MASTER

The Influence Function and Its Application to Data Validation

Michael R. Chernick

OAK RIDGE NATIONAL LABORATORY
OPERATED BY UNION CARBIDE CORPORATION · FOR THE DEPARTMENT OF ENERGY

REPRODUCTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Printed in the United States of America. Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road, Springfield, Virginia 22161
Price: Printed Copy \$4.00; Microfiche \$3.00

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, contractors, subcontractors, or their employees, makes any warranty, express or implied, nor assumes any legal liability or responsibility for any third party's use or the results of such use of any information, apparatus, product or process disclosed in this report, nor represents that its use by such third party would not infringe privately owned rights.

Contract No. W-7405-eng-26

THE INFLUENCE FUNCTION AND ITS APPLICATION TO DATA VALIDATION

Michael R. Chernick
Regional and Urban Studies Section
Energy Division

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

Date Published - August 1979

OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37830
Operated by
UNION CARBIDE CORPORATION
for the
DEPARTMENT OF ENERGY

LB
DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

THE INFLUENCE FUNCTION AND ITS APPLICATION TO DATA VALIDATION

M. R. Chernick

ABSTRACT

Hampel's influence function has been used by Devlin, Gnanadesikan, and Kettenring to detect bivariate observations which have unusual influence on estimates of correlation. In the validation of energy data systems such observations may sometimes be considered to be outliers. The identification of such outliers may be valuable for the detection of errors in a data base.

When data are used in regression equations, those observations which have the greatest effect on the multiple correlation coefficient or the regression coefficients are of interest. The contours of constant influence are derived for the multiple correlation coefficient in the case of regressing two variables on a third.

In some problems the analytic form of the influence function may be difficult to derive. In such cases the empiric estimator of the influence function, as proposed by Mallows, may be useful for detecting outliers.

For FPC form 4 power plant data, the correlation between generation and consumption is a parameter of interest to users of the data. Estimates of the contours of constant influence were determined and used to detect outliers with respect to bivariate correlation.

1. INTRODUCTION

The influence function was defined by Hampel (1974) as a tool in assessing robust estimators. Devlin, Gnanadesikan, and Kettenring (1975) illustrate the potential use of the influence function for detecting outlying multivariate observations. The concept is that for estimators of parameters in a multivariate population an influence function can usually be defined which indicates where in the n -dimensional space of observations,

the observed vectors would have a large effect on the value of the estimator of the parameter. If we have a sample of observation vectors and we find that one of these vectors has a much greater effect on the estimator than any of the others we may consider that vector to be an outlier. If further consideration brings forth other information indicating that the suspected observation is an outlier we might decide that the estimate obtained by omitting the observation is more reasonable. Other alternative actions would be replacement with a new observation, imputation (i.e., replacement with a value considered likely for the conditions under which the observation was taken) or correction if it is possible to determine the error in the generation of the observation. In any case it is good statistical practice to note in any reporting of the data how outliers were treated and what the value of the original observation was.

Devlin, Gnanadesikan, and Kettenring (1975) illustrate how the influence function can be applied to detect outliers with respect to bivariate correlation. Some of the results for bivariate correlation will be given in Section 2.

In multivariate analysis several types of correlation are of interest. In data validation we are usually considering multivariate observations and users of the data will often compute bivariate correlations or do regression analyses in which the multiple correlation coefficient is used as a measure of goodness of fit for the regression equations. For the case of regressing two variables on a third variable, the influence function is derived in Section 3.

Mallows (1974) gives various estimates of the influence function which can be applied to a given data set. These estimates can be used to determine which observations could be considered as outliers. The

"empiric" estimate of the influence function seems to be one such estimator with desirable properties and general applicability.

In the instances of bivariate correlation given in Section 2 and multiple correlation given in Section 3 the influence function is derived as a function of some unknown parameters. These derivations are useful in that they give information about the directions, in the n -dimensional observation space, in which the influence function increases most rapidly. They also give information about the directions in which the influence function increases little or not at all. For a given data set the influence function can be approximated by replacing the unknown parameters with their sample values.

In the case of bivariate correlation the contours of constant value of the influence function are hyperbolae in the plane. These contours depend on the mean vector, the component standard deviations, and the correlation coefficient. Using the sample estimates for these parameters, one can obtain approximations to these contours which can be superimposed over a scatter plot of the data. Such a graphical procedure is a useful method for spotting outliers.

