Communications for Statistical Applications and Methods
2019, Vol. 26, No. 5, 463–471

https://doi.org/10.29220/CSAM.2019.26.5.463
Print ISSN 2287-7843 / Online ISSN 2383-4757

A robust method for response variable transformations using dynamic plots

Han Son Seo[1,a]

[a]Department of Applied Statistics, Konkuk University, Korea

Abstract

The variable transformations are useful ways to guarantee the functional relationships in the model. However, the presence of outliers may undermine the accuracy of transformation. This paper deals with response transformations in the partial linear models under the existence of outliers. A new procedure for response transformation and outliers detection is proposed. The procedure uses a sequential method for identifying outliers and dynamic graphical methods for an appropriate transformation. The graphical tools make it possible to catch diagnostic information by monitoring the movement of points in the data. The procedure is illustrated with several examples. Examples show that visual clues regarding the optimal transformation, the fittness of the model and the outlyness of the observations can be checked from the series of plots.

Keywords: diagnostics, dynamic plots, outliers, partial linear models, variable transformations

1. Introduction

A statistical model is often characterized by its mean function. However, often the correct form for the mean function does not follow the assumed one. A transformation of variables is needed to achieve an assumed mean function in the transformed scale. We consider a partial linear model in regards to the problem of transformation. A standard linear regression model is a basic tool for analyzing statistical data and is widely used due to its simplicity. In some problems the linear relationship between the response and all covariates are not known. A partial linear model is more flexible by incorporating the nonlinear functional relationship in a general linear model. The model with response transformations in partial linear model is given as,

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

where $Y^{(\lambda)}$ is a transformed response variable, λ is a transformation coefficient, X is a matrix of explanatory variables, h is a curvature function of a explanatory variable Z and ε is an error term. The transformation family used most is the scaled power family defined for positive Y by $Y^{(\lambda)} = (Y^{\lambda} - 1)/\lambda$ when $\lambda \neq 0$ and $Y^{(\lambda)} = \log(Y)$, when $\lambda = 0$. Seo (2009) and Seo and Yoon (2009) suggested a graphical method for capturing the curve and estimating the transformation coefficient in the model (1.1).

In this paper we deal with response transformations in the partial linear models under the existence of outliers. The existence of outliers in the data is a common problem in a statistical analysis. Many

[1] Department of Applied Statistics, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea.
E-mail: hsseo@konkuk.ac.kr

Published 30 September 2019 / journal homepage: http://csam.or.kr

approaches for detecting multiple outliers are suggested in a linear model, for example sequential procedures (Hadi and Simonoff, 1993), high-breakdown methods (Rousseeuw, 1984; Yohai, 1987) and forward searches (Atkinson, 1994). In the discussion of outliers, the difference between outliers and influential observations in estimating transformations should be understood. An influential observation is one whose deletion has a large effect on the transformation estimates. Cheng (2005) suggested a robust method for response transformation against influential observations in a linear model. They used the least trimmed squares estimator and the trimmed likelihood estimator. An outlier is an observation that diverges from an overall pattern formed by the transformed data. Seo *et al.* (2012) suggested several procedures for detecting outliers in the process of response transformation in a linear model. Seo and Yoon (2013) used maximum trimmed likelihood estimators to overcome the bad effects caused by outliers in fitting a partial linear model with response transformation.

This paper presents a graphical procedure for a robust transformation using an outlier detection method and dynamic plots. Section 2 suggests a dynamic graphical method for the transformation and the outlier detection in a partial linear model. The method involves augmented partial residual plot for specifying the curvature and a sequential procedure for detecting outliers. Section 3 provides several examples with artificial data and real data to illustrate the suggested method. Section 4 contains some concluding remarks.

