The Korean Journal of Applied Statistics (2009)
22(1), 189–195

A Dynamic Graphical Method for Transformations and Curvature Specifications in Regression

Han Son Seo[1] · Min Yoon[2]

[1]Dept. of Applied Statistics, Konkuk University; [2]Dept. of Applied Statistics, Konkuk University,

(Received December 2008; accepted January 2009)

Abstract

A dynamic graphical procedure is suggested to estimate optimal response transformation parameter and a curvature function of covariates in the regression model. Augmented partial residual plot is chosen for specifying a curvature. The proposed method is compared with a different approach (Seo, 2007) and is investigated efficiency by applying it to the real and the artificial data. The method is also extended to the 3D graphical situations.

Keywords: Augmented partial residual plot, dynamic graphics, response transformation.

1. Introduction

There are several assumptions underlying a standard linear regression analysis. The linear regression model can be more applicable by allowing response transformation and nonlinear relationship between a response and some explanatory variables. Incorporating the possibility of transforming the response and curvature function of some explanatory variables, the regression model can be represented as follows.

$$Y^{(\lambda)} = \alpha + X\beta + h(Z) + \epsilon, \tag{1.1}$$

where Y is the response, λ is an unknown power, X^T is $p_1 \times 1$ and Z^T is $p_2 \times 1$ vector of covariates and h is unknown function with $E(\epsilon|\, x_1, x_2) = 0$.

In this article we concern with the problem of estimating optimal response transformation parameter λ and capturing a curvature h simultaneously. Theoretical solution to this problem is very difficult and complicated. If we take advantage of recent advances in computer technology we can interact with the data easily and can find the solution empirically.

The basic idea is to fit the model (1.1) with a fixed value of λ and $\hat{h}$, a graphically estimated function of h, and then change λ according to some control parameter so that the optimal λ and corresponding $\hat{h}$ is found by monitoring the fitness of the model.

If we fix λ in (1.1), we can restrict attention to the problem of specifying the curvature h.

This paper was supported by Konkuk University in 2008.

[1]Corresponding author: Professor, Dept. of Applied Statistics, Konkuk University, 1 Hwayang-Dong, Gwangjin-Gu, Seoul 143-701, Korea. E-mail:hsseo@konkuk.ac.kr

Consider the following model:

$$Y = \beta_0 + X\beta_1 + g(Z) + \epsilon, \tag{1.2}$$

where Y is the response, X^T and Z^T are vectors of covariates and g is unknown function. For visualizing g in (1.2) many graphical methods are suggested including added variable plot (Chamber *et al.*, 1983, p.272), partial residual plot (Larsen and Mccleary, 1972; Weisberg, 1985), augmented partial residual plot (Mallows, 1986) and CERES plot (Cook 1993).

For the construction of partial residual plots we fit a linear model

$$Y = a_0 + Xa_1 + Zb + \text{error} \tag{1.3}$$

and estimate coefficients by minimizing a convex objective function :

$$\left(\hat{a_0}, \hat{a_1}, \hat{b}\right) = \arg\min L_n(a_0, a_1, b), \tag{1.4}$$

where $L_n(a_0, a_1, b) = 1/n \sum_{i=1}^{n} L(y_i - a_0 - x_i a_1 - z_i b)$ and L is a convex objective function. A partial residual plot for Z is defined as $e + Z\hat{b}$ on the vertical axis and Z in the horizontal axis.

Augmented partial residual plots are designed to depict g better than partial residual plots. Augmented partial residual plot for Z is the plot of $e + \hat{\phi_1} Z + \hat{\phi_2} Z^2$ versus Z, which is constructed from the following model containing a quadratic term in Z and estimates coefficients from (1.4)

$$Y = \rho_0 + X\rho_1 + \phi_1 Z + \phi_2 Z^2 + \text{error}. \tag{1.5}$$

CERES plots come from the fact that $E(X|Z)$ should be included in the function of Z as a special case to depict f accurately. It is constructed based on the model

$$Y = a_0 + Xa_1 + E(X|Z)b + \text{error}. \tag{1.6}$$

CERES plots is described as the plot of $e + E(X|Z)\hat{b}$ versus Z where $E(X|Z)$ are modeled either parametrically or nonparametrically and estimates are obtained from (1.4). Estimate $\hat{a_1}$ in (1.4) converges almost surely to β_1 in (1.2), and consequently $e + E(X|Z)\hat{b}$ converges to constant$+g(Z)+\epsilon$.