In some situations the parameters of interest may be so complex that a derivation of the influence function would be difficult. In such cases the empiric estimate is a useful tool for determining outliers. A description of the empiric estimate and some of its properties is given in Section 4.

We shall now give a definition of the influence function. The influence function depends on the parameter we are estimating, the observation vector whose influence we are measuring, and the distribution function of

a random observation vector. The parameter can be considered as a functional of the distribution function F and is commonly written $T(F)$. The estimator usually can be expressed as $T(F_m)$ where F_m is the empiric distribution function (i.e., $F_m(\underline{y})$ is the fraction of the m sample vectors $\underline{x}$ whose coordinates are all less than or equal to the coordinates of $\underline{y}$). The influence function is defined by the following equation whenever the limit on the right hand side exists:

$$I(F, T(F), \underline{x}) = \lim_{\epsilon \rightarrow 0} \frac{T((1 - \epsilon)F + \epsilon \delta_{\underline{x}}) - T(F)}{\epsilon} \quad (1)$$

In equation (1) $\underline{x} = (x_1, x_2, \dots, x_m)$ is the point of interest in the observation space, ϵ is a positive real number, and $\delta_{\underline{x}}$ is the distribution function which has all its probability mass concentrated at the point $\underline{x}$. For most practical applications the influence function will be well defined. Hampel (1974) gives a more general definition which allows the influence function to be vector valued. His definition allows for the consideration of several parameters simultaneously. This general definition will not be considered in this paper.

2. THE INFLUENCE FUNCTION FOR BIVARIATE CORRELATION

In the case of bivariate correlation $\underline{x} = (x_1, x_2)$ and

$$\begin{aligned} T(F) &= \frac{\int x_1 x_2 dF(x_1, x_2) - \int x_1 dF(x_1, \infty) \int x_2 dF(\infty, x_2)}{\sqrt{[\int x_1^2 dF(x_1, \infty) - (\int x_1 dF(x_1, \infty))^2][\int x_2^2 dF(\infty, x_2) - (\int x_2 dF(\infty, x_2))^2]}} \\ &= \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1) \text{Var}(X_2)}} \end{aligned}$$

(X_1, X_2) is a random vector with distribution function F . The correlation coefficient $T(F)$ shall also be denoted as ρ . Let $\mu_1 = E(X_1)$, $\mu_2 = E(X_2)$, $\sigma_1^2 = \text{Var}(X_1)$, and $\sigma_2^2 = \text{Var}(X_2)$. It is convenient to define

$$Y_1 = \frac{X_1 - \mu_1}{\sigma_1}$$

and

$$Y_2 = \frac{X_2 - \mu_2}{\sigma_2}$$

Since the transformation from (X_1, X_2) to (Y_1, Y_2) does not affect the correlation, $I(F, T(F), (x_1, x_2)) = I(H, T(H), (y_1, y_2))$ where

$$y_1 = \frac{x_1 - \mu_1}{\sigma_1}, \quad y_2 = \frac{x_2 - \mu_2}{\sigma_2}$$

and H is the distribution function for (Y_1, Y_2) . Now, since $T(F) = T(H) = \rho$, we shall denote the influence function by $I(H, \rho, (y_1, y_2))$. $E_H(Y_1)$, $E_H(Y_2)$, $\text{Var}_H(Y_1)$, and $\text{Var}_H(Y_2)$ shall denote the means and variances of Y_1 and Y_2 .

$$\text{Proposition 1: } I(H, \rho, (y_1, y_2)) = \frac{y_1 y_2 - \rho(y_1^2 + y_2^2)}{2}$$

where H is any bivariate distribution function for which $E_H(Y_1) = E_H(Y_2) = 0$ and $\text{Var}_H(Y_1) = \text{Var}_H(Y_2) = 1$.

$$\text{Proof: By definition } I(H, \rho, (y_1, y_2)) = \lim_{\epsilon \rightarrow 0} \frac{T(G) - T(H)}{\epsilon}$$

where $G(u_1, u_2) = (1 - \epsilon)F(u_1, u_2) + \epsilon L(u_1 \geq y_1, u_2 \geq y_2)$ and $L(u_1 \geq y_1, u_2 \geq y_2) = 1$ if $u_1 \geq y_1$ and $u_2 \geq y_2$, is 0 otherwise. In general for any statement A , $L(A) = 1$ if A is true, $L(A) = 0$ if A is false.

