unknown function. If $f(X_1, X_2)$ belongs to the class of generalized additive models (Hastie and Tibshirani, 1990 P86) then backfitting algorithm can be used to fit it.

In this article a method for fitting the model of (1.2) is proposed. We suggest to transform a response variable and to use CERES plot for the visualizing the curvature. We assume that $p_2 = 1$ for the simplicity of the problem. Section 2 discusses CERES plots and inverse response plot. Section 3 proposes a new procedure for fitting nonadditive models and applies it to the real and the artificial data. In Section 4 concluding remarks are given.

2. CERES plots and inverse response plot

2.1 Visualizing the curvature via CERES plots

Two dimensional plot is denoted as $\{h, v\}$ with the understanding that h is assigned to the horizontal axis and v is assigned to the vertical axis. We assume model (1.1) and are interested in uncovering the curvature in the regression. Consider the model

$$Y = a_0 + a_1^T X_1 + b^T m(X_2) + error \qquad (2.1)$$

where $m(X_2) = E(X_1 \mid X_2)$, and construct a plot $\{X_{2i},\ e_i + \widehat{b^T} m(X_{2i})\}$ $i = 1, \cdots, n$. $E(X_1 \mid X_2)$ can be obtained parametrically by setting $E(X_1 \mid X_2) = B^T h(X_2)$ where B is a matrix of unknown coefficients and $h(X_2)$ is a user-specified vector-valued function of X_2. Alternatively $E(X_1 \mid X_2)$ could be estimated by using nonparametric regression. Coefficient estimates are obtained by minimizing a convex objective function :

$$(\widehat{a}_0, \widehat{a}_1, \widehat{b}) = \arg\min L_n(a_0, a_1, b) \qquad (2.2)$$

where $L_n(a_0, a_1, b) = \frac{1}{n}\sum_{i=1}^{n} L(y_i - a_0 - a_1^T x_i - b^T E(X_1 \mid X_{2i}))$ and L is a convex objective function. Based on the model (2.1) and estimates in (2.2), the estimate $\widehat{a_1}$ converges almost surely to a_1, and consequently $e_i + \widehat{b^T} m(X_{2i})$ converges to $constant + g(X_{2i}) + \varepsilon_i$. The plot $\{X_2\ ,\ e + \widehat{b^T} E(X_1 \mid X_2)\ \}$ is referred to as CERES plot, an abbreviated acronym for "Combining Conditional Expectations and RESiduals". If $E(X_1 \mid X_2)$ is linear in X_2 CERES plots are partial residual plots and if $E(X_1 \mid X_2)$ is quadratic in X_2 then CERES plot are same as an augmented partial residual plots. When the conditional expectations $E(X_1 \mid X_2)$ are all linear, partial residual plot would be enough to examine the curvature in regression problem. But if $E(X_1 \mid X_2)$ is nonlinear, augmented partial residual plots and CERES plots should be considered and CERES plots enhance the resolution of a week trend of curvature in an augmented partial residual plot.

CERES plot was developed under the assumption that X_1 enters the model linearly. If this assumption is not satisfied CERES plot for X_2 may contain a notable bias. This kind of bias was mentioned as leakage effect by Chamber et al. (1983, p.306). To see how the leakage effect comes about, consider a CERES plot for X_2 based on OLS estimation in a situation in which X_1 may contribute nonlinearly,

$$y = a + g_1(x_1) + g_2(x_2) + error \tag{2.3}$$

where $E(g_j) = 0$, $E(X_j) = 0$, and the errors are *iid*. It is shown that a CERES plot for X_2 which is constructed from (2.1) will display g_2 as indicated previously only if X_1 and X_2 are independent or if $g_1(x_1)$ is linear in x_1. Thus if $g_1(x_1)$ is nonlinear and X_1 and X_2 are sufficiently dependent then the effect of g_1 may cause a notable bias in the CERES plot for X_2, even if g_2 is linear. So when the model of (1.2) is assumed CERES plot is not an appropriate tool for uncovering the curvature of X_2. Cook(1993) suggested that higher order terms and transformations could be incorporated in x_1. We consider using inverse response plot for transforming response variable y in an effort to satisfy the assumption of linearity.