2. A robust method with dynamic plots

Determining the coefficient of optimal response transformation in a partial linear model is difficult to solve analytically. We suggest an exploratory procedure of observing related plots for many transformations. Plots for each transformation need the estimation of a partial linear model and the detection of outliers. Graphical methods are used for the estimation of a partial linear model. Many graphical methods are suggested to specify the curvature in a partial linear model including added variable plot (Chamber *et al.*, 1983, p.272), partial residual plot (Larsen and McCleary, 1972; Weisberg, 2005), augmented partial residual plot (Mallows, 1986) and CERES plot (Cook, 1993). An augmented partial residual plot is known more effective than other graphical methods including inverse response plots (Cook and Weisberg, 1994). Augmented partial residual plots are constructed based on the linear model with X, Z and quadratic terms of Z as covariates, $Y = \rho_0 + X\rho_1 + \phi_1 Z + \phi_2 Z^2 + \text{error}$. The coefficients of the model are estimated by minimizing a convex objective function L, $L_n = (\rho_0, \rho_1, \phi_1, \phi_2) = \sum_{i=1}^{n} L(y_i - \rho_0 - x_i\rho_1 - z_i\phi_1 - z_i^2 1\phi_2)/n$. Augmented partial residual plot for Z is the plot of $e + \hat{\Phi}_1 Z + \hat{\Phi}_2 Z^2$ versus Z. Augmented partial residual plots are expected to depict the curvature h better than partial residual plots which are constructed from the linear model $Y = \rho_0 + X\rho_1\phi_1 Z + \text{error}$. Another graphical method, CERES plots use a model $Y = a_0 + Xa_1 E(X|Z)b + \text{error}$. CERES plots are a larger class of plots including partial residual plot and augmented partial residual plot as special ones. The performance of CERES plots depends on the accuracy of the estimation of $E(X|Z)$.

With a fixed transformation and the estimated curvature in the model (1.1) an outlier detection method designed for linear models can be applied. In this paper a sequential procedure proposed by Hadi and Simonoff (1993) is used, which consists of three steps , constructing a clean set, calculating residuals and testing for outliers. For the construction of initial clean subset Hadi and Simonoff (1993) suggested two methods. The first method fits the model of p explanatory variables with p observations and selects a new set of $(p+1)$ observations corresponding to the $(p+1)$ smallest residuals. An initial clean is obtained by repeating the process until to get int$[(n + p - 1)/2]$ observations. The second method uses the backward-stepping approach (Rosner, 1975; Simonoff, 1984, 1988) applying single linkage clustering (Hartigan, 1981). Hadi and Simonoff (1993) showed that the first method was

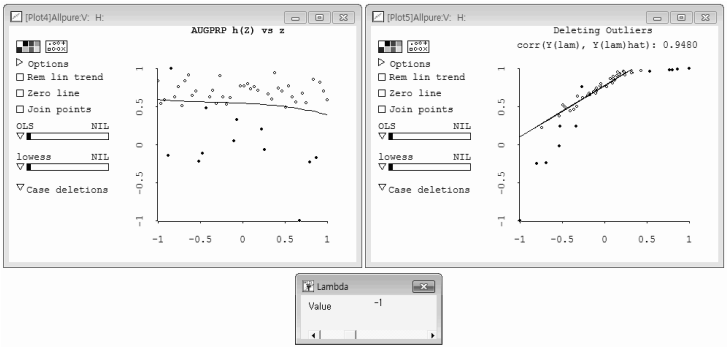


Figure 1: *Dynamic plots: An augmented partial residual plot, a forward response plot and a lambda control-slider.*

more effective. The main algorithm for detecting outliers is as follows. Given a clean subset M of size s a diagnostic measure d_i are calculated, which are the internally studentized residuals for the observations included in a clean subset and the scaled prediction errors for others.

$$d_i = \begin{cases} \dfrac{\left(y_i - x_i^T \hat{\beta}_M\right)}{\hat{\sigma}_M \sqrt{1 - x_i^T \left(X_M^T X_M\right)^{-1} x_i}}, & \text{if } i \in M, \\ \dfrac{\left(y_i - x_i^T \hat{\beta}_M\right)}{\hat{\sigma}_M \sqrt{1 + x_i^T \left(X_M^T X_M\right)^{-1} x_i}}, & \text{if } i \notin M. \end{cases}$$

