L_1-*Statistical Procedures and Related Topics*
IMS Lecture Notes – Monograph Series (1997) Volume 31

Interactive graphical modeling methods for regression

Nate Wetzel

State University of New York, Binghamton, USA

Abstract: Identification of curvature in regression models is an important aspect of data analysis. Partial residual plots have played a major role. Recently a new class of plots has been developed. They are called CERES plots and include partial residual plots as a special case. Implementation of these plots necessitates modeling the relationships between certain covariates. If these relationships are linear, a partial residual plot is obtained. However, if the relationships are nonlinear, the more general CERES plot is obtained. Generalized additive models (GAM) are another method for identifying and estimating curvature. Again, implementation of a GAM requires modeling the relationships between covariates and the response. Here, we motivate and describe key features of interactive, graphical methods which construct CERES plots and/or GAMs.

Key words: Partial residual plots, CERES plots, Generalized Additive Models, XLISP-STAT, S-PLUS.

AMS subject classification: 62G07.

1 Introduction

Conditional expectation residual plots (CERES plots, see Cook, 1993) and Generalized Additive Models (GAMs, see Hastie and Tibshirani, 1990) have been developed in the literature as diagnostic and modeling tools for regression analysis. These methods are designed to detect curvilinear relationships between selected covariates and the response variate in regression. When used interactively, these methods can help detect outliers, give information about possible heteroscedasticity.

In this paper, we outline the basic theory and assumptions underlying CERES plots and GAMs. Using simulated data, we then illustrate how these methods are used and implemented. Both of these methods rely on the use of scatterplot smoothers. The examples are intended to highlight the usefulness of an implementation which (1) shows the data and associated scatterplot smoothers and (2) has an interactive interface so that smoothers can be easily changed and results compared.

2 CERES plots and GAMs - a primer

Consider the regression model (given X_1 and X_2) $Y = \alpha_0 + g_1(X_1) + g_2(X_2) + \epsilon$, where α_0 is an unknown constant and g_1 and g_2 are unknown functions with $E(g_i(X_i)) = 0$ and $E(\epsilon|X1, X_2) = 0$. In general, X_1 and X_2 may be random vectors, but for the purposes of this paper, X_1 and X_2 are random variables. In other words, for the purposes of this paper, there are two predictor variables.

The idea behind CERES plots (see Cook, 1993) is that if g_1 is the identity function and $E(X_1|X_2)$ is known, then a CERES plot will display the function g_2. This display will be with error and possible vertical shift. In practice, $E(X_1|X_2)$ is unknown, so we estimate it by smoothing the plot of X_1 versus X_2, and then estimate g_2 by smoothing the CERES plot which is obtained by assuming that our estimate of $E(X_1|X_2)$ is correct. An implementation of CERES plots using the XLISP-STAT software (see Tierney, 1990) is given in a paper by Wetzel (1996).

GAMs have the additional assumption that ϵ is independent of (X_1, X_2), and the basic idea is that if we know g_1 then $E(Y - \alpha_0 - g_1(X_1)|X_2) = g_2(X_2)$. In practice g_1 is unknown, so we use an iterative algorithm to estimate g_1, then g_2, then g_1, etc. An implementation of GAMs is given in the S-PLUS software.

The theory underlying both CERES plots and GAM is powerful; however, when used in practice, we need the implementations to be interactive enough so that we can be critical users. When using the above methods in exploratory data analysis, we need to be able to look 'behind the scenes' to, in the CERES case, see the smooth which estimates $E(X_1|X_2)$, and in the GAM case, see the iterative process. In order to critically use these methods, we must be able to adjust and see the new results quickly. This need is demonstrated in the next section.

3 The need for interactive methods

In this section, we will look at a few examples which illustrate the need for interactive methods when using either CERES plots and/or GAM.

