THE INFLUENCE CURVE

Gabrielle Kelly

The influence curve of an estimator measures how much an individual observation changes the value of the estimator. Thus, in any estimation problem the role the i th data point say, plays in the analysis, can be made exact. This intuitively appealing idea of Hampels (1974) initiated interest in the influence curve. Now there is a substantial theory on its properties and uses in statistics, of which I give here a preliminary account.

To understand the purpose and nature of the influence curve, we must think of parameters and their corresponding estimators as functionals. Consider a real-valued functional $T(\cdot)$ defined on the space of distribution functions and let the parameter of interest be $\theta = T(F)$ (usually F denotes the 'true' underlying distribution function). To fix ideas look at the following examples.

Example 1: (i) The mean functional is given by

$$\mu(G) = \int x \, dG(x)$$

provided the integral exists.

(ii) The variance functional is given by

$$\sigma^2(G) = \int x^2 \, dG(x) - \left[\int x \, dG(x)\right]^2$$

again provided the integral exists.

In these examples the parameters of interest might be the true mean $\mu(F)$ and the true variance $\sigma^2(F)$. To look at estimators we have to consider $X_1, \ldots, X_n$ a random sample from a population with distribution function $F(\cdot)$. The empirical distribution function of these X's, is $F_n(\cdot)$, where

$$F_n(t) = [X_i\text{'s} \leq t]/n \qquad -\infty < t < \infty.$$

In many estimation problems the estimator $\hat{\theta}$ can be put in the same functional form as θ i.e. if $\theta = T(F)$ then $\hat{\theta} = T(F_n)$. Again we can look at familiar examples.

Example 2: (i) The estimator corresponding to the mean is

$$\hat{\mu} = \mu(F_n) = \int x \, dF_n(x) = \sum_{i=1}^{n} X_i/n = \bar{X},$$

i.e. the usual sample mean.

(ii) The estimator corresponding to the variance is

$$\hat{\sigma}^2 = \sigma^2(F_n) = \sum_{i=1}^{n} (X_i - \bar{X})^2/n.$$

We now define the influence curve of a functional $T(\cdot)$ at a point G, $IC(T, G ;\cdot)$, as follows. Let $W = (1-\varepsilon)G + \varepsilon\delta_z$ be a perturbation of G by δ_z, the distribution function for the point mass of one at z, i.e.

$$\delta_z(x) = \begin{cases} 0, & x \leq z \\ 1, & x \geq z \end{cases}$$

Then

$$IC(T, G ; z) = \lim_{\varepsilon \to 0} \frac{T(W) - T(G)}{\varepsilon},$$

$$= \frac{d}{d\varepsilon} T(W) \Big|_{\varepsilon=0},$$

provided the limit exists for every $z \in R$. (It is also known as the Gâteau differential of T at δ_z). This derivative measures the effect on the functional T of a small (infinitesimal) change in the weight the distribution function G gives to the point z. Thus, when we consider the estimator $T(F_n)$, its influence curve $IC(T, F_n ; z)$ measures the "influence" on the estimator of an additional observation at the point z. To see this look at the influence curve of the mean.

Example 3: Denote the mean of F by $\mu(F)$. Then we have

$$W = (1-\varepsilon)F + \varepsilon\delta_z ,$$

$$\mu(W) = \int x \, d[(1-\varepsilon)F + \varepsilon\delta_z],$$

$$= (1-\varepsilon)\mu(F) + \varepsilon z ,$$

and

$$\frac{d}{d\varepsilon}\mu(W)\Big|_{\varepsilon=0} = z - \mu(F).$$

So

$$IC(\mu, F ; z) = z - \mu(F).$$

In particular for the sample mean

$$IC(\mu, F ; z) = z - \bar{x} .$$

Thus the effect on the sample mean of an additional observation is directly proportional to the value of the observation as is shown in Fig. 1.

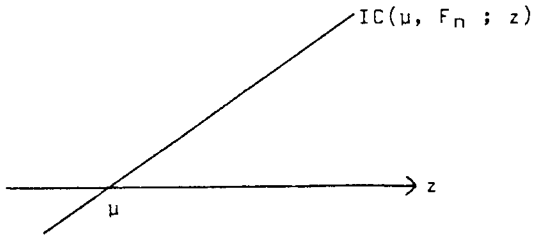

FIGURE 1

The Influence Curve of the Sample Mean.