The set of outlier candidates is determined by the absolute value of d_i, $|d_i|$. If we let $|d|_{(j)}$ be the j^{th} order statistic of the d_j in ascending order the outlier candidates are the $(n - s)$ observations corresponding to $|d|_{(s+1)}, \ldots, |d|_{(n)}$. The test for outlyingness is done by comparing the statistic $|d|_{(s+1)}$ with $t_{(\alpha/2(s+1),s-k)}$. If the null hypothesis is rejected a new clean subset that consists of the first $(s + 1)$ ordered observations is formed and the process is repeated.

For an exploratory analysis to figure out the transformation, the curvature and outliers simultaneously animation techniques in plotting is usually used (Seo and Yoon, 2009). We use augmented partial residual plots and forward response plots which are animated as λ changes. The plots contain diagnostic measures calculated from the data excluding detected outliers. Visual clues regarding the fittness of the model and the outlyingness of the observations are checked from the series of plots to determine the optimal transformation. The suggested procedure is summarized as follows.

- Fix a value of λ and estimate $h(Z)$ by an augmented partial residual plot.
- Use Hadi-Simonoff 's procedure with $Y^{(\lambda)}$, X, and $\widehat{h(z)}$ to identify outliers.
- Calculate the fitted values $\hat{Y^{(\lambda)}}$ with a clean subset of $Y^{(\lambda)}$, X, and $\widehat{h(z)}$.
- Draw an augmented partial residual plot and a forward response plot ($\hat{Y}^{(\lambda)}$ vs. $Y^{(\lambda)}$) in which clean cases and outliers are marked with different symbols.
- Change λ smoothly and check the fitness of the model from a forward response plot and capture a curve $h(Z)$ from augmented partial residual plot.

Table 1: Generated data from the model (3.1)

Case #	X_1	X_2	Z	Y	Case #	X_1	X_2	Z	Y
1	5.54	5.11	−1.00	123059	26	5.30	4.35	0.02	16334
2	5.05	5.98	−0.96	154554	27	5.37	4.34	0.06	17862
3	4.12	6.38	−0.92	85194	28	5.15	4.56	0.10	17025
4	4.55	7.61	−0.88	367573	29	5.39	4.46	0.14	18966
5	4.44	4.88	−0.84	24570	30	5.01	5.03	0.18	24532
6	3.99	4.55	−0.80	9388	31	8.06	4.76	0.22	368547
7	5.57	5.13	−0.76	76244	32	4.25	4.31	0.27	5744
8	6.21	5.31	−0.71	172704	33	6.00	4.98	0.31	65102
9	4.56	4.66	−0.67	15879	34	4.22	5.55	0.35	19864
10	5.57	3.41	−0.63	11464	35	6.77	3.95	0.39	55872
11	4.50	5.84	−0.59	44631	36	5.00	5.00	0.43	24839
12	4.39	4.95	−0.55	15226	37	4.49	5.33	0.47	22878
13	4.45	3.73	−0.51	4719	38	3.76	6.51	0.51	38361
14	6.04	6.61	−0.47	392211	39	4.81	5.19	0.55	29470
15	5.03	3.70	−0.43	7701	40	4.52	4.45	0.59	10586
16	4.21	5.70	−0.39	24558	41	5.64	5.51	0.63	100145
17	4.77	5.06	−0.35	21159	42	5.74	1.80	0.67	2926
18	4.05	5.67	−0.31	17971	43	4.89	5.64	0.71	60870
19	6.42	4.14	−0.27	44608	44	5.42	5.67	0.76	101969
20	4.97	4.19	−0.22	10798	45	6.70	6.43	0.80	984019
21	4.18	5.89	−0.18	24561	46	4.20	4.48	0.84	12227
22	5.13	5.29	−0.14	30737	47	3.38	4.32	0.88	4709
23	3.62	5.39	−0.10	7731	48	6.27	4.70	0.92	133900
24	6.09	5.68	−0.06	132488	49	4.28	3.73	0.96	7554
25	5.93	3.47	−0.02	11718	50	4.45	6.24	1.00	122214

- Stop at which the clean cases in a forward response plot show a linear trend.