$$T(G) = \frac{\int u_1 u_2 dG(u_1, u_2) - \epsilon y_1 \epsilon y_2}{\sqrt{(\int u_1^2 dG(u_1, \infty) - \epsilon^2 y_1^2)(\int u_2^2 dG(\infty, u_2) - \epsilon^2 y_2^2)}}$$

where

$$G(u_1, \infty) = (1 - \epsilon) H(u_1, \infty) + \epsilon L(u_1 \geq y_1)$$

and

$$G(\infty, u_2) = (1 - \epsilon) H(\infty, u_2) + \epsilon L(u_2 \geq y_2) .$$

Note that when (Y_1, Y_2) has distribution function G , $E_G(Y_1) = \epsilon y_1$ and $E_G(Y_2) = \epsilon y_2$ since $E_H(Y_1) = E_H(Y_2) = 0$. We see that

$$\begin{aligned} \int u_1 u_2 dG(u_1, u_2) &= (1 - \epsilon) \int u_1 u_2 dH(u_1, u_2) + \epsilon \int u_1 u_2 dL(u_1 \geq y_1, u_2 \geq y_2) \\ &= (1 - \epsilon) \rho + \epsilon y_1 y_2 . \end{aligned}$$

We also have

$$\int u_1^2 dG(u_1, \infty) = 1 - \epsilon + \epsilon y_1^2$$

and

$$\int u_2^2 dG(\infty, u_2) = 1 - \epsilon + \epsilon y_2^2$$

since

$$\int u_1^2 dH(u_1, \infty) = \int u_2^2 dH(\infty, u_2) = 1 .$$

So

$$\frac{T(G) - T(H)}{\epsilon} = \frac{\epsilon y_1 y_2 + (1 - \epsilon) \rho - \epsilon^2 y_1 y_2}{\epsilon h(\epsilon)} - \frac{\rho}{\epsilon}$$

where

$$h(\epsilon) = (1 - \epsilon + \epsilon y_1^2 - \epsilon^2 y_1^2)^{\frac{1}{2}} (1 - \epsilon + \epsilon y_2^2 - \epsilon^2 y_2^2)^{\frac{1}{2}} .$$

Now

$$\lim_{\epsilon \rightarrow 0} \frac{T(G) - T(H)}{\epsilon} = y_1 y_2 + \rho \lim_{\epsilon \rightarrow 0} \frac{1 - \epsilon - h(\epsilon)}{\epsilon h(\epsilon)} .$$

Since $h(\epsilon) \rightarrow 1$ as $\epsilon \rightarrow 0$ both $1 - \epsilon - h(\epsilon)$ and $\epsilon h(\epsilon)$ tend to zero as $\epsilon \rightarrow 0$.

L'Hospital's rule is applicable and hence we get

$$\lim_{\epsilon \rightarrow 0} \frac{1 - \epsilon - h(\epsilon)}{\epsilon h(\epsilon)} = \lim_{\epsilon \rightarrow 0} \frac{-1 - h'(\epsilon)}{\epsilon h'(\epsilon) + h(\epsilon)}.$$

Differentiation of h yields

$$\begin{aligned} h'(\epsilon) = & \frac{1}{2} (1 - \epsilon + (\epsilon - \epsilon^2)y_1^2)^{-\frac{1}{2}} (1 - \epsilon + (\epsilon - \epsilon^2)y_2^2)^{\frac{1}{2}} (y_1^2(1 - 2\epsilon) - 1) \\ & + \frac{1}{2} (1 - \epsilon + (\epsilon - \epsilon^2)y_2^2)^{-\frac{1}{2}} (1 - \epsilon + (\epsilon - \epsilon^2)y_1^2)^{\frac{1}{2}} (y_2^2(1 - 2\epsilon) - 1). \end{aligned}$$

So

$$\lim_{\epsilon \rightarrow 0} h'(\epsilon) = \frac{y_1^2 + y_2^2 - 1}{2}.$$

Therefore

$$\lim_{\epsilon \rightarrow 0} \frac{-1 - h'(\epsilon)}{h'(\epsilon) + h(\epsilon)} = \frac{-(y_1^2 + y_2^2)}{2}.$$