3.1 The influence of the choice of smoother in GAMs

The first example involves randomly generated data with the following distributions: (1) $X_1 \sim N(0,1)$, (2) $X_2|X_1 \sim N(4+.25*X1,.01)$, (3) $Y = 6 + X_1^2 + \log(X_2 - 3)$.

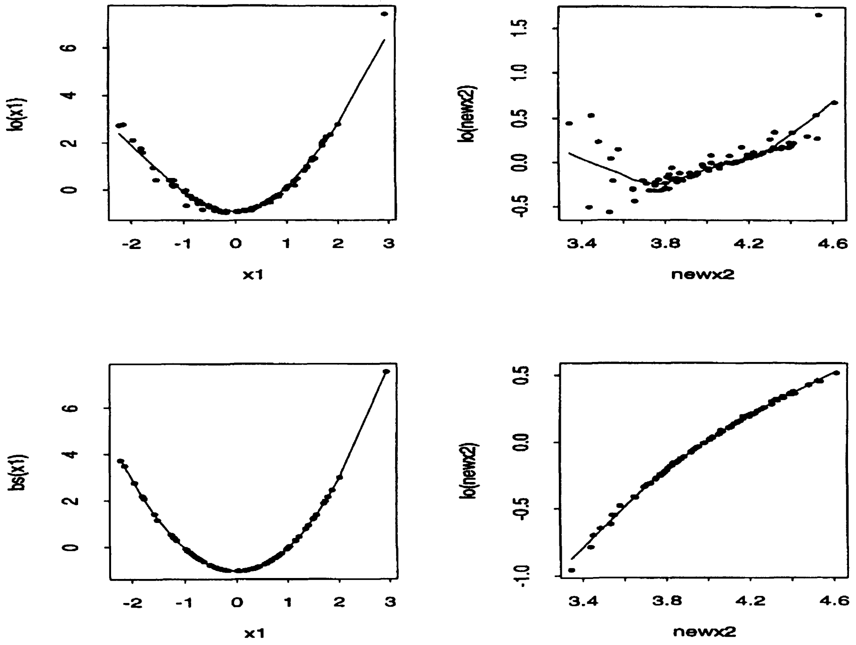

Figure 1: Estimated additive functions from GAM.

Two Generalized Additive Models were fitted and the summary plots from S-PLUS are given in Figure 1. The top two plots are from the fit given by the Splus command `gam(y ~ lo(x1) + lo(x2))` and the bottom two plots from the command `gam(y ~ bs(x1) + lo(x2) , bf.maxit=20)`. Here `lo` and `bs` correspond to a loess fit and b-spline fit, respectively.

The summary plots from S-PLUS show the estimated functions $\hat{g}_1$ and $\hat{g}_2$ as well as the points used to estimate these curves. The horizontal axes have the predictors X_1 and X_2, and the vertical axes have what can be thought of as partial residuals. They are partial residuals, but are weighted in a non-trivial way. As analysts, we are to know that if the points are closer to the curve, then the fit will have smaller residual sum of squares.

Notice that the predicted function of X_1 in both cases appears to be quadratic. In the first GAM, we see a loess curve, and in the second a b-spline fit. However, the predicted functions for X_2 are very different. In the first case it appears that Y is dependent on X_2 quadratically, and in the second, we see the true logarithmic relationship.

This shows that the choice of smoothers in GAM is very important. Ideally, a dedicated data analyst would see that in the first GAM, the loess smooth for X_1 is under-fitting for both the small and large values of X_1. An interactive interface should allow them to interactively change the smoother for X_1. An interactive investigation of the 'outlier', when $X_1 \approx 3$, may also be informative. In this case, deleting the 'outlier' does not significantly change the predicted model, while changing the smoother does.

