Journal of the KSQC
Vol.21, No.1, June 1993

1993년 6월 韓國品質管理學會誌 제21권 제1호

A Study on Applications of Regression Diagnostic Method to Technometrics, and the Statistical Quality Control *

Soon - Kwi Kim **

ABSTRACT

This article is concerned with procedures for detecting one or more outliers or influential observations in a linear regression model. A test procedure, based on recursive residuals is proposed and developed

The power of the test procedure to identify one or more outliers is investigated through simulation, and its relevance to the number and configuration of the outlier.

1. Introduction

Consider the standard form of the linear regression model

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

* This paper was supported in part by NON DIRECTED RESEARCH FUND, Korea Research Foundation. 1991

** Kangnung University

where $\underline{Y} = (y_1, \cdots, y_n)'$ is an $n \times 1$ vector of observations on the dependent variable ; $X = (\underline{x}_1', \cdots, \underline{x}_n')'$ is an nxp matrix of explanatory variables, possibly including the intercept term ; $\underline{\beta}$ is a $p \times 1$ vector of unknown parameters ; and $\underline{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)'$ is an $n \times 1$ vector of independent normal random variables with mean 0 and unknown variance σ^2. For $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$, the ordinary least squares(OLS) estimator of $\underline{\beta}$, all information about outliers is contained in the vector of regression residuals

$$\underline{e} = \underline{Y} - X\hat{\underline{\beta}} = (I-H)\underline{Y},$$

where $H = (h_{ij}) = X(X'X)^{-1}X'$ and under the null hypothesis

$$\underline{e} \sim N(\underline{0}, (I-H)\sigma^2),$$

a degenerate distribution, since the idempotent matrix $(I-H)$ has rank $n-p$. One should note that even though all information about outliers is contained in $\underline{e}$, these ordinary residuals do have some major defects they are not independent, and in general they do not all have the same variance. This results in the effect of structural change, for example, being smeared over all the residuals. It also implies that the distribution of the residual is dependent on the particular design matrix under consideration.

A scaled varsion of e_i can be defined as

$$r_i = \frac{e_i}{s(1-h_{ii})}, \quad i = 1, \cdots, n$$

where $s^2 = \underline{e}'\underline{e} / (n-p)$ is the residual mean square estimate of σ^2. The r_i is usually called the studentized resida1. Another version of the e_i is often called the externally studentized residual and is defined as

$$r_i = \frac{e_i}{s_{(i)}(1-h_{ii})^{1/2}}, \quad i = 1, \cdots, n$$

where $s^2_{(i)}$ is the residual mean square estimate of σ^2 without the ith observation. Noting that $r_i^* = [(n-p-1)/(n-p-r_i^2)]^{1/2}$, we see that r_i^* is a monotonic transformation of r_i. While the above scaled versions of the residuals e_i have approximately unit variance, they are still correlated.

The detection and examination of outliers that are influential are the object of the outlier detection procedures, since such outliers have a catastrophic effect on the regression.

Outliers that are not influential may be retained in the data set without changing the regression equation greatly. Of course, different observations may be influential in different calculations. Cook's D is a well-known measure of influence of the ith observation on the center of the confidence ellipsoids or, equivalently, on $\underline{\beta}$. It is defined as

$$D_i = \frac{(\hat{\underline{\beta}}-\hat{\underline{\beta}}_{(i)})'(X'X)(\hat{\underline{\beta}}-\hat{\underline{\beta}}_{(i)})}{pS^2}$$

where $\hat{\underline{\beta}}_{(i)}$ is the ordinary least squares estimate of $\underline{\beta}$ when the ith observation is omitted.

An alternative class of measures of the influence of the ith observation is based on the change in the volume of the confidence ellipsoid when the ith observation is omitted. On the other hand, omitting an observation with a large residual will result in a large reduction in the residual sum of squares, SSE. The influence of the ith observation can be measured by combining these two ideas, one such measure is, suggested by Andrews and Pregibon (1978),

$$AP_i = \frac{SSE_{(i)}\det(X_{(i)}'X_{(i)})}{SSE\det(X'X)} = \frac{\det(Z_{(i)}'Z_{(i)})}{\det(Z'Z)}$$

where the augmented matrix Z is formed by adding $\underline{y}$ to X. AP_i measures the relative change in $\det(Z'Z)$ due to the omission of the ith observation. Omitting an observation that is far from the center of the data will result in a large reduction in the determinant and thereby a large increase in the volume. Hence, small values of AP_i are associated with deviant or influential observations. Regardless of which is actually the case, it is desirable to isolate subsets of the observations producing small AP_i for further scrutiny.