Here, as is often the case, the influence curve is easy to compute. Note that the distribution functions can be multivariate. In such cases, the point z corresponds to a vector-valued observation. The influence curve has also been defined for vector-valued functionals. For example, the functional given by $T(G) = (\mu(G), \sigma^2(G))^T$ has a vector-valued influence curve defined at F as the pointwise limit:

$$IC(T, F ; z) = \frac{d}{d\varepsilon}T(W)\Big|_{\varepsilon=0} ,$$

$$= (IC(\mu, F ; z), IC(\sigma^2, F; z))^T,$$

provided the limit exists for every z. Now we examine various aspects of the influence curve to gain insight into the nature of an estimator.

Firstly, the shape of the influence curve provides information about the robustness properties of an estimator. In the example above we see the influence curve is unbounded reflecting the fact that the sample mean is sensitive to extremely large or small observations. In contrast to this the influence curve of the median is a step function.

Example 4: The median functional, $m(\cdot)$, is given by

$$m(F) = \tfrac{1}{2}(\sigma^* + \sigma^{**}),$$

where

$$\sigma^* = \sup\{x \mid F(x) \leq \tfrac{1}{2}\} \text{ and } \sigma^{**} = \inf\{x \mid F(x) \geq \tfrac{1}{2}\}.$$

Then

$$m(F_n) = \text{median } \{X_1, \ldots, X_n\},$$

is the sample median. The influence curve (see Fig. 2) is given by

$$IC(m, F ; z) = \begin{cases} -\dfrac{1}{f(F^{-1}(\frac{1}{2}))} , & z \quad F^{-1}(\tfrac{1}{2}) \\ 0, & \text{otherwise}, \end{cases}$$

where

$$f(x) = \frac{d}{dx}F(x)$$

This influence curve is bounded and thus is not sensitive to extreme observations and is robust in this sense.

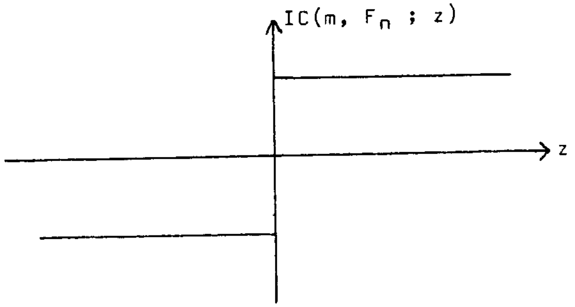

FIGURE 2

The Influence Curve of the Sample Median.

The influence curve is used to derive new estimators with pre-specified robustness properties. The study of various norms connected with the influence curve leads to estimators which are "best" over a large class of distribution functions. This breaks with Fisher's classical theory of estimation which looks for the best estimator with respect to one particular distribution function. The interested reader is referred to Huber (1977).

The influence curve plays an important role in asymptotic theory. The asymptotic variance of an estimator, for example, can be written in terms of the influence function. The usual delta method formula for calculating the asymptotic variance of an estimator $T(F_n)$ is in fact

$$\int IC^2(T, F ; x)\, dF(x)/n.$$

It can be estimated in the usual way by replacing F with F_n i.e. by

$$\sum_{i=1}^{n} IC^2(T, F_n ; X)/n^2.$$

This is because for a wide range of estimators, $T(F_n)$, Von Mises (1947) showed that

$$T(F_n) = T(F) + \sum_{i=1}^{n} IC(T, F ; X_i)/\sqrt{n} + R_n,$$

where

$$P\{\sqrt{n}|R_n| > \varepsilon\} < \varepsilon,$$

for n large. This means

$$\sqrt{n}(T(F_n) - T(F)) = \sum_{i=1}^{n} IC(T, F ; X_i)/\sqrt{n}.$$

So

$$\sqrt{n}(T(F_n) - T(F)) \overset{p}{\to} 0,$$

where $\overset{p}{\to}$ denotes convergence in probability. Then by the Central Limit Theorem

$$\sqrt{n}(T(F_n) - T(F) \overset{d}{\to} N(0, V(F)),$$

where $\overset{d}{\to}$ denotes convergence in distribution and

$$V(F) = \int IC^2(T, F ; x)\, dF(x).$$

Thus, we have

$$\text{Variance } (\sqrt{n}T(F_n)) \to V(F).$$

A simple illustration of this is as follows.