Hence

$$\lim_{\epsilon \rightarrow 0} \frac{T(G) - T(H)}{\epsilon} = \frac{y_1 y_2 - \rho(y_1^2 + y_2^2)}{2}.$$

Although this proposition is given by Devlin, Gnanadesikan, and Kettenring (1975), who attribute it to Mallows, this author has not seen it proved anywhere in the literature. We note that if F is any bivariate distribution function for which $E_F(X_1)$, $E_F(X_2)$, $E_F(X_1^2)$, and $E_F(X_2^2)$ are all finite, so that a correlation between X_1 and X_2 exists,

$$I(F, \rho, (x_1, x_2)) = I(H, \rho, (y_1, y_2)) = y_1 y_2 - \frac{\rho(y_1^2 + y_2^2)}{2}$$

where

$$y_1 = \frac{x_1 - \mu_1}{\sigma_1}, \quad y_2 = \frac{x_2 - \mu_2}{\sigma_2}$$

and $\mu_1 = E_F(X_1)$, $\mu_2 = E_F(X_2)$, $\sigma_1^2 = \text{Var}_F(X_1)$, and $\sigma_2^2 = \text{Var}_F(X_2)$. H is the distribution of the random vector (Y_1, Y_2) where

$$Y_1 = \frac{X_1 - \mu_1}{\sigma_1} \text{ and } Y_2 = \frac{X_2 - \mu_2}{\sigma_2}.$$

Consider the equation

$$y_1 y_2 - \frac{\rho(y_1^2 + y_2^2)}{2} = C \quad (2)$$

where C is some constant. Let

$$v_1 = \frac{1}{\sqrt{2}} \left[\frac{y_1 + y_2}{\sqrt{1 + \rho}} + \frac{(y_1 - y_2)}{\sqrt{1 - \rho}} \right]$$

and

$$v_2 = \frac{1}{\sqrt{2}} \left[\frac{y_1 + y_2}{\sqrt{1 + \rho}} - \frac{(y_1 - y_2)}{\sqrt{1 - \rho}} \right].$$

After some algebra we see that equation 2 is equivalent to

$$(1 - \rho^2) v_1 v_2 = 2C. \quad (3)$$

This shows us that the contours of constant influence are hyperbolae with axes along the principal axes, the standardized sum and difference of y_1 and y_2 .

For a sample of size n from a bivariate population let r be the usual sample estimate of the correlation coefficient. We would use

$(1 - r^2)u_{i1}u_{i2}$ as an estimate of the influence function at the i th observation point $X_i = (X_{i1}, X_{i2})$ where

$$u_{i1} = \frac{1}{\sqrt{2}} \left[\frac{y_{i1} + y_{i2}}{\sqrt{1 + r}} + \frac{(y_{i1} - y_{i2})}{\sqrt{1 - r}} \right],$$

$$u_{i2} = \frac{1}{\sqrt{2}} \left[\frac{y_{i1} + y_{i2}}{\sqrt{1 + r}} + \frac{(y_{i1} - y_{i2})}{\sqrt{1 - r}} \right],$$

$$y_{i1} = \frac{x_{i1} - m_1}{s_1}, \quad y_{i2} = \frac{x_{i2} - m_2}{s_2}$$

and m_1 , m_2 , s_1^2 , and s_2^2 are the sample estimates of μ_1 , μ_2 , σ_1^2 , and σ_2^2 respectively. Since $(1 - r^2) u_{i1}u_{i2}$ is approximately equal to $(n - 1)d$ where d is the difference between the estimate of ρ with the i th observation included and the estimate with the i th observation deleted, we see that the influence on the i th observation is approximately the influence function estimate divided by $n - 1$. In general, a rough statement would be that for any observation the influence function estimate measures the effect of the observation on the estimator scaled by a factor of the sample size. The justification for the previous statement will be given in Section 4 on the empiric estimate of the influence function.

In the validation study of the FPC form 4 data it was discovered that the correlation between the variables consumption and generation was considered to be of interest. We decided to include the influence function for bivariate correlation between generation and consumption as an outlier detection tool in our analysis of the data.

