Influence Functions for Fun and Profit

Jay Kahn
Ross School of Business, University of Michigan

July 10, 2015

An influence function tells you the effect of a change in one observation on an estimator. It's' useful in studying model robustness and calculating variance-covariance matrices for certain types of estimators, especially when more straightforward methods become hard to implement.

1 Definition

Let $(\Omega, \mathcal{S}, P)$ be a probability space with random variables $X_1, \ldots, X_n : (\Omega, \mathcal{S}) \to (\mathcal{X}, \Sigma)$ i.i.d. random variables defined on it. Finally, suppose we're interested in an estimator, $\hat{\theta} : (\mathcal{X}^n, \Sigma^n) \to (\Gamma, \mathcal{A})$.

One way to begin is to use the concept of a "contaminated" distribution function:

Definition 1.1. *Suppose F is a distribution on Σ. The **contaminated distribution function** is defined:*

$$F_\epsilon(x|G) = (1-\epsilon)F + \epsilon G \tag{1}$$

where δ_x is the probability measure on Σ which assigns probability 1 to $\{x\}$ and 0 to all other elements of Σ.

Imagine sampling F as pulling individuals from a barrel. The distribution $F_\epsilon(x|G)$ is called a contaminated distribution because sampling from it is like sampling from a barrel of F where some number of individuals from G have slipped into the barrel. This concept alreay has a clear use for robustness applications. But there are even more useful restrictions of contaminated distribution functions:

Definition 1.2. *Suppose F is a distribution on Σ. The **ϵ-contaminated distribution function** is defined:*

$$F_\epsilon(x) = (1-\epsilon)F + \epsilon\delta_x = F_\epsilon(x|\delta_x) \tag{2}$$

where δ_x is the probability measure on Σ which assigns probability 1 to $\{x\}$ and 0 to all other elements of Σ.

Now instead of talking about some members of G slipping into the distribution, it's as if we have too many of individual x in the barrel. Later on, this will allow us to talk about oversampling of one particular observation. Using the ϵ-contaminated distribution we can define the influence function fairly easily:

Definition 1.3. *The **influence function** of $\hat{\theta}$ at F, $\psi_i : \mathcal{X} \to \Gamma$ is defined:*

$$\psi_{\hat{\theta},F}(x) = \lim_{\epsilon \to 0} \frac{\hat{\theta}(F_\epsilon(x)) - \hat{\theta}(F)}{\epsilon} \tag{3}$$

In other words, the influence function is the marginal effect of oversampling x on a particular estimator for an uncontaminated distribution.

In a more general setting, we can discuss a certain type of derivative known as the Gâteaux derivative:

Definition 1.4. *The **Gâteaux derivative** of $\hat{\theta}$ at F in the direction G is defined:*

$$L_F(x) = \lim_{\epsilon \to 0} \frac{\hat{\theta}((1-\epsilon)F + \epsilon G) - \hat{\theta}(F)}{\epsilon} \tag{4}$$

So the influence function is clearly the Gâteaux derivative of $\hat{\theta}$ at F in the direction of δ_x. Recasting influence functions in this light means that we have many useful properties of influence functions which come from the fundamental theorem of calculus applied to Gâteaux derivatives. Most importantly, we can make use of the chain rule:

Theorem 1.5 (Chain rule for influence functions)**.** *Suppose we have an estimator $\hat{\theta}(F)$ such that $\hat{\theta}(F) = T(\hat{\theta}_1(F), \ldots, \hat{\theta}_n(F))$. Then:*

$$\psi_{\hat{\theta},F}(x) = \sum_{i=1}^{n} \frac{\partial T}{\partial \hat{\theta}_i} \psi_{\hat{\theta}_i}(x) \tag{5}$$

2 Examples

These properties of the influence function already allow us to derive influence functions for a wide variety of estimators. In what follows I will use X or Y to denote a random random variable, following a joint distribution F. Particular values of X or Y will be denoted with lower case letters, x or y.