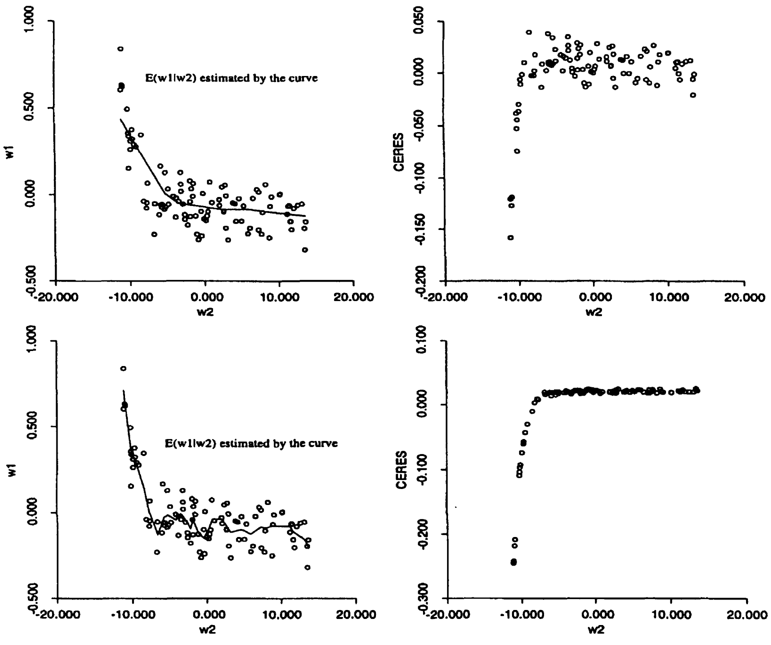

Figure 2: Estimated $E(W_1|W_2)$ and associated CERES plot.

3.2 Looking at intermediate plots for CERES

As demonstrated in Wetzel (1996), the intermediate step of estimating $E(X_1|X_2)$ has a great influence on the resulting prediction. Here, we generate data similar to Wetzel (1996): (1) $W_2 \sim$ Uniform[1,26], (2) $W_1|W_2 \sim N(1/W_2, .01)$, (3) $Z = W_1 + 1/(1+\exp(-W_2))$. In this example $g_2(w_2) = 1/(1+e^{-w_2})$. A graph of the function g_2 after being shifted both horizontally and vertically, looks exactly like the lower right plot in Figure

2. The plots in Figure 2 show two choices of smoother used to estimate $E(W_1|W_2)$ and the associated CERES plot. These plots were obtained using the XLISP-STAT package and the code developed in Wetzel (1996). The leftmost plot shows the smooth used to estimate $E(W_1|W_2)$ and the rightmost plot shows the CERES plot. (W_1 and W_2 are centered in all plots.) As analysts, the CERES plot should be smoothed and the resulting smoother used to estimate g_2. Again, the further the points in the CERES plot are from the smooth, the larger the residual sum of squares for a final fit.

The plots in Figure 2 show that the choice of smoother has a large effect of our perception of the amount of noise in the prediction of $g_2(W_2)$. Again, a dedicated data analyst would be able to interactively adjust the smoother and see that for a coarse smooth for $E(W_1|W_2)$, the apparent noise in the prediction of $g_2(W_2)$ is reduced. Although experience has shown that a coarse smooth for $E(W_1|W_2)$ often results in a more accurate display of g_2, the point here is that the user should easily be able to experiment with smoothers and parameters. In this case, experimentation allows us to *see* that it is not the coarseness of the smooth that makes the second CERES plot give a more accurate smooth, but instead it is the fact that the coarser smooth is closer to the truth for small values of W_2. In fact, if we estimate $E(W_1|W_2)$ with a piecewise linear using only two lines, the CERES plot looks virtually the same as the the lower right plot of Figure 2.

3.3 Influential Points

Imagine that in our first example, we had an error in measurement in the observation where X_1 is largest. This point is already a suspected outlier, but imagine that instead of a response value of 15.04, a response of 19.04 was recorded. We fit the same GAMs used in section 3.1, and the plots in Figure 3 are obtained.