3. THE INFLUENCE FUNCTION FOR MULTIPLE CORRELATION

In this section we will consider three random variables Y , X_1 , and X_2 . We are interested in the multiple correlation between Y and the pair (X_1, X_2) . The multiple correlation is defined as the square of the correlation between Y and that linear combination of X_1 and X_2 which has the maximum correlation with Y . Let

$$\Sigma = \begin{pmatrix} \sigma_Y^2 & \sigma_{YX_1} & \sigma_{YX_2} \\ \sigma_{YX_1} & \sigma_{X_1}^2 & \sigma_{X_1X_2} \\ \sigma_{YX_2} & \sigma_{X_1X_2} & \sigma_{X_2}^2 \end{pmatrix}$$

be the covariance matrix for (Y, X_1, X_2) . We partition Σ writing

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where $\sigma_1^2 = \sigma_Y^2$, $\Sigma_{12} = \Sigma_{21}^t = (\sigma_{YX_1}, \sigma_{YX_2})$

and

$$\Sigma_{22} = \begin{pmatrix} \sigma_{X_1}^2 & \sigma_{X_1X_2} \\ \sigma_{X_1X_2} & \sigma_{X_2}^2 \end{pmatrix}$$

The multiple correlation R^2 (sometimes R is called the multiple correlation) is shown in Anderson (1958) p. 86 to be

$$\frac{\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}}{\sigma_Y^2}.$$

Computations show

$$\Sigma_{22}^{-1} = \frac{1}{\sigma_{X_1}^2 \sigma_{X_2}^2 - \sigma_{X_1 X_2}^2} \begin{pmatrix} \sigma_{X_2}^2 & -\sigma_{X_1 X_2} \\ -\sigma_{X_1 X_2} & \sigma_{X_1}^2 \end{pmatrix}.$$

So, after matrix multiplications, we find that

$$\begin{aligned} R^2 &= \frac{1}{(\sigma_Y^2 \sigma_{X_1}^2 \sigma_{X_2}^2 - \sigma_{X_1 X_2}^2)} \left[\sigma_{X_2}^2 \sigma_{Y X_1}^2 - 2 \sigma_{Y X_2} \sigma_{X_1 X_2} \sigma_{Y X_1} + \sigma_{X_1}^2 \sigma_{Y X_2}^2 \right] \\ &= \frac{\rho_{Y X_1}^2 + \rho_{Y X_2}^2 - 2 \rho_{Y X_1} \rho_{Y X_2} \rho_{X_1 X_2}}{1 - \rho_{X_1 X_2}^2}. \end{aligned}$$

$\rho_{Y X_1}$, $\rho_{Y X_2}$, and $\rho_{X_1 X_2}$ are the bivariate correlation between Y and X_1 , Y and X_2 , and X_1 and X_2 respectively. R^2 is well defined as long as Y , X_1 , and X_2 all have finite second moments. R^2 depends only on the bivariate correlations and therefore we may standardize Y , X_1 , and X_2 without affecting R^2 .

Let

$$R^2(\epsilon) \text{ denote } T\left((1 - \epsilon)F + \epsilon \delta_{(Y, X_1, X_2)}\right).$$

Let

$$H(\epsilon) = \frac{R^2(\epsilon) - R^2}{\epsilon}.$$

We see that $H'(0) = I(F, R^2, (Y, X_1, X_2))$.

Now

$$H'(0) = \left. \frac{dR^2(\epsilon)}{d\epsilon} \right|_{\epsilon} = 0.$$

So, by the rules of calculus,

$$\begin{aligned}
 H'(0) = \frac{1}{(1 - \rho_{x_1 x_2}^2)^2} & \left[(1 - \rho_{x_1 x_2}^2)^2 (\rho_{y x_1} \rho_{y x_1}'(0)) \right. \\
 & + \rho_{y x_2} \rho_{y x_2}'(0) - \rho_{y x_1} \rho_{y x_2} \rho_{x_1 x_2}'(0) \\
 & - \rho_{y x_1} \rho_{x_1 x_2} \rho_{y x_2}'(0) - \rho_{x_1 x_2} \rho_{y x_2} \rho_{y x_1}'(0) \\
 & \left. + (\rho_{y x_1}^2 + \rho_{y x_2}^2 - 2\rho_{y x_1} \rho_{y x_2} \rho_{x_1 x_2}) 2\rho_{x_1 x_2} \rho_{x_1 x_2}'(0) \right] .
 \end{aligned} \tag{4}$$