Recursive residuals have frequently been suggested for testing model fit and model assumptions(see Brown, Durbin and Evans(1975), Hawkins(1980), Galpin and Hawkins (1984)) illustrate the use of plots of the CUSUMS of the recursive residuals and plots of the CUSUMS of the square roots of the absolute values of the standardized recursive residuals to check the model assumptions mentioned earlier, to investigate the effect of an omitted variable and to detect outliers. A new procedure, introduced by Kianifard and Swallow (1989), uses recursive residuals, calculated on observations that have been ordered according to their studentized residuals, values of Cook's D, or another regression diagnostic of the user's choice. Like conventional regression diagnostics, Hawkins(1991) proposes analogous single-case diagnostics for use with recursive fitting. Further more, since recursive fitting focuses on the compatibility of various subregressions with the full regression, cumulative measures that assess the leverage, mean compatibility, and influence of all excluded cases

have been suggested. (Hawkins(1991)). The recursive residuals are defined in section 2, where a simple interpretation of them is also given.

In this article, we propose a procedure based on recursive residual for identifying outliers or influential observations in a linear regression model. The procedure is presented in section 3, and its properties investigated in a simulation study described in section 4. Finally, conclusions and summaries are listed in section 5.

2. Recursive Residuals

Let $\underline{b}_r$ be the least squares estimate of $\underline{\beta}$ based on the first r observations. Let X and $\underline{Y}$ be partitioned accordingly, so that $X_r' = (\underline{X}_1, \cdots, \underline{X}_r)$ and $Y_r' = (y_1, \cdots, y_r)$. Assume that $X_r'X_r$ is non-singular. Then $\underline{b}_r = (X_r'X_r)^{-1}X_r'\underline{Y}_r$, and the recursive residuals can be defined as in two algebraically equivalent forms

$$w_r = \frac{y_r - \underline{x}_r'\underline{b}_{r-1}}{\sqrt{1+\underline{x}_r'(X_{r-1}'X_{r-1})^{-1}\underline{x}_r}} \quad , \quad r = p+1, \cdots, n$$

and

$$w_r = \frac{y_r - \underline{x}_r'\underline{b}_r}{\sqrt{1-\underline{x}_r'(X_r'X_r)^{-1}\underline{x}_r}}$$

Recursive residuals are not defined for cases 1, $\cdots$, p since the reduction the sample to fewer than p points would give a singular design matrix. Brown, Durbin and Evans suggest using the first p data points as the base set, but this is by no means the only sensible choice. One could, in fact, calculate them using any k of the n observations as a basis. The name recursive residual derives from the fact that w_r can be obtained from w_{r-1} by means of an updating formula.

On the assumptions that the errors are identically and independently distributed as $N(0, \sigma^2)$ the recursive residuals are also independently and identically distributed as $N(0, \sigma^2)$ so that if the model assumptions are satisfied, the normal probability plot should show a straight line through the origin.

The calculation of the recursive residuals may appear to be a time-consuming operation, involving the fitting of $n-p$ regressions, but the computations are performed much more efficiently using the following updating formulas(Brown et. al., 1975) :

$$(X_r'X_r)^{-1} = (X_{r-1}'X_{r-1})^{-1} - dd'/(1+\underline{x}_r'\underline{d}),$$

$$\underline{b}_r = \underline{b}_{r-1} + (X_r'X_r)^{-1}\underline{x}_r(y_r - \underline{x}_r'\underline{b}_{r-1})$$

$$S_r = S_{r-1} + w_r^2, \quad r = p+1, \cdots, n$$

where $\underline{d} = (X_{r-1}'X_{r-1})^{-1}\underline{x}_r$ and S_r is the residual sum of square based on r observations.