The error in measurement actually serves to allow the loess smoother for X_1 to begin to capture the true parabolic relationship between X_1 and Y for positive values of X_1. GAM. Notice that although the estimated functions for the model fit by the Splus command `gam(y ~ lo(x1) + lo(x2))` do not differ much from those in Figure 1, the observations with values of X_2 between 4.2 and 4.6 are fit much better when we have the error in measurement.

This shows that single points may be highly influential in the estimated fit as well as the perception of fit.

4 Interactive Methods

The above examples illustrate that there is a need for interactive methods which will aid the data analyst in understanding the regression. Such methods should:

- allow the user to interactively change the smoothers.
- allow the user to investigate the influence of individual points.
- not be too cumbersome for practical use.

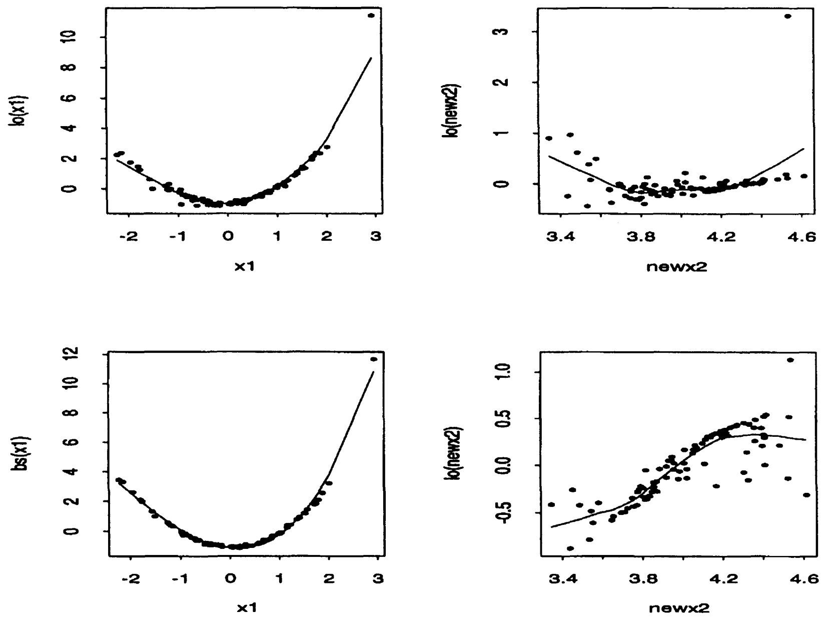

Figure 3: Estimated additive functions from GAM.

In Wetzel (1996), such a set of methods was developed for CERES plots. The estimation of $E(X_1|X_2)$ is displayed and the user may easily change smoothing parameters, and/or the points used in the calculation. A set of reasonable defaults were established, but the analyst was presented with all of the relevant plots. Smoothers can be changed and points deleted and after a few mouse clicks, all of the plots are updated.

An outline of a similar set of methods for GAM is described below. Since the GAM procedure is iterative, we need to be able to monitor the process through all of the iterations. This process is described by the following algorithm (see Hastie and Tibshirani, 1990).

- initialize: $g_1 = g_1^0$, $g_2 = g_2^0$.
- cycle: g_1^i = apply a smoother to the plot of $Y - g_2^{i-1}(X_2)$ versus X_1,
  g_2^i = apply a smoother to the plot of $Y - g_1^i(X_1)$ versus X_2,

- continue until g_1 and g_2 don't change

Interactive methods for GAM should allow the user to see each of the plots which are smoothed. An analyst's perception of the appropriateness of these smoothes will allow that analyst to proceed with another iteration. If the smooth is seen to be inappropriate, a different smoother is chosen, and then another iteration is performed. Similarly, if a point is deemed an outlier, the analyst will take appropriate action and continue with another iteration. The software should keep track of which smoothers were used at which iteration as well as which points were used. At some point, we need to 'continue until g_1 and g_2 don't change.' At this point, the observations which will be used and the smoothers should be fixed. We have not proven a result, but it seems clear that the first few steps of the iterative procedure should have little bearing on convergence results.