During the procedure the curve $h(Z)$ is not estimated again with a clean subset because $h(Z)$ does not depends on the existence of outliers under the mean-shifted outlier model. For performing suggested procedure customized dynamic plots are coded by using XLISP-STAT (Tierney, 1990).

Figure 1 shows an augmented partial residual plot and a forward response plot and a slider for changing λ from -2 to 2. Variables in the plot are standardized so all observations to be between -1 and 1. Outliers are symbolized with solid dot (•) in two plots and a correlation coefficient of variables in the forward response plot are also displayed for reference.

3. Examples

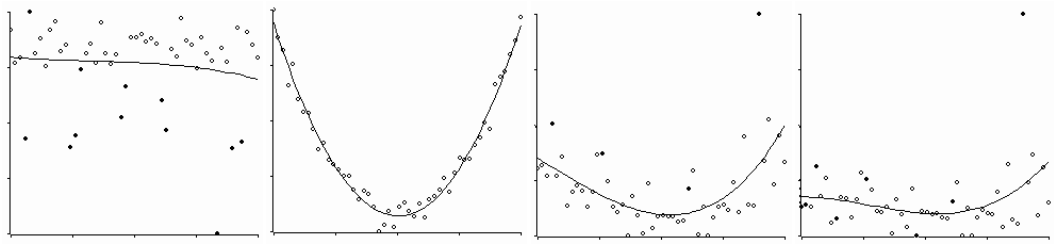
Example 1. Artificial data without outliers

A dataset shown in Table 1 is artificially generated according to the model,

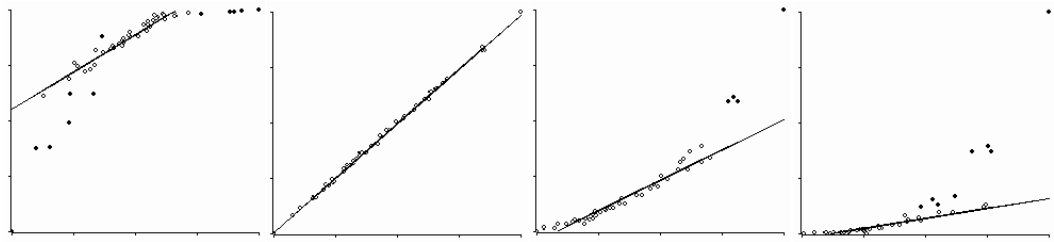
$$Y = \exp\left(X_1 + X_2 + Z^2 + \varepsilon\right), \tag{3.1}$$

where variables X_1 and X_2 follows the bivariate normal distribution $(\, x_1 \;\; x_2 \,)^T \sim \text{MVN}\left(\binom{5}{5}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)$, Z has equally spaced values between -1 and 1 and ε is a normal random variate with mean 0 and standard deviation 0.05. Figure 2 contains dynamic plots for selected values of λ. Some plots reflect the occurrence of a swamping effect.

For example, when $\lambda = 0.5$ four observations 4, 14, 31, 45 are detected as outliers. Dynamic forward response plots after removing detected outliers indicate that when $\lambda = 0$ there is no outlier in the data and the model fits well ($R^2 = 0.98^2$). This results coincides with the model (3.1) from which



(a) Augmented partial residual plots



(b) Forward response plots

Figure 2: *Dynamic plots with* $\lambda = -1, 0, 0.5, 1$ *(from left).*

the dataset is generated. We also see that when $\lambda = 0$ the augmented partial residual plot captures curvature function successfully.