It makes their use so attractive that the interpretation of the recursive residuals as showing the effect of successively deleting points from the data set, in addition to their propert of independence. Because of this interpretation, recursive residuals are very flexible, and they can be used directly as the basis for diagnostics. If there are any misfits from the model assumptions, then all of the ordinary residuals may be affected by it. If the misfits from model is confined to the portion of the data set, then all other irrelevant recursive residuals will not be affected by the departure. This may be a base-line for recursive residuals to be able to be used to detect any model misspecifications that may be hard to identify with diagnostics based on the OLS residuals. And, they would be seen to have potential for the study of outliers, although on progress on this front is evident. There is a major difficulty in that the labelling of the observations is usually done at random or in relation to some concomitant variable rather than adaptively in response to the observed sample values.

Recursive residuals have been used by Brown et. al.,(1975) in testing for structural change over time, and by Harvey and Collier(1977) in testing for possible model misspecifications. Galpin and Hawkins(1984) proposed the use of recursive residuals in graphical procedures in checking the model assumptions of normality, homoscedasticity and so on. Du Toit, Steyn and Stumpf(1986) provided programs for calculating and plotting recursive residuals : they uses PROC MATRIX and PROC PLOT in SAS. Kianifard and Swallow(1989) suggested using recursive residuals, calculated on adaptively-ordered observations for the detection of outliers. More recently, Hawkins(1991) proposes single-case diagnostics analogous to conventional regression case diagnostics for use with recursive fitting. Further more, cumulative measures that a ssess the leverage, mean compatibility, and influence of all excluded cases are also defined and researched by him.

3. The Test Procedure

For a given set of n observations, there are $n!/p!$ different sets of recursive residuals. Which set is actually computed depends on the result of two connected decisions. The first concerns which p observations should be used to form the basis : the second concerns now the remaining $n-p$ observations should be ordered.

The recursive residual w_i provides a measure of outlyingness but for formal assessment needs to be studentized. By studentizing it using s_{i-1}, Hawkins(1991) suggests the measure

$$t_i = \frac{w_i}{s_{i-1}}$$

for the detection of outliers where s_{i-1}^2 is the usual error mean square estimate of σ^2 based on the first $i-1$ observations. The measure t_i follows a t-distribution with $i-p-1$ df and the test based on t_i is optimal for the compatibility of case i with its predecessors (Hawkins 1980). To remedy different degrees of freedom(df) of the measure t_i, he proposes a measure

$$u_i = \pm \frac{8v+1}{8v+3}(v \log_e [1+t_i^2/v])^{1/2}$$

by use of an approximating normalizing transformation, where $v = i-p-1$ is the df of s_{i-1} and u_i takes the sign of t_i. The fact that u_i approximates $N(0, 1)$ distribution closely provides an easy check for the detection of outliers from the regression of its predecessors.

We suggest the following procedure for ordering the observations, and for calculating recursive residuals and test statistics :

1. Compute values of a proper regression diagnostic(e. g., the studentized residual or Cook's D) for each of the n observations, when fitted the regression model to the data.
2. Order the observations according to the values of the chosen diagnostic.
3. Use the first p observations in the ordered set to from the basis.
4. Compute recursive residuals w_j for $j = p+1, \cdots, n$.
5. Compute the statistics w_j / s_{j-1}, for $j = p+2, \cdots, n$.
6. Calculate the statistics u_j, $j = p+2, \cdots, n$, comparing the computed values with values of $N(0, 1)$

The recursive residuals w_j are iid $N(0, \sigma^2)$ random variables. Hence the statistics w_j / s_{j-1} would have an exact t distribution because s_{j-1}, an estimate of σ^2, was independent of w_j. However, when the observations are adatively ordered by the ordering variable which is not independent of the recursive residuals, the normality of them will be voided.

We point out here two important aspects in regard to ordering the observations. First, outliers or influential observations can be expected to appear late in the sequence of recursive residuals, so the w_j for data points that precede them will not be affected by them, reducing the potential for masking and swamping. Second : by ordering the observations adaptively, outliers will not appear among the first p ordered observations, that is, the

ordering yields a clean basis set for calculating recursive residuals.

We consider three diagnostics according to which the observations could be ordered These diagnostics represent different classes of regression diagnostics. The studentized residual, r_i, was picked up because it is more widely used through statistical packages Cook's D and AP_i are well known, but represent a different class of diagnostics. (Kianifard and Swallow 1989).