For finding optimal λ and specifying $h(Z)$ Seo (2007) suggested a graphical method using inverse response plot (Cook and Weisberg, 1994). But the method is limited in its application since the inverse response plot requires several conditions, for example, monotonicity of the function h. We suggest a new graphical method for estimating λ and revealing h using augmented partial residual plot as a tool for specification of h.

Whatever the form of f is, an augmented partial residual plot can display f better than a partial residual plot. CERES plots are designed to work well even when predictors are arbitrary noise function of each other or when $E(X|Z)$ are neither linear nor quadratic. But Many examples show that the efficiency of CERES plots depends sensitively on the accuracy of the estimated value of $E(X|Z)$.

In Section 2 a new procedure for estimating λ and h in (1.1) is proposed and is compared with other method through examples. Section 3 contains discussion.

2. A Dynamic Graphical Method

Under the formulation of a regression model in which covariates enter nonlinearly with a power transformed response variable Seo (2007) suggested a graphical procedure for estimating an optimal

response transformation parameter and specifying the curvature function. The procedure using inverse response plot (Cook and Weisberg, 1994) for curvature specification, is summarized briefly as follows.

If we exchange $Y^{(\lambda)}$ and $h(Z)$ in (1.1) the model can be expressed as

$$h(Z) \approx \alpha^* + \beta^{*\ T} + \kappa Y^{(\lambda)} + \epsilon^*. \tag{2.1}$$

For a fixed value of λ an inverse response plot of Z with covariates X and $Y^{(\lambda)}$ is constructed and is used to estimate the curve $h(Z)$. After fitting the regression model $Y^{(\lambda)}$ on X and $\hat{h_\lambda}(Z)$ the forward response plot $\{\ \hat{Y}^{(\lambda)}, Y^{(\lambda)}\}$ is drawn, where $\hat{Y}^{(\lambda)}$ is the fitted value. The best value of λ, denote it as λ^* and corresponding $\hat{h}_{\lambda^*}$ are determined from the inverse response plot and forward response plot as changing λ smoothly. This procedure is limited due to the fact that inverse response plot is designed to estimate only a monotonic function h and needs covariates following an elliptically contoured distribution (Eaton, 1986). We propose to use augmented partial residual plot for capturing $h(Z)$ to overcome the limitations as addressed above. Advantages of using augmented partial residual plot over other competing graphical methods are explained in Section 1.

For a fixed value of λ the new procedure applies augmented partial residual plot to the model (1.1) for estimating h. Then $\hat{Y}^{(\lambda)}$ is calculated from the regression Y on X and $\hat{h_\lambda}(Z)$. After plotting augmented partial residual plot and forward response plot simultaneously the optimal value of λ and corresponding estimated curve of $h(Z)$ are determined by monitoring points in the plots, which change dynamically as λ changes.

More detailed procedure is outlined as follows.

1 Fix the value of λ.

2 Do a linear regression $Y^{(\lambda)}$ on X and Z and make an augmented partial residual plot of Z.

3 Estimate $h_\lambda(Z)$ using a polynomial function from the augmented partial residual plot.

4 Do a regression $Y^{(\lambda)}$ on X and $\hat{h_\lambda}(Z)$ and calculate the fitted values, $\hat{Y}^{(\lambda)}$.

5 Make a forward response plot, $\{\hat{Y}^{(\lambda)}, Y^{(\lambda)}\}$.

6 Change λ smoothly from -2 to 2 and stop at which points on forward response plot show a linear trend.

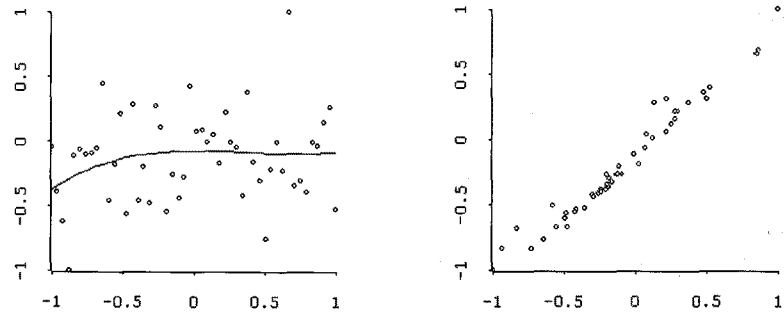
Fixing λ as the optimal value $h(Z)$ can be estimated by using power family function $h(Z) = Z^\delta$ for some $\delta \in (-2, 2)$ through a similar process done before.