Example 2. Artificial data with outliers

Some observations in the Table 1 are modified to include outliers. Three outliers are planted at the 10^{th}, 20^{th}, and 30^{th} observations by changing their y values as 64, 98, 531 respectively. For this contaminated data dynamics plots are constructed with or without performing a test for outliers. Figure 3 contains forward response plots for several values of λ. Forward response plots without performing outlier-test suggest the optimal transformation as $\lambda = 0.5$ wrongly. But the forward response plot with outlier-test shows a strong linear trend ($R^2 = 0.97^2$) when $\lambda = 0$ and detected the 10^{th}, 20^{th}, and 30^{th} observations as outliers.

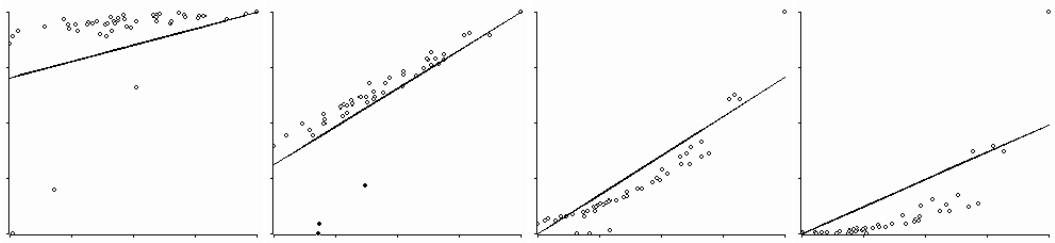
Example 3. A real data (Nitrogen in lakes data)

Nitrogen in lakes data (Atkinson and Riani, 2000, p.297) include 29 observations on the amount of nitrogen in US lakes with the variables, X_1: average influent nitrogen concentration, X_2:water retention time and Y: mean annual nitrogen concentration. Stromberg (1993) and Atkinson and Riani (2000) analyzed data using the following nonlinear model and diagnosed the 10^{th} and the 23^{rd} observations as outliers.

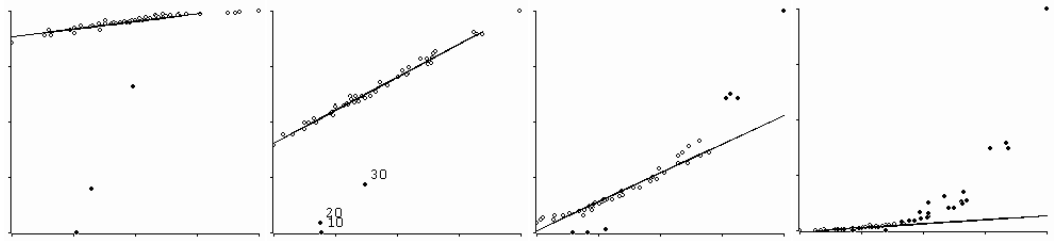
$$y_i = \frac{x_{x_{1i}}}{1 + \beta_1 x_{2i}^{\beta_2}} + \varepsilon_i, \quad i = 1, \ldots, 29.$$

Stromberg (1993) fit the model using least median of squares estimate and MM estimate. The fitted model yielded $R^2 = 0.66^2$ after removing the points 10 and 23 as outliers.

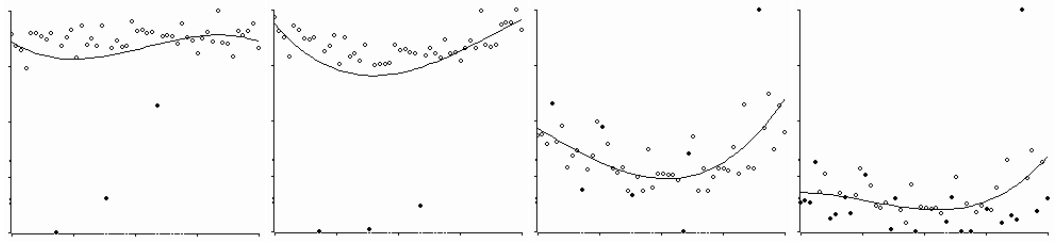
We fit a model (1.1) and conducted the dynamic graphical procedure. Augmented partial response plots and forward response plots are shown in Figure 4. Judging from forward response plots the candidates of optimal estimate of λ is 1 or 0. When $\lambda = 1$ the forward response plot shows a strong linear relationship and the model detected more than 15 points as outliers. When $\lambda = 0$, however,