2.2 Response transformation using Inverse Response Plot

Assume that there is an monotonic function t such that

$$t(y) = \beta_0 + \beta^T X + error \tag{2.4}$$

and consider the problem finding a response transformation $t(y)$. If $k\beta$ were known the plot $\{y, k\beta^Tx\}$ will provide a visualization of an appropriate transformation. According to the different value of k, the transformations displayed may not be same, but they are related linearly and are all satisfy (2.4). Cook and Weisberg (1994) suggested an inverse response plot taking the following approach. Since exact value of $k\beta$ is not so sure in most cases maximum likelihood-type regression based on a linear model $y = \alpha_0 + \alpha^Tx + e$ is considered as a practically useful estimators of $k\beta$. The maximum likelihood estimate of (α_0, α) is obtained similarly to (2.2) by minimizing an objective function $n^{-1}\sum L(\alpha_0 + \alpha^Tx_i, y_i)$, where $L(m, y)$ is convex in m for each y. Ordinary least squares is an example of possible estimation method. Following Li & Duan (1989), $\hat{\alpha}$ is a consistent estimator of $k\beta$ if $E(x \mid \beta^Tx)$ is linear in β^Tx. This condition hold for all β if and only if x has an elliptically contoured distribution (Eaton, 1986). Since we can get a consistent estimate of $k\beta$ using Li & Duan's results, we now assume that β is known in the argument of population case.

We consider the plot $\{y, \beta^Tx\}$ for obtaining a good impression of an appropriate transformation. The plot $\{y, \beta^Tx\}$ is useful when, at least approximately, $E(\beta^Tx \mid y) = t(y)$. Since $t(y)$ is assumed to be monotonic the following holds :

$$\begin{aligned} E(\beta^Tx \mid y) &= E(\beta^Tx \mid t(y)) \\ &= t(y) - E(\varepsilon \mid \beta^T + \varepsilon) \end{aligned} \tag{2.5}$$

Thus to satisfy the condition $E(\beta^Tx \mid y) = t(y)$, it requires that $E(\varepsilon \mid \beta^Tx + \varepsilon)$ should be linear in $t(y)$. This condition will hold if (β^Tx, ε) follows an elliptically contoured distribution (Cambanis, Huang and Simons, 1981).

To measure the degree of linearity, the population correlation coefficient between $E(\beta^Tx \mid t(y))$ and $t(y)$

$$\rho = \frac{Cov[E(\beta^Tx \mid t(y)), t(y)]}{Var(E(\beta^Tx \mid t(y)))^{1/2}\ Var(t(y))^{1/2}} \tag{2.6}$$

can be used. Once $t(y)$ is estimated from the inverse plot applicability of the method can be checked by calculating a sample correlation between $\widehat{t(y)}$ and an estimate of $E(\beta^Tx \mid y)$ which is obtained by smoothing the plot $\{\widehat{t(y)}, \widehat{\beta^T}x\}$.

3. Transfomation and visualization in the fitting of nonadditive models

3.1 Linear approximation

We now consider a method for fitting the nonadditive model of (1.2). We assume that there is a strictly monotonic transformation of the response, with which nonadditive model (1.2) can be expressed as one of generalized additive models. In other words we assume that there is a strictly monotonic function t for some h which satisfies the following relation

$$t(y)=\alpha+\beta_1^T x_1+h(x_2)+\varepsilon \tag{3.1}$$

and that h is sufficiently smooth for a simple linear approximation to h to work well locally. Once an appropriate transformation t is found, the curvature h can be fitted by one of known methods. We suggest to use an inverse response plot for visualizing t and to use a CERES plot for h.

To use the inverse response plot for finding transformation t, model (3.1) should be expressed linearly. For a linear representation of h, locally linear approximation method is used (Johnson and McCulloch, 1987). We first partition observations by their x_2 value. A set of n observations is partitioned into subsets so that within each subset the values of the variable x_2 do not vary much relative to the overall variation in x_2. Within each subset of our partitioning scheme we will assume that h is linear and the slope and intercept will be allowed to vary among subsets. Once the partition is chosen we then have the following model

$$t(y_{ij}) = \beta_1^T x_{ij} + a_i + b_i(x_{2ij}-\overline{x_{2i}}) + \varepsilon_{ij} \qquad i=1,\cdots,k \quad j=1,\cdots,n_i \tag{3.2}$$

where y_{ij} is the value of j th response variable in the i th subset, n_i is the number of observations in the i th subset and k is the number of subsets. At (3.2) we have used the approximation $h(x_{2ij}) \approx a_i + b_i(x_{2ij}-\overline{x_{2i}})$ where $\overline{x_{2i}}$ is the mean to the x_2 values of the observations in the i th subset. Using one subscript for each variable, x_{2m} and $\overline{x_{2l}}$ denote the m th value of x_2 variable and mean of x_2 values of the observations in the l th subset respectively. Two n by k matrices D_a and Z_b are defined such that value of $(m\ ,\ l)$th element of D_a and Z_b is, respectively, 1 and $(x_{2m}-\overline{x_{2l}})$ if x_{2m}belongs to l th subset, or zero otherwise. Taking locally linear approximation model (3.1) is written as

$$t(y) \approx \alpha + \beta_1^T x_1 + D_a a + Z_b b + \varepsilon. \tag{3.3}$$

Now a response transformation $t(y)$ can be estimated by using inverse response plot with covariates x_1, a and b in (3.3). And with the transformed response variable $\hat{t}(y)$ CERES plot for x_2 is used for uncovering $h(x_2)$.