The proposed method is compared with other approaches and is extended to the 3D graphical situations. All programs are coded under the environment of the package ARC (Cook and Weisberg, 1994) which is made by using Xlisp-stat (Tierney, 1990).

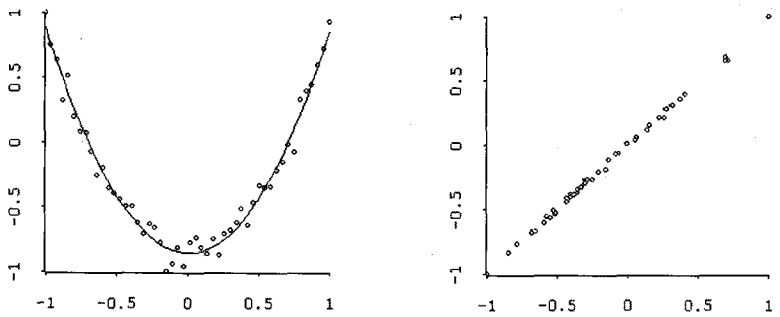
EXAMPLE 2.1. (2D). For the example, 50 observations were generated according to the model

$$Y = \exp(X_1 + X_2 + Z^2 + \epsilon),$$

where X_1 and X_2 are independent normal random variables with mean 5 and variance 1, Z contains equally spaced values between -1 and 1, and ϵ follows normal distribution with mean 0 and standard deviation 0.05. Following the notations used in model (1.1), we have $\lambda = 0$ and $h(Z) = Z^2$. Both by using inverse response plot and augmented partial residual plot, we see the strongest linear trend of points in forward response plot when $\lambda = 0$. But as seen in Figure 2.1(a) and (b) inverse response



(a) Inverse response plot and forward response plot of $\lambda = 0$



(b) Augmented partial residual plot and forward response plot of $\lambda = 0$

Figure 2.1

plot fails to capture curvature function $h(z)$ whereas augmented partial residual plot captures it successively.

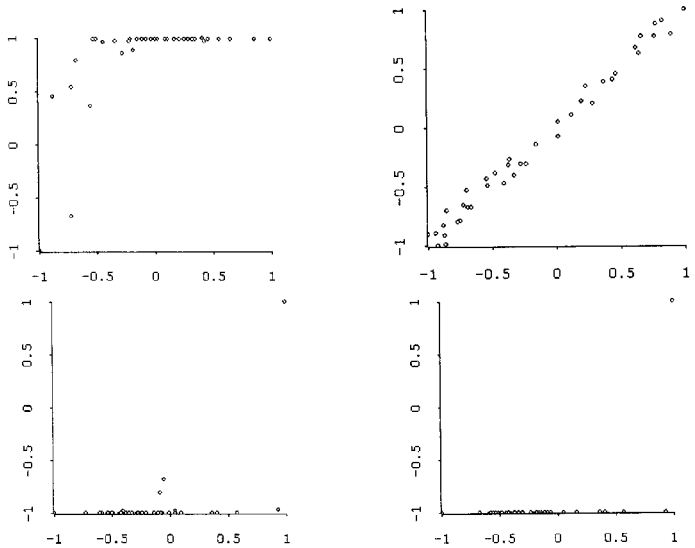
We can extend the proposed method to the following model by using 3D augmented partial residual plot:

$$Y^{(\lambda)} = \alpha + X\beta + g(Z_1, Z_2) + \epsilon.$$

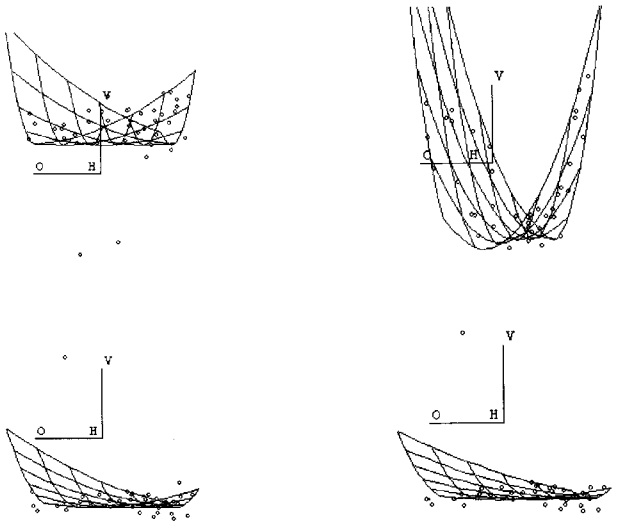
Throughout the examples 3D augmented partial residual plot based on (1.5) is used. We display $\hat{g}(Z_1, Z_2)$ on the vertical axis (V-axis), and Z_1 and Z_2 on the two horizontal axes (H-axis and O-axis). 3D plot can be rotated about each of three axes. 3D rotation about the vertical axis is nothing more than rapidly updating the 2D plot y versus a liner combination of Z_1 and Z_2 in small steps. We usually stop rotation about vertical axis when 2D plot shows the most evident trend. We call this plot as *the best view plot*. 3D plot has a slider bar reflecting the change of λ. As holding down the mouse button on the slider, scroll bar is moved, the display value of λ is changed and so is the plot dynamically. Changes occurred in 3D plots are observed as λ changes.

EXAMPLE 2.2. (3D). An artificially created sample of 40 observations was used. X_1, X_2 and W are independent uniform random variables on the interval $(40, 80)$, $(10, 60)$ and $(10, 60)$ respectively. For each value of W, Z_1 is randomly generated from uniform distribution with range $(0, W)$ and Z_2 is calculated as $W - Z_1$. Error term ϵ follows standard normal distribution and Y is generated by the model:

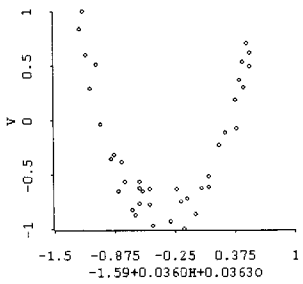
$$\log Y = 0.31X_1 + 0.71X_2 + (Z_1 + Z_2 - 10)(Z_1 + Z_2 - 60) + \epsilon.$$



(a) Forward response plot of $\lambda = -0.2, 0, 0.3, 1$ (From top left, clockwise)

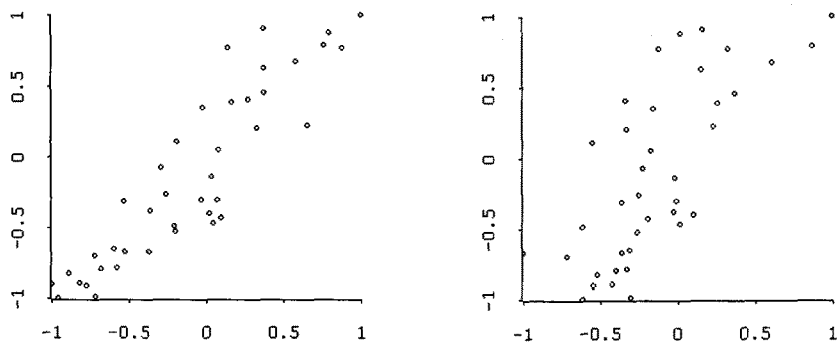


(b) Augmented partial residual plot of $\lambda = -0.2, 0, 0.3, 1$ (From top left, clockwise)



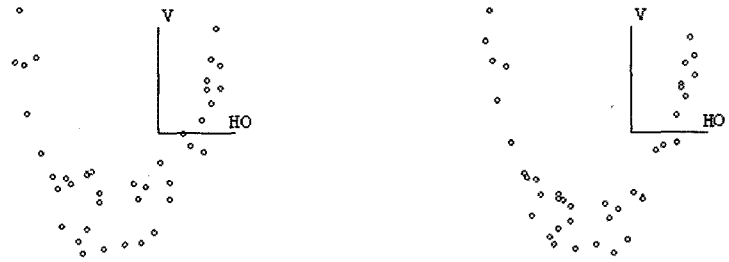
(c) Best plot of 3D augmented partial residual plot of $\lambda = 0$