4. Simulation Results

We now present simulation results (i) to illustrate the gains in power with ordering, and (ii) to evaluate the performance of the proposedtest procedure with three alternative ordering variables. We adapted a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $n = 25$ in all simulations discussed here. Of course, the residuals are not influenced by the particular values of β_0 and β_1 used : we set $\beta_0 = 0$, $\beta_1 = 1$. The $x's$ were generated as uniform (0, 1) variables multiplied by 15. The errors ε_i were generated as $N(0, 1)$ random variables. We made necessary modifications according to various magnitude and location of outliers as described below. All results are based on 1,000 simulated samples of $n = 25$ each, with the $x's$ are based $\varepsilon's$ regenerated after every 100 samples to avoid possible periodicity of those variables. The same 10 sets of $x's$ and $\varepsilon's$ are used throughout all simulations. For a given case, alternatives are compared with the same set of 1,000 samples. The size of the test is $\alpha = 0.05$ throughout.

4. 1 Gains in Power with Ordering

In Table 1, the performance of the recursive procedure is summarized according to whether or not the observations are ordered by a diagnostic measure. In case the sample was to be not ordered, a single outlier, two or three outliers were created as follows. Three random observation numbers were selected from number 3−25. If there were two outliers in the data set, for example, the first two random observation numbers would be used. This ensured that the outliers would be tested, not be part of the bases set. An amount δ was then added to the two generated x-values for that observations, respectively in place of a simulated error term : so using $\delta = (3, 3)$, for example, is equivalent to placing the two outliers 3 standard deviations above the line. In cases where the data set would be ordered by a diagnostic measure, the two outliers, for example, were created by adding δ to the 24th and 25th generated x-value in place of a simulated error.

Table 1.

PROCORR and PROINC when up to three outliers were planted at distances δ_i from the true line, and the observations not ordered, and ordered by the Studentized residuals, or Cook's D or AP_i.

Outlier pattern (δ_i's)	Not ordered		Studentized		D_i		Descending AP_i	
	PROCORR	PROINC	PROCORR	PROINC	PROCORR	PROINC	PROCORR	PROINC
(2.5)	.547	.034	.895	.226	.876	.216	.910	.176
(3)	.742	.031	.986	.226	.980	.217	.985	.173
(4)	.883	.027	1.000	.223	1.000	.219	1.000	.165
(3, 3)	.592	.023	.947	.217	.909	.203	.954	.159
(3, 3, −3)	.533	.016	.924	.214	.872	.201	.932	.145

The entries in Table 1 and Table 2 are defined as follows :
PROCORR = the proportions of correctly identified outliers,
PROINC = the proportions of good observations (inliers) incorrectly identified as outliers.

The results show that when ordering is used, both the power to detect outliers and PROINC increase highly for the cases shown. Of course, if we had not prevented outliers in unordered data from falling into the basis set where they would have been untested and thus undetected, the increase in PROCORR with ordering would have been far larger. Also it is worth noting when ordering is used, the test statistic is very sensitive to the detection of outliers shown.

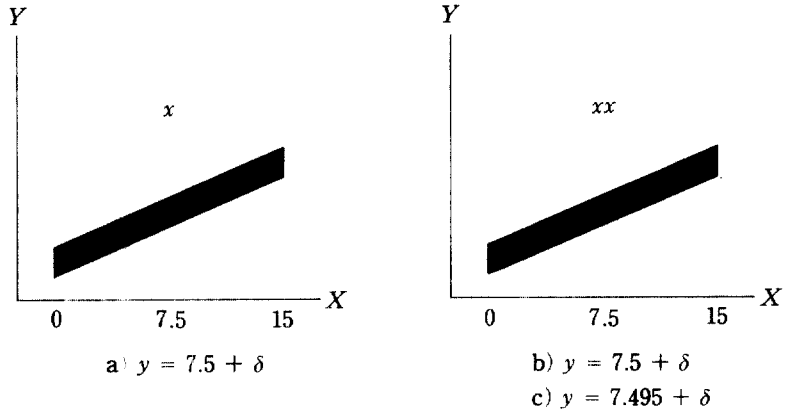
4. 2 Power under Outlier Patterns and the Choice of Ordering Variables

Figure 1 shows various outlier patterns that might be of interest. Inliers are assumed to lie in a box represented by the parallelogram, and the symbol(x) points out an outlier. Table 2 summarizes the results of the simulations for each configuration of data set in Figure 1 in turn. Each outlier is created by adding a quantity δ to the x-values indicated in Figure 1. The performance of the recursive procedure is then summarized in Table 2 for increasing δ. The entries in the table were as defined before.