(a) Forward response plots without performing outliers-test



(b) Forward response plots with performing outliers-test



(c) Augmented partial residual plots corresponding to forward response plots in (b)

Figure 3: *Dynamic plots with* $\lambda = -0.5, 0, 0.5, 1$ *(from left).*

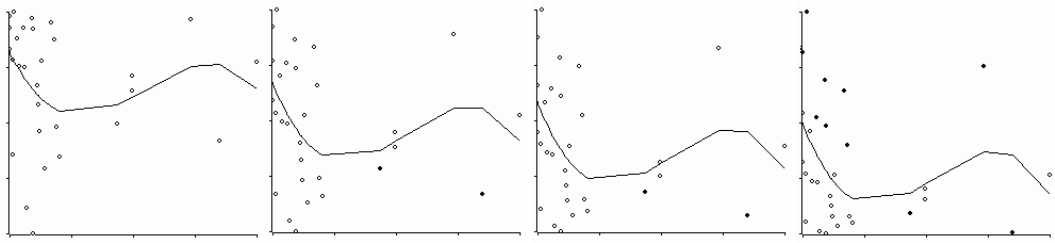
only two points 10, 23 are detected as outliers and the estimated model fits the data well ($R^2 = 0.70^2$). R^2 is lower when $\lambda = 0$, however, the log-transformation is more appropriate because more than one third observations in the data is too much to exclude. A linear model can also be used for the analysis of nitrogen in lake data since the curve in the partial linear model is estimated parametrically. The same results is achieved by using maximum trimmed likelihood estimators (Seo and Yoon, 2013).

Example 4. Artificial data (for exploratory analysis)

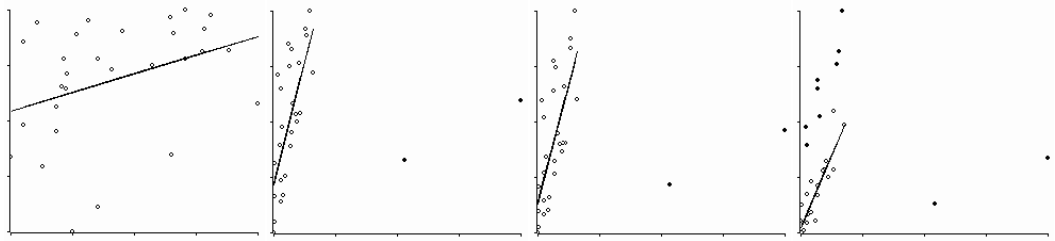
This example is to illustrate the effectiveness of the exploratory procedure using dynamic plots. Eighteen observations are generated from the model,

$$\sqrt{Y} = X_1 + X_2 + \log Z + \varepsilon, \tag{3.2}$$

where covariates X_1, X_2, error term ε follow the same distributions as ones in the Example 1, and variable Z has equally spaced values between 1 and 2. Two outliers, the 19^{th} and 20^{th} observations, are generated from the mean shifted model defined by adding 2 to the model (3.2). Table 2 contains the generated data. Applying the suggested method to the generated data, the optimal transformation is estimated as $\lambda = 1$ and 8 observations $(2, 4, 5, 6, 14, 16, 18, 19)$ are detected as outliers ($R^2 = 0.92^2$). Figure 5 shows the augmented partial residual plot and the forward response plot with $\lambda = 1$.