Example 5: For the mean, by Example 3, we have

$$IC(\mu, F ; x) = x - \mu(F).$$

Thus,

$$\int IC^2(\mu, F ; x)\, dF(x) = \int (x-\mu)^2\, dF(x) = \sigma^2(F).$$

So here,

$$V(F) = \sigma^2(F),$$

which is estimated by

$$\sigma^2(F_n) = \int IC^2(\mu, F_n ; x)\, dF_n(x)$$
$$= \sum_{i=1}^{n} (X_i - \overline{X})^2/n.$$

Our formula for the estimated variance of $\overline{X}$ is then

$$\widehat{\text{Variance}}\ (\overline{X}) = \sum_{i=1}^{n} (X_i - \overline{X})^2/n^2.$$

A more recent development concerning the influence curve is its use in outlier detection. In many statistical problems it is assumed that the form of the underlying distribution function F is known apart from an unknown parameter θ. The assumed model is then denoted by $F(\cdot\ ; \theta)$. A familiar example is the normal distribution with unknown mean θ. Let $\hat{\theta} = \theta(F_n)$ be the estimator and

$$\sum_{i=1}^{n} IC^2(\theta, F_n ; X_i)/n^2$$

its estimated variance. We denote this as

$$\hat{V}ar\ (\hat{\theta}) = \sum_{i=1}^{n} IC^2(\theta, F_n ; X_i)/n^2.$$

The statistic

$$D_i = IC^2(\theta, F_n ; X_i)/\sum_{i=1}^{n} IC^2(\theta, F_n ; X_i)$$

can be interpreted as a measure of the 'goodness of fit' of the i th data point to the model $F(\cdot\ ; \theta)$. It can be shown that for n large $D \overset{d}{\sim} F(1, n-1)$, where $F(1, n-1)$ is the F-distribution with 1 and n-1 degrees of freedom. The symbol "$\overset{d}{\sim}$" denotes "is asymptotically distributed as" when $n \to \infty$. Thus D_i provides a measure of fit of the i th data point in terms of descriptive levels of significance. For p vector-valued influence curves, IC^2 in D_i is replaced by $IC^T IC$ and then $D_i \overset{d}{\sim} F(p, n-p)$. Another interesting interpretation of D_i is as follows. Let $\hat{\theta}_{-i}$ be $\hat{\theta}$ with the i th observation omitted. Now

$$(n-1)(\hat{\theta}-\hat{\theta}_{-i}) = \left.\frac{\theta(W) - \theta(F)}{\varepsilon}\right|_{\varepsilon = \frac{1}{n-1}}, \qquad (1)$$

and this together with the definition of the influence curve implies

$$D_i \overset{d}{\sim} \frac{(\hat{\theta} - \hat{\theta}_{-i})}{Var\ \hat{\theta}}.$$

{Equation (1) also provides the connecting link between the influence curve and the jackknife; c.f. Miller (1974).} This can be used in the following way. Let $F(1, n-1, 1-\alpha)$ denote the $(1-\alpha)$ th probability point of the $F(1, n-1)$ distribution. Then for example, if $D = F(1, n-1, 5)$, removal of the i th data point moves the estimator to the edge of the 50% confidence region for θ based on $\hat{\theta}$. Measures of large residuals from regression models surveyed by Atkinson (1982) can be shown to be all versions of the statistic D_i above.

The following example serves as a demonstration of the use of D_i. No attempt at a complete analysis is made.

<u>Example 6</u>: Miller (1982) presented simultaneous pairs of measurements of serum kanamycin levels in blood samples drawn from twenty babies. One of the measurements was obtained by a heelstick method (X), the other using an umbilical catheter (Y). The heelstick method had been customarily used but due to the necessity of frequently drawing samples, this left neonates with badly bruised heels. The aim of the experiment was to see if the two methods measured the same levels except for error variability. If true, this would eliminate the unnecessary trauma to the newborn of repeated venapunctures. Since both measurements

are subject to error, an error in variables rather than regression analysis is used (c.f. Kendall and Stuart). It was assumed the true F was the bivariate normal and that the points followed a line with unknown slope β and intercept α. The parameter of interest then is $\underset{\sim}{\theta} = (\alpha, \beta)^T$ and the influence curve is bivariate.

The twenty pairs of heelstick and catheter values are presented in Table 1.