Equation (4) simplifies to yield

$$\begin{aligned}
 H'(0) = \frac{2}{(1 - \rho_{x_1 x_2}^2)^2} & \left[\rho_{x_1 x_2}'(0) \left\{ \rho_{x_1 x_2} (\rho_{y x_1}^2 - \rho_{y x_1} \rho_{y x_2} \rho_{x_1 x_2} \right. \right. \\
 & + \rho_{y x_2}^2) - \rho_{y x_1} \rho_{y x_2} \left. \right\} + \rho_{y x_1}'(0) (\rho_{y x_1} - \rho_{x_1 x_2} \rho_{y x_2}) (1 - \rho_{x_1 x_2}^2) \\
 & \left. + \rho_{y x_2}'(0) (\rho_{y x_2} - \rho_{x_1 x_2} \rho_{y x_1}) (1 - \rho_{x_1 x_2}^2) \right]
 \end{aligned} \tag{5}$$

where

$$\rho_{x_1 x_2}'(0) = x_1 x_2 - \rho_{x_1 x_2} \frac{(x_1^2 + x_2^2)}{2}$$

$$\rho_{y x_1}'(0) = y x_1 - \rho_{y x_1} \frac{(y^2 + x_1^2)}{2}$$

and

$$\rho_{y x_2}'(0) = y x_2 - \rho_{y x_2} \frac{(y^2 + x_2^2)}{2} .$$

In the interesting special case where $\rho_{x_1 x_2} = 0$, this simplifies to

$$H'(0) = \left[2 y x_2 \rho_{y x_2} + y x_1 \rho_{y x_1} - x_1 x_2 \rho_{y x_1} \rho_{y x_2} \right] - \rho_{y x_1}^2 (y^2 + x_1^2) \quad (6)$$

$$- \rho_{y x_2}^2 (y^2 + x_2^2) .$$

In any case we see from equations (5) and (6) that the surfaces of constant influence are hyperbolic.

4. THE EMPIRIC INFLUENCE FUNCTION

The empiric version of the influence function is defined as follows:

$$I_E(m, T, \underline{x}) = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left[T((1 - \epsilon)F_m + \epsilon \delta_{\underline{x}}) - T(F_m) \right]$$

where F_m is the empiric distribution function for the sample of size m .

We would be particularly interested in $I_E(m, T, \underline{x}_i)$ for $i = 1, 2, \dots, m$ where $\underline{x}_i$ is the i th observation vector from a sample of size m .

The empiric influence function has some desirable large sample properties as pointed out by Mallows (1974). Included among these are the following:

P. 1. Consistency

P. 2. Approximate independence

Consistency means $I_E(m, T, \underline{x})$ converges in probability to $I(T(F), F, \underline{x})$ at each $\underline{x}$ as $m \rightarrow \infty$. Approximate independence means that $I_E(m, T, \underline{x}_i)$ and $I_E(m, T, \underline{x}_j)$ are asymptotically independent random variables for each $i \neq j$ as $m \rightarrow \infty$. Mallows does not give a mathematically precise definition of approximate independence.

These two properties justify the use of influence function contours superimposed over a scatter plot of the data for reasonably large samples. We see that in large samples it is reasonable to interpret the value of the influence function at an observation as being a measure of the effect of that observation on the estimate independent of the other observations.

The empiric influence function also has the property that it is a statistic (i.e., it can be evaluated for a given sample and does not depend on unknown parameters). It also can be evaluated at all vectors $\underline{x}$ in the n -dimensional observation space and not just at the sample points.

Suppose we have $m + 1$ observation vectors and let F_m be the empiric distribution function for $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_m$. Consider $I_E(m, T, \underline{x}_{m+1})$ for some T of interest.

$$I_E(m, T, \underline{x}_{m+1}) = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left[T((1 - \epsilon)F_m + \epsilon\delta_{\underline{x}_{m+1}}) - T(F_m) \right].$$

Now if m is large

$$\frac{1}{m+1}$$

is small and if we replace ϵ by

$$\frac{1}{m+1}$$

we get an approximation to the limit.

$$T\left(\frac{m F_m + \delta_{\underline{x}_{m+1}}}{m+1}\right)$$

where F_{m+1} is the empiric distribution function for all $m+1$ observations. We see that $I_E(m, T, X_{m+1}) \approx (m+1)[T(F_{m+1}) - T(F_m)]$. Note that $T(F_{m+1})$ is our estimate when X_{m+1} is included in the sample and $T(F_m)$ is the estimate when X_{m+1} is deleted. Hence $T(F_{m+1}) - T(F_m)$ represents the effect of the observation X_{m+1} on the estimate and our remark that the influence function is approximately this effect multiplied by the sample size is justified.