(a) Augmented partial residual plots



(b) Forward response plots

Figure 4: *Dynamic plots with* $\lambda = 1, 0, 0.5, 1$ *(from left).*

Table 2: Generated data from the model (3.2)

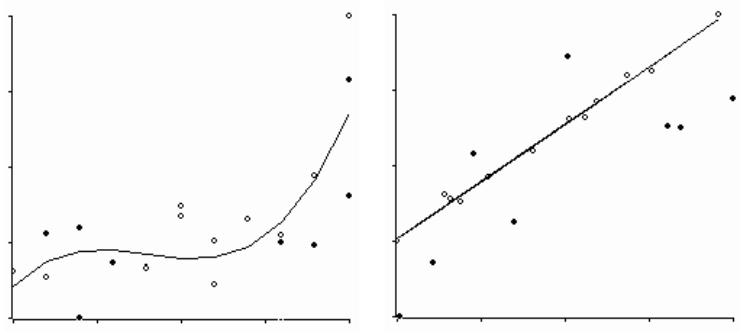
Case #	X_1	X_2	Z	Y	Case #	X_1	X_2	Z	Y
1	7.3	4.4	1.0	138.9	11	4.6	4.7	1.6	97.5
2	4.7	5.7	1.1	117.1	12	6.0	5.8	1.6	131.7
3	4.2	6.0	1.1	100.6	13	3.4	4.8	1.7	80.9
4	5.6	7.1	1.2	157.8	14	3.2	3.4	1.8	49.3
5	5.3	4.5	1.2	79.0	15	4.0	6.0	1.8	107.4
6	4.2	3.9	1.3	71.9	16	6.2	4.7	1.9	128.3
7	5.5	5.3	1.4	118.3	17	5.5	6.0	1.9	151.9
8	6.3	5.2	1.4	132.3	18	5.3	5.8	2.0	139.9
9	7.2	4.2	1.5	150.1	19	4.7	4.0	2.0	128.9
10	4.5	4.5	1.5	98.6	20	5.2	5.3	2.0	174.9

From the augmented partial residual plots for many values of λ it seemed that the observations 19 and 20 were masked. Figure 6 shows augmented partial residual plots for selected values of λ. When $\lambda = -1$ observations 19 and 20 (marked as x in the plot) are masked by the detected outlier, observation 14. Observations 19 and 20 seemed out of the trend in augmented partial residual plots with $\lambda = 0$ and 0.5.

Considering these information two observations 19 and 20 are deleted from the data and the optimal value of λ is now estimated as 0.5 ($R^2 = 0.93^2$). Figure 7 shows the forward response plots with $\lambda = 0.5$ after deleting points 19 and 20.

4. Concluding remarks

The problem of outliers detection and response transformation in a partial linear model is difficult to handle analytically. An exploratory procedure is suggested as a unified method to solve the problem. A procedure combining outlier detection methods and graphical techniques is proposed to provide an appropriate variable transformation robust to outliers. Diagnostic measures are calculated from



(a) Augmented partial residual plot (b) Forward response plot

Figure 5: *Dynamic plots with* $\lambda = 1$.

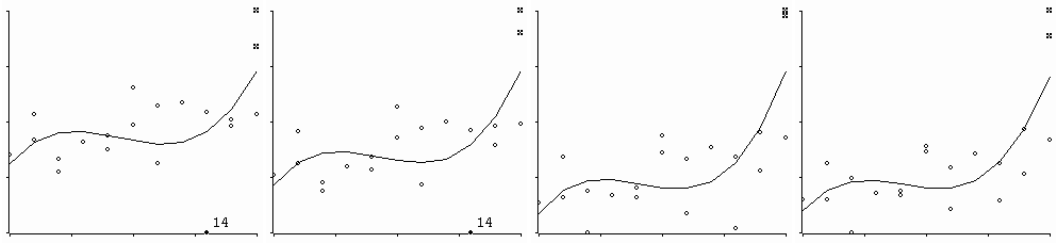
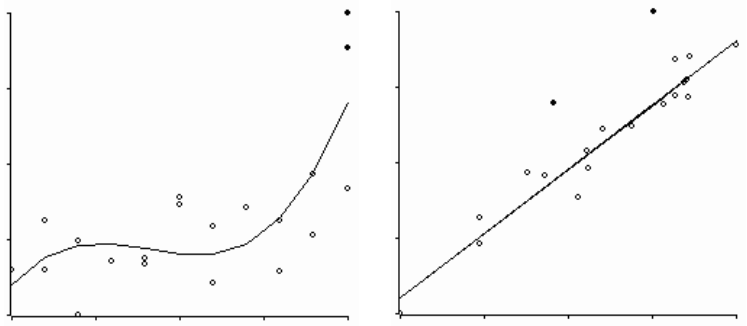