Baby	Heelstick	Catheter
1	23.0	25.2
2	33.2	26.0
3	16.6	16.3
4	26.3	27.2
5	20.0	23.2
6	20.0	18.1
7	20.6	22.2
8	18.9	17.2
9	17.8	18.6
10	20.0	16.4
11	26.4	24.8
12	21.8	26.8
13	14.9	15.4
14	17.4	14.9
15	20.0	18.1
16	13.2	16.3
17	28.4	31.3
18	25.9	31.2
19	18.9	18.0
20	13.8	15.6

TABLE 1

Serum kanamycin levels in blood samples drawn simultaneously from an umbilical catheter and a heel venapuncture in twenty babies

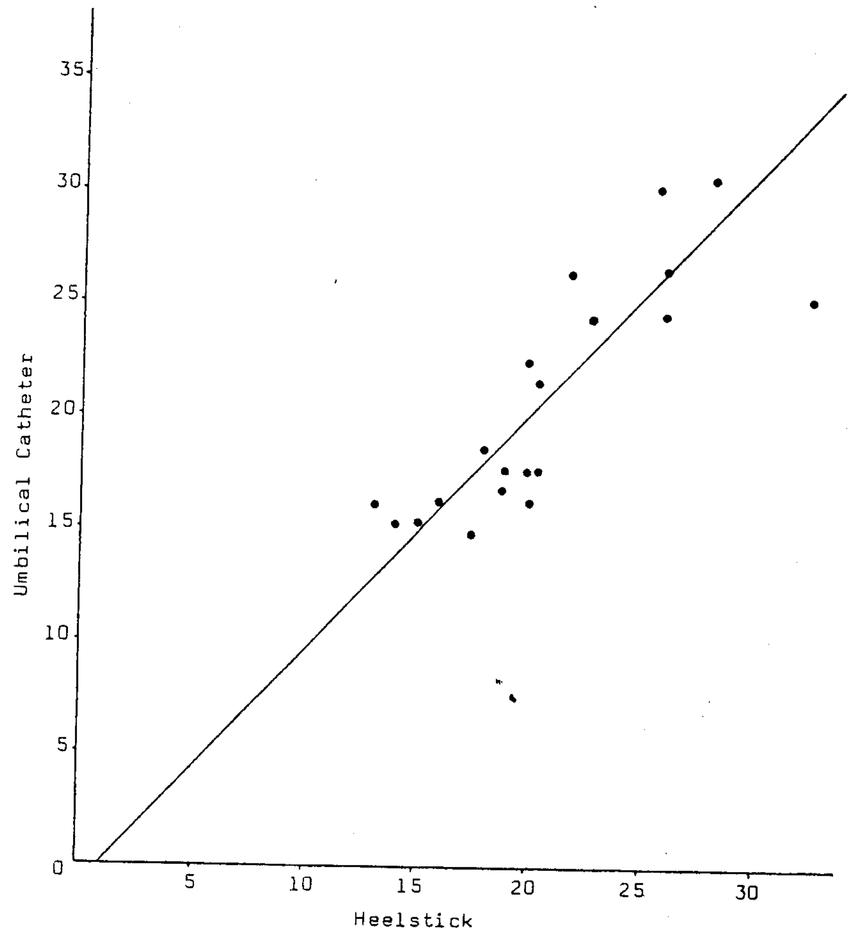

FIGURE 3

Plot of the Twenty Pairs of Serum Kanamycin levels from Table 1

TABLE 2: Estimates of the influence curve and deleted estimates, for each data point

Baby	Intercept: $IC(\alpha, F_n; Z)$ Sample Influence	Intercept: $\hat{\alpha}_{-i}$ Deleted Estimate	Slope: $IC(\beta, F_n; Z)$ Sample Influence	Slope: $\hat{\beta}_{-i}$ Deleted Estimate	D_i Normalised Influence
1	-31.68	- .96	.26	1.06	.012
2	61.66	-5.26	-3.36	1.29	.768
3	- 1.64	-1.07	.06	1.07	.001
4	- 1.06	-1.10	.08	1.07	.000
5	1.07	-1.21	.09	1.07	.051
6	- 6.34	- .83	.20	1.06	.046
7	.77	-1.20	.03	1.07	.012
8	- 7.35	- .77	.26	1.06	.041
9	3.36	-1.34	- .12	1.08	.010
10	-14.71	- .39	.52	1.04	.175
11	8.02	-1.61	- .49	1.10	.025
12	-10.85	- .59	.74	1.03	.088
13	4.21	-1.40	-0.17	1.08	.009
14	-14.93	- .34	.59	1.04	.117
15	- 6.34	- .83	.20	1.06	.046
16	23.83	-2.57	- .98	1.13	.290
17	-16.20	- .14	.88	1.01	.045
18	-30.47	.58	1.68	.97	.145
19	- 3.75	- .96	.13	1.06	.013
20	14.40	-2.01	- .59	1.04	.103