REFERENCES

- Anderson, T. W. 1958. *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons.
- Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. 1975. "Robust estimation and outlier detection with correlation coefficients." *Biometrika* 62, 531-545.
- Hampel, F. R. 1974. "The influence curve and its role in robust estimation." *J. Amer. Statist. Assoc.* 69, 383-393.
- Mallow, C. L. 1974. "On some topics in robustness." Unpublished paper.

INTERNAL DISTRIBUTION

1. J. T. Arehart	33. C. R. Weisbin
2-16. M. R. Chernick	34. G. W. Westley
17. R. M. Davis	35-36. Central Research Library
18. W. Fulkerson	37. Document Reference Section
19. W. E. Lever	38-39. Laboratory Records
20. G. E. Liepins	40. Laboratory Records (RC)
21-31. A. S. Loeb1	41. ORNL Patent Office
32. T. Takayama	

EXTERNAL DISTRIBUTION

42. Assistant Manager, Office of Energy Research and Development, DOE-ORO
43. Richard Beckman, Los Alamos Scientific Laboratory, MS-600, P. O. Box 1663, Los Alamos, NM 87545
44. Martha Bentley, Librarian, Florida State Energy Office, 301 Bryant Building, Tallahassee, FL 32301
45. Brian Berry, Gund Hall, Room 333, Harvard University, Cambridge, MA 01720
46. Larry Bray, Tennessee Valley Authority, 270LB, Knoxville, TN 37902
47. Paul Cho, Regional Assessment Division, E-201, Department of Energy, Washington, D.C. 20545
48. Craig H. Cranston, Office of Energy Information Validation, Department of Energy, 12th and Pennsylvania Avenue, N.W., Washington, D.C. 20461
49. Gary de Mik, MERT Division, Oak Ridge Associated Universities, P. O. Box 117, Oak Ridge, TN 37830
50. Dr. Asil Gezen, Transportation Research and Economics Association, Suite 888, 1901 N. Fort Myer Drive, Arlington, VA 22209
51. Dr. Thomas B. Griswold, Department of Energy, Capital Plaza Tower, Frankfort, KY 40601
52. Mildred Kikta, Energy Information Administration, 12th and Pennsylvania Avenue, N.W., Washington, D.C. 20461
53. Claire E. Leslie, Science Applications, Inc., P. O. Box 843, Oak Ridge, TN. 37830
54. Gus Whiton, Evaluation Research Corporation, 8330 Old Court House Road, Vienna, VA 22180
55. Nancy C. Lewis, Arkansas Energy Conservation and Policy Office, 960 Plaza West Bldg., Little Rock, AR 72205
56. Oficina de Energía, Apartado 41089 Estación Minillas, Santurce, Pureto Rico 00090

57. Barbara Otto, Regional Energy Information Center, P. O. Box 35228, Dallas, TX 75235
58. Walter Page, Department of Economics, West Virginia University, Morgantown, WV 26505
- 59-68. Brian Poole, Energy Information Administration, Department of Energy, 12th and Pennsylvania Avenue, Washington, D.C. 20461
69. Steve Reynolds, Energy Information Administration, Department of Energy, 12th and Pennsylvania Avenue, Washington, D.C. 20461
70. Dr. Kerry Smith, Resources for the Future, 1755 Massachusetts Avenue, N.W., Washington, D.C. 20036
71. David Sircle, 226 Capitol Blvd. Building, Suite 707, Nashville, TN 37219
72. Richard Stevens, Texas Energy Advisory Council, 7703 North Lamar, Austin, TX 78752
73. Lester Taylor, Department of Economics, University of Arizona Tucson, AZ 85721
74. Ronald Wyzga, Electric Power Research Institute, 3412 Hillview Avenue, P. O. Box 10412, Palo Alto, CA 94303
- 75-101. Technical Information Center, P. O. Box 62, Oak Ridge, TN 37830