Figure 6: *Augmented partial residual plots with* $\lambda = -1, -0.5, 0, 0.5$ *(from left). Two x-marked points correspond to observation 19 and 20 which are not detected as outliers.*



(a) Augmented partial residual plot (b) Forward response plot

Figure 7: *Dynamic plots with* $\lambda = 0.5$ *after deleting observations 19 and 20. The deleted points 19 and 20 are marked as* $\bullet$.

the data excluding outliers and are plotted. Examples show that it is possible to examine the role of observations in the diagnostic point of view through the dynamic plots.

The suggested procedure uses a sequential method to detect outliers, an augmented plot for estimating a curvature and the forward response plot for observing fitness. The performance of the proposed procedure depends on outlier detection methods and curvature estimation. If a dataset is large and has enough repeated observations to estimate $E(X|Z)$ accurately, CERES plots can be a better tool for estimating a curvature in a partial linear model.

Acknowledgements

This paper was supported by Konkuk University in 2018.

References

Atkinson AC (1994). Fast very robust methods for the detection of multiple outliers, *Journal of the American Statistical Association*, **89**, 1329–1339.

Atkinson AC and Riani M (2000). *Robust Diagnostic Regression Analysis*, Springer, New York.

Chambers JM, Cleveland WS, Kleiner B, and Tukey P (1983). *Graphical Methods for Data Analysis*, Duxbury Press, Boston.

Cheng T (2005). Robust regression diagnostics with data transformations, *Computational Statistics and Data Analysis*, **49**, 875–891.

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

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

Hadi AS and Simonoff JS (1993). Procedures for the identification of multiple outliers in linear models, *Journal of the American Statistical Association*, **88**, 1264–1272.

Hartigan JA (1981). Consistency of single linkage for high-density clusters, *Journal of the American Statistical Association*, **76**, 388–394.

Larsen WA and McCleary SJ (1972). The use of partial residual plots in regression analysis, *Technometrics*, **14**, 781–790.

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

Rosner B (1975). On the Detection of Many Outliers, *Technometrics*, **17**, 217–227.

Rousseeuw PJ (1984). Least median of squares regression, *Journal of American Statistical Association*, **79**, 871–880.

Seo HS (2009). A visual procedure for optimal response transformation and curvature specifications, *Optimization and Engineering*, **10**, 301–312.

Seo HS, Lee GY, and Yoon M (2012). Robust response transformation using outlier detection in regression model, *The Korean Journal of Applied Statistics*, **25**, 205–213.

Seo HS and Yoon M (2009). A dynamic graphical method for transformations and curvature specifications in regression, *The Korean Journal of Applied Statistics*, **22**, 189–195.

Seo HS and Yoon M (2013). Regression diagnostics for response transformations in a partial linear model, *Journal of the Korean Data & Information Science Society*, **24**, 33–39.

Simonoff JS (1984). The calculation of outlier detection statistics, *Communications in Statistics, Part B-Simulation and Computation*, **13**, 275–285.

Simonoff JS (1988). Detecting outlying cells in two-way contingency tables via backwards-stepping, *Technometrics*, **30**, 339–345.

Stromberg AJ (1993). Computation of high breakdown nonlinear regression parameters, *Journal of the American Statistical Association*, **88**, 237–244.

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

Weisberg S (2005). *Applied Linear Regression*, Wiley, New York.

Yohai VJ (1987). High breakdown-point and high efficiency robust estimate for regression, *The Annals of Statistics*, **15**, 642–656.

Received February 24, 2019; Revised June 10, 2019; Accepted June 11, 2019