In order to illustrate these methods, we return to our first example. Using Xlisp-Stat and an initialization of $g_1^0 = 0$ and $g_2^0 = 0$, we obtain the plots in Figure 4. The ordering of these plots is left to right, and top to bottom. The curves shown in the plots are the smoothes used to estimate g_1^i and g_2^i. The indicated numerical argument for lowess is the value used to call the `lowess` function in Xlisp-Stat. No weighting is used in this example. For example, the plot in the upper right hand corner shows $Y - \overline{Y} - g_1^1(X_1)$ versus X_2, where the smooth shown in the upper left plot is used for g_1^1.

We notice that the lowess smooth used in the first iteration under estimates at the extremes, so in the second iteration we use a 2nd degree polynomial and obtain better estimates for both g_1 and g_2.

The code for such methods is currently underway; the first plot in Figure 4 was created with a command from the keyboard, but the other five plots in Figure 4 were created with a series of mouse clicks. A final mouse click had the iteration continue until a crude convergence criteria was met. The final fits do not appear significantly different from the third row of plots in Figure 4.

5 Discussion

Finally, it should be clear that there is a connection between CERES plots and GAMs. Both find nonlinear relationships between the response and predictors. CERES plots assume that all of the predictors act linearly except for one. GAMs add additional assumptions to the errors. Berk and Booth (1995) compare CERES and GAMS to each other as well as other methods. Also, since at each stage in the iterative process GAMs use partial residual plots, and partial residual plots assume that the relationship between predictors is at most linear, strong nonlinear relationships between the predictors may results in poor GAM performance. An approach to combining these two ideas is being investigated (see Croos-Dabrera, 1994). Implementation of these methods should allow interaction as described above.

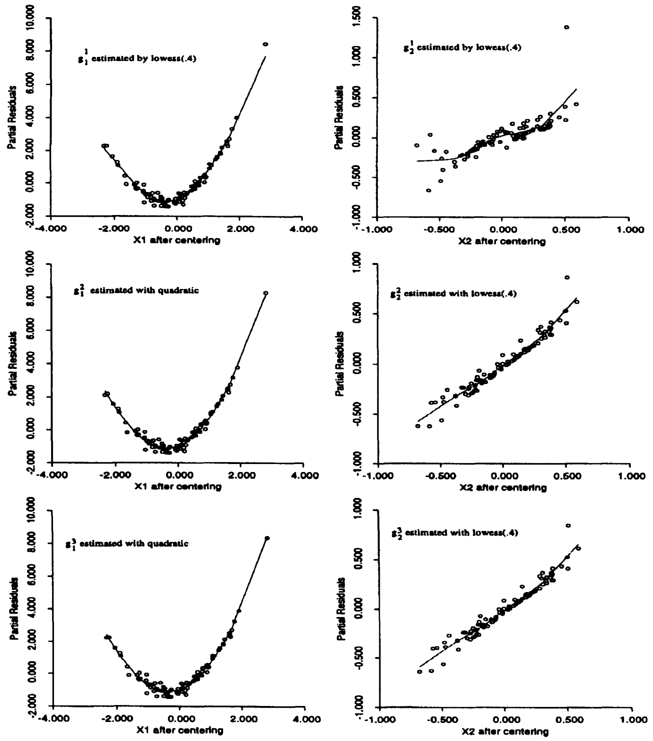

Figure 4: Sequential Partial Residual Plots.

References

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

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

[3] Croos-Dabrera, R. (1994). Graphical analysis of curvature in semiparametric generalized linear models. Ph.D. dissertation, University of Minnesota - School of Statistics.

[4] Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models.* New York: Chapman and Hall.

[5] Tierney, L. (1990). *LISP-STAT.* New York: Wiley.

[6] Wetzel, N. (1996). Graphical data modeling methods using CERES plots. *J. Statist. Comput. Simul.* **54,** 37–44.