In Figure 1(a), the outlier occurs near the center of the independent variable. It is worth nothing that the power of the test procedure gets nearly close to 1 for δ as small as 3 and PROINC also increases highly for the cases shown. Figure 1(b) has both outliers at the center of the range and with the same values of δ. This causes serious masking effect. In the cases considered in Figure 1(a)−1(b), the choice of the ordering variable does not make an appreciable difference in the power of the recursive procedure. It also has to be pointed out that the procedure is very sensitive to detecting outliers and nearly avoids the joint

effects of groups of observations.

In practice, any one of the ordering statistics can be used to reorder the observations before computing recursive residuals.



a) $y = 7.5 + \delta$

b) $y = 7.5 + \delta$

c) $y = 7.495 + \delta$

Figure 1. Outlier Patterns for Table 2

Table 2.

Outlier pattern (δ_i's)	Studentized		D_i		Descending AP_i	
	PROCORR	PROINC	PROCORR	PROINC	PROCORR	PROINC
(a) 2.5	.945	.227	.982	.209	.931	.177
3	.998	.227	.998	.195	.996	.172
4	.983	.226	1.000	.217	.998	.166
(b) (2.5, 2.5)	.824	.217	.909	.192	.796	.165
(3, 3)	.971	.218	.979	.202	.966	.160
(4, 4)	1.000	.210	1.000	.200	1.000	.143

5. Summary and Conclusions

Recursive estimation is a technique for updating estimates of regression coefficients with addition of each observation and produces recursive residuals that are uncorrelated with zero mean and constant variance. They are particularly effective for diagnosis when the

assumptions of the regression model do not hold for the full data set but hold for the data set except the outliers. The test procedure uses the diagnostic, proposed by Hawkins(1991) for use with recursive fitting. The key to using recursive residuals is an ordering such that departures from the model are confined to the cases near one end of the data set.

The power of the test procedure is investigated through simulation. The proposed procedure are compared with alternative forms of various outlier patterns. It is shown that when ordering is used, the use of recursive residuals increases power significantly and nearly covers the masking effect when multiple outliers are present.

REFERENCES

1. Andrews, D. F., and Pregibon, D. (1978), "*Finding Outliers That Matter*", J. R. Statist. Soc. B, 40, 85-93
2. Brown, R. L., Durbin, J., and Evans, J. M. (1975), "*Techniques for Testing the Constancy of Regression Relation ships Over Time*", J. R. Statist. Soc. B, 37, 149-163
3. Chatterjee, S. and Hadi, A. S. (1988), "*Sensitivity Analysis in Linear Regression*", New York : Wiley.
4. Du Toit, S.M.C., Steyn, A.G.W., and Stumpf, R. H. (1986), "*Graphical Exploratory Data Analysis*", New York : Springer-Verlag.
5. Farebrocher, R. W. (1978), "*An Historical Note on Recursive Residuals*", J. R. Statist. Soc. B, 40, 373-375.
6. Galpin, J. S., and Hawkins, D. M. (1984), "*The Use of Recursive Residuals in Checking Model Fit in Linear Regression*", American Statistician, 38, 94-105.
7. Harvey, A. C., and Collier, P. (1977), "*Testing for functional misspecification in regression analysis*", Journal of Econometrics 6, 103-109.
8. Hawkins, D. M. (1980), "*The Identification of Outliers*", London : Chapman & Hall.
9. Hawkins, D. M. (1991), "*Diagnostics for use with regression recursive residuals*", Technometrics 33, 221-234.
10. Kianifard, F., and Swallow, W. H. (1989), "*Using Recursive Residuals, Calculated on Adaptively-Ordered Observations, to Identify Outliers in Linear Regression*", Biometrics 45, 571-585.