Figure 2.2



(a) linear + quadratic (b) linear + interaction

Figure 2.3. Forward response plot using different function for estimating points on augmented partial residual plot



(a) linear (b) linear+interaction

Horizontal: $-1.590 + .03595\text{H} + .03631\text{O}$ Horizontal: $-1.591 + .03610\text{H} + .03619\text{O}$

Figure 2.4. Best view plot of different types of 3D added variable plots (curve fitting is done by linear + interaction + quadratic)

Four frames of $\{\hat{Y}^{(\lambda)}, Y^{(\lambda)}\}$ in Figure 2.2(a) correspond to $\lambda = -0.2$, 0, 0.3, 1 respectively. Values in axes are scaled so that the minimum and maximum values are -1 and 1 respectively. Plots in Figure 2.2(b) are the corresponding augmented partial residual plot of $\{Z, \hat{Z}_\lambda\}$. The superimposed line on each plot in Figure 2.2(b) stands for $\hat{h}(Z)$, estimated by using a full quadratic function. When $\lambda = 0$, plot of $\{\hat{Y}^{(\lambda)}, Y^{(\lambda)}\}$ shows the most strong linear trend visually.

Four flames of Figure 2.2(c) are the best view of 3D augmented partial residual plot for $\lambda = 0$. Their corresponding horizontal linear combinations are $0.0360X_1 + 0.0363X_2$. The ratio of coefficients, 0.0360/0.0363, approximately equals to the true value of 1.

Figure 2.3 shows forward response plots of $\lambda = 0$ when we estimated $g(Z_1, Z_2)$ from 3D augmented partial residual plot using $\hat{g}(Z_1, Z_2) = \hat{a_1}Z_1 + \hat{a_2}Z_1^2 + \hat{a_3}Z_2 + \hat{a_4}Z_2^2$ and $\hat{g}(Z_1, Z_2) = \hat{a_1}Z_1 + \hat{a_2}Z_2 + \hat{a_3}Z_1Z_2$ respectively. The linearity is less clearer than when a full quadratic function is used for the estimation of $\hat{g}$ in augmented partial residual plot. Figure 2.4 contains best view plots of 3D added variable plots constructed from the model $Y = \rho_0 + X\rho_1 + \phi_1 Z_1 + \phi_2 Z_2$ and $Y = \rho_0 + X\rho_1 + \phi_1 Z_1 + \phi_2 Z_2 + \phi_3 Z_1 Z_2$. The shape of the function is less obvious in the best view plots at which the ratios of two coefficients are close to 1.

3. Discussion

This paper studies the graphical method of curvature specification in the model of transformed response variable and suggests to use dynamic augmented partial residual plot. The advantage to dynamic graphical approach is that the behavior of observations can be seen in the process of searching the optimal transformation and estimating curvature function. Thus outliers or influential cases in the context of transformation or curvature fitting can also be detected. Unlike inverse response plot approach the proposed procedure can be extended to 3D situations. When collinearity among predictors is very severe graphical methods discussed in section 1 fail to perceive the curve. In the highly collinearity situation, the plot of residual against predictor can be used as Berk and Booth (1995) recommended .

References

Berk, K. N. and Booth D. E. (1995). Seeing a curve in multiple regression, *Technometrics*, **37**, 385–398.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. (1983). *Graphical Methods for Data Analysis*, Duxbury Press, Boston.

Cook, R. D. (1993). Exploring partial residual plots, *Technometrics*, **35**, 351–362.

Cook, R. D. and Weisberg, S. (1994). Transforming a response variable for linearity, *Biometrika*, **81**, 731–737.

Eaton, M. L. (1986). A characterization of spherical distributions, *Journal of Multivariate Analysis*, **20**, 272–276.

Larsen, W. A. and Mccleary, S. J. (1972). The use of partial residual plots in regression analysis, *Technometrics*, **14**, 781–790.

Mallows, C. L. (1986). Augmented partial residual plots, *Technometrics*, **28**, 313–320.

Seo, H. S. (2007). A visual procedure for response transformations and curvature specifications, *The 7th International Conference on Optimization: Techniques and Applications (ICOTA7)*, CD-ROM.

Tierney, L. (1990). *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*, John Wiley & Sons, New York.

Weisberg, S. (1985). *Applied Linear Regression*, John Wiley & Sons, New York.