3.2 Examples

Two examples are proposed and the related programs are coded by using Xlisp-stat (Tierney, 1990).

Example 1. (Artificial data) For the example, 20 observations were generated according to the model

$$y = (5 + 3x_1 + 1/x_2^2)^2 ,$$

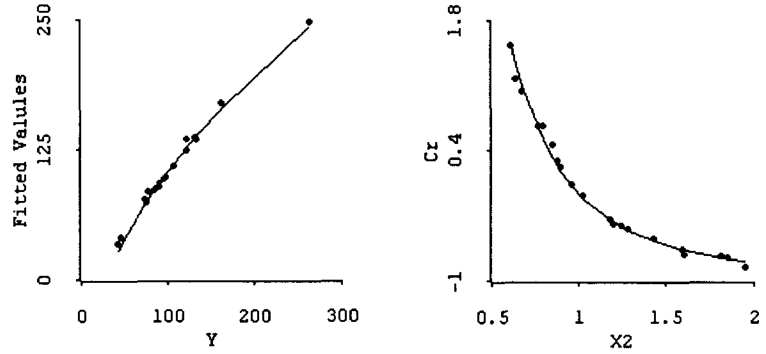
where $x_1 = 1/x_2 + N(0,1)$, $x_2 = 0.3u_1 + 0.2u_2$, u_1 and u_2 are uniform random variables on the interval (1, 3) and (1, 8) respectively. Following the notations in model (3.1), we have $t(y) = \sqrt{y}$, $h(x_2) = 1/x_2^2$. Table 1 contains the data and a partitioning scheme.

Table 1. The data and Partition Scheme for Example 1.

case	y	x1	x2	partition	case	y	x1	x2	partition
1	131	1.29	0.619	1	11	89.2	1.25	1.19	3
2	264	2.95	0.647	1	12	108	1.57	1.2	3
3	162	1.86	0.68	1	13	91.6	1.31	1.25	3
4	122	1.44	0.767	1	14	96.8	1.41	1.29	3
5	47.8	0.111	0.795	1	15	87.3	1.28	1.43	4
6	43.8	0.0814	0.854	1	16	90.3	1.37	1.6	4
7	78.5	0.856	0.881	1	17	123	1.9	1.61	4
8	85.8	1.011	0.902	2	18	75.4	1.13	1.82	5
9	91.6	1.17	0.968	2	19	75.6	1.14	1.86	5
10	98.9	1.33	1.03	2	20	134	2.1	1.96	5

Conditional expectations $E(x_1 \mid x_2)$ are estimated by using LOESS curves. Figure 1 (A) shows the inverse response plot with the given partitioning scheme. The superimposed line on the figure 1 is obtained from the ordinary least squares regression of y on $\sqrt{y}$. The curve indicate that $\sqrt{y}$ transformation with $corr(\hat{y}, \sqrt{y})$ being 0.99, is a strong candidate for achieving linearity in (3.1). $Corr(E(\beta^T \mid \sqrt{y}), \sqrt{y})$, the measurement of linearity between $E(\beta^T x \mid t(y))$ and $t(y)$, is 0.98, which is big enough to guarantee the applicability of the inverse response plot to this data. With different partitioning schemes we have similar plots.

Figure 1 (B) is CERES plot for x_2 after transforming response variable into $\sqrt{y}$. The superimposed line is $g-\bar{g}$.



(A) (B)

Figure 1. (A) Inverse response plot of $\{ y , \hat{y} \}$ for the example 1. The line on the plot is obtained from the ordinary least squares regression of $\hat{y}$ on $\sqrt{y}$.

(B) CERES plot for x_2 of Example 1 with transformed response $\sqrt{y}$. The superimposed line is $g-\bar{g}$ where $g= (1/x_2)^2$.

When the model of (1.1) is assumed, constructing a variety of CERES plots seems common way to defect curvature. But deciding which variables correspond to x_2 is not straightforward (Cook 1993). We propose an example impling that the new method is helpful to decide which variable corresponds to x_2.