A graphical display of these twenty pairs is reproduced in Fig. 3. The estimates of intercept and slope from the analysis are

$$\hat{\theta} = -1.16, \quad \hat{\beta} = 1.07$$

The line with $\hat{\alpha}$ and $\hat{\beta}$ is drawn in Fig. 3. Table 2 presents the influence curve of the slope and intercept at each data point. The estimates of α and β obtained by deleting each data point in turn are also tabled as well as the values of D_i. We see Babies 2 and 16 have the largest values of the influence curve and have a negative influence on the slope estimate. If we look again at Fig. 3 we realise how difficult it is to detect and agree on what an 'outlier' is, without some objective measure. From Table 3, we have $D_2 = .688 = F(2, 18; .45)$, so removal of Baby 2 moves the estimate of $\underset{\sim}{\theta}$ to approximately the edge of a 55% confidence region around $\hat{\underset{\sim}{\theta}}$. Removal of Baby 16 moves the estimate of $\underset{\sim}{\theta}$ to the edge of a 40% confidence region around $\hat{\underset{\sim}{\theta}}$.

The influence curve is easy to explain and interpret in consultancy work and we could make the argument that it become an integral part of all data analysis. For this reason, I have emphasised mathematical rigor less than intuitive meaning in this article. There are still many open mathematical details, like regularity conditions, to be addressed. Many other influence curves of widely used estimators, need to be derived, studied and interpreted.

References

Atkinson, A.C. (1982). Regression Diagnostics, Transformations and Constructed Variables (with discussion). J. Royal Statist. Soc. B 44, 1-36.

Efron, B. (1981). Nonparametric Estimates of Standard Error:

the Jackknife, the Bootstrap and Other Methods. Biometrika 68, 589-599.

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. Siam: Philadelphia.

Hampel, F.R. (1974). The Influence Curve and its Role in Robust Estimation. J. Amer. Stat. Assoc. 69, 383-393.

Huber, P.J. (1977). Robust Statistical Procedures. Siam: Philadelphia.

Kendall, M.G. and Stuart, A. (1961). The Advanced Theory of Statistics, Vol. 2. Hafner: New York.

Miller, R., Jr., editor (1980). Biostatistics Casebook. Wiley.

Von Mises, R. (1947). On the Asymptotic Distribution of Differentiable Statistical Functions. Ann. of Math. Statist. 18, 309-348.

Department of Statistics,
University College,
Cork.

FEIGENBAUM'S NUMBER

J. Kennedy

In 1975 Mitchell J. Feigenbaum [2] of the Los Alamos National Laboratory, whose work concerns the transition from periodic to aperiodic bahaviour, discovered a new universal constant which has since been called Feigenbaum's number. He had been using a programmable calculator to examine the iteration of one-parameter families of maps of a finite interval into itself. One map he looked at was $x \to f_B(x) = Bx(1-x)$; another was $x \to B \sin \pi x$, both on the interval $[0,1]$. Feigenbaum observed some common features of the parameter dependence of these maps which he suspects would not have been noticed had the calculations been carried out on a large computer rather than a small calculator. The theory of these maps has been extended by Pierre Collet of Paris, Jean-Pierre Eckmann of Geneva and H. Koch of Harvard. The topic is reviewed in the book by Collet and Eckmann [3] on which this note is based.

For the most part, Collet and Eckmann consider mappings $x \to f(x)$ which are C^1-unimodal. A mapping f of the interval $[-1,1]$ into itself is C^1-unimodal if f is continuous; $f(0)=1$; f is strictly decreasing on $(0,1]$ and strictly increasing on $[-1,0)$; and f is once continuously differentiable with $f'(x) \neq 0$ when $x \neq 0$.

Denoting by f^0 the identity, $f^1 = f$, $f^2 = f \circ f$, $f^n = f \circ f^{n-1}$ the sets of iterates of points $x \in [-1,1]$

$$O_f(x) = \{x, f(x), f^2(x), f^3(x) \ldots\}$$

are called the orbits of f. A point $x \in [-1,1]$ is called a periodic point for f if $O_f(x)$ is a finite set. The cardinality of this set is called the period of x and $O_f(x)$ is called the periodic orbit of x. It is then also the periodic orbit of