Example 2. (Real data) The ozone pollution data given by L. Brieman (1985) were discussed by many authors (Fried man and Silverman, 1989; Hastie and Tibshirani 1990, p.294). The data consist of atmospheric ozone concentration (Y) from eight daily meteorological measurements (X) made in the Los Angeles basin for 330 days in 1976. For each covariate we decide the appropriate transformation of response variable for (3.1) from an inverse response plot. IBTP with response transformation of $y^{0.7}$ is decided as x_2, with the highest value of $corr(\hat{y} , t(y))=0.78$ and $corr(E(\beta^T x \mid t(y)) , t(y))=0.99$, among eight variables for a variety partitioning schemes. Figure 2 shows the inverse response plot and regression line of y on $y^{0.7}$. This implies that with the transformed response variable and IBTP as x_2, linearity of the model as in (3.1) is well guaranteed. Figure 3 (A) is CERES plot for IBTP and Figure 3 (B) is CERES plot for IBTP with the transformed response variable by $y^{0.7}$. The apparent systematic effect of IBTP has been reduced.

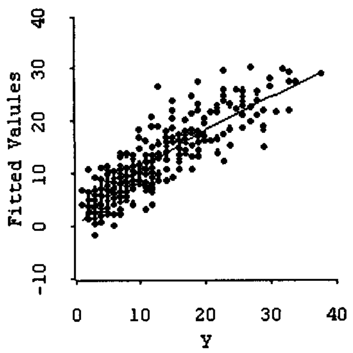
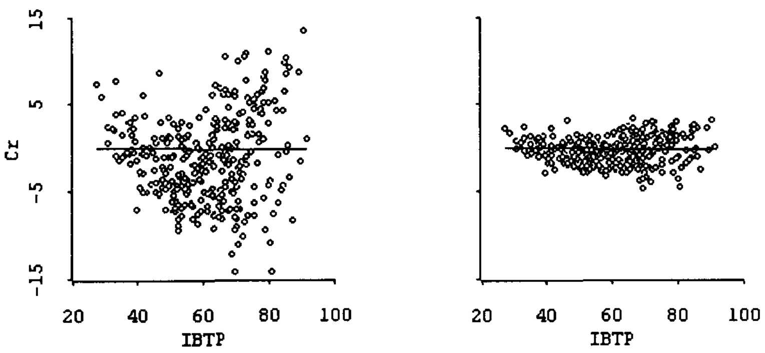


Figure 2. Inverse response plot of { y , $\hat{y}$ } for the ozone pollution data.
The line on the plot is obtained from the ordinary least squares regression of $\hat{y}$ on $y^{0.7}$.



(A) (B)

Figure 3. CERES plots for IBTP with (A) y and (B) $y^{0.7}$

4. Remarks

Methodology suggested in this article consists of two procedures, partitioning data for transforming response variable and using CERES plots. Partitioning scheme does effect on only the decision of response transformation. It is difficult to determine the optimal choice of partition schemes analytically to balance all factors. However, since there is no difficulty trying various partition schemes, if the outcome is insensitive to changes in partition scheme, then we are reassured. Usefulness of CERES plot depends on how much the estimation of $E(x_1 \mid x_2)$ is accurate.

References

[1] Berk, K. N., and Booth D. E. (1995). Seeing a curve in multiple regression, *Journal of the American Statistical Association,* Vol. 37, 385-398.

[2] Brieman, L., and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion), *Journal of the American Statistical Association.* Vol. 80, 580-619.

[3] Cambanis, S., Huang, S., and Simons, G. (1981). On the theory of elliptically contoured distributions, *Journal of Multivariate Analsis,* Vol. 7, 368-385.

[4] Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. (1983). *Graphical methods for data analysis,* Duxbury Press, Boston.

[5] Cook, R. D. (1993). Exploring partial residual plots. *Technometrics* Vol. 35, 351-362.

[6] Cook, R. D. and Weisberg, S. (1994). Transforming a response variable for linearity, *Biometrika,* Vol. 81, 731-737.

[7] Eaton, M. L. (1986). A Characterization of spherical distribution, *Journal of Multivariate Analysis,* Vol. 20, 272-276.

[8] Friedman, J. H., and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion), *Technometrics,* Vol. 31, 3-40.

[9] Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized additive models.* Chapman and Hall, New York.

[10] Johnson, B. W. and McCulloch, R. E. (1987). Added-variable plots in linear regression, *Technometrics,* Vol. 29, 427-433.

[11] Larsen, W. A. and McCleary, S. J. (1972). The use of partial residual plots in regression analysis, *Technometrics,* Vol. 14, 781-790.

[12] Li, K. C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics,* Vol. 17, 1009-1052.

[13] Mallows, C. L. (1986). Augmented partial residual plots. *Technometrics.* Vol. 28, 313-320.

[14] Tierney, L. (1990). *Lisp-stat,* Wiley, New York.

[15] Weisberg, S. (1985). *Applied linear regression,* Wiley, New York.