2.1 Mean of a distribution

Let $\hat{\theta} = \mathrm{E}[X]$, and denote the expectation with respect to the current distribution function $F(x)$ as $E_F(x)$. Then:

$$\begin{aligned} \hat{\theta}(F_\epsilon(x)) &= \mathrm{E}_{F_\epsilon(x)}[X] \\ &= (1-\epsilon)\,\mathrm{E}_{F_\epsilon(x)}[X] + \epsilon x \end{aligned}$$

Therefore, applying the definition of the influence function above:

$$\Rightarrow \psi_{\hat{\theta}}(x) = x - \mathrm{E}[X] \tag{6}$$

Here we can see a feature of influence functions we'll derive later, which is:

$$\mathrm{E}\left[\psi_{\hat{\theta}}(x)\right] = \mathrm{E}[x - \mathrm{E}[X]] = 0$$

Among other things, this is a useful way to check that there are no errors in calculating an influence function: if the mean is not equal to zero (or within the computer language's zero tolerance), then the influence function has been calculated incorrectly.

2.2 Variance and covariance of a distribution

Suppose we're now interested in $\hat{\theta}(F) = \mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$. Then by the Chain rule for influence functions:

$$\begin{aligned} \psi_{\hat{\theta}}(x) &= (x^2 - \mathrm{E}[X^2]) - 2\,\mathrm{E}[X](x - \mathrm{E}[X]) \\ &= (x - \mathrm{E}[X])^2 - \mathrm{Var}[X] \end{aligned} \tag{7}$$

Similarly, for the covariance of X and Y, $\hat{\theta}(F) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$.

$$\begin{aligned} \psi_{\hat{\theta}}(x) &= \big(xy - \mathrm{E}[XY]\big) - \big(x - \mathrm{E}[X]\big)\,\mathrm{E}[Y] - \big(y - \mathrm{E}[Y]\big)\,\mathrm{E}[X] \\ &= (x - \mathrm{E}[X])(y - \mathrm{E}[Y]) - \mathrm{Cov}(X, Y) \end{aligned} \tag{8}$$

In the three cases above, you can see a form start to develop. The influence function for variance is just the moment condition that defines variance evaluated at x, minus the population estimate of variance. This is really just an extension of the influence function for the mean defined above.

2.3 Simple linear regression

Often times instead of summary statistics, we're interesting in coefficients from regressions. Take the simplest case of regression, a single-variable linear regressions with a slope and an intercept. In this case, $\hat{\theta}(F) = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$. A final application of the chain rule is necessary to find the influence function for this estimator:

$$\begin{aligned}
\psi_{\hat{\theta}}(x,y) &= \frac{(x-\text{E}[X])(y-\text{E}[Y]) - \text{Cov}(X,Y)}{\text{Var}(X)} - \frac{\left((x-\text{E}[X])^2 - \text{Var}(X)\right)\text{Cov}(X,Y)}{\left(\text{Var}(X)\right)^2} \\
&= \frac{(x-\text{E}[X])(y-\text{E}[Y]) - \beta(x-\text{E}[X])^2}{\text{Var}(X)} \\
&= \frac{(x-\text{E}[X])}{\text{Var}(X)}\left[(y-\text{E}[Y]) - \beta(x-\text{E}[X])\right] \qquad (9)
\end{aligned}$$

This equation isn't quite as simple as the previous examples, but notice that it is still the case that $\text{E}\left[\psi_{\hat{\theta}}(x,y)\right] = 0$. I'll cover more complicated examples of regressions later on.

2.4 Median

We can also apply influence functions to estimators that aren't based on expectations, though these influence functions are generally harder to derive. For instance, the median is a statistic, M_X, which solves:

$$0.5 = F(M_X)$$

It is convenient to start with the influence function for the median of a uniform number over the unit interval, $\psi_{\tilde{M_X}}(x)$. The uncontaminated median is 0.5. The contaminated median solves:

$$0.5 = (1-\epsilon)M_X + \epsilon \mathbb{1}_{M_X > x}$$

This equation has three possible solutions:

$$M_X = \begin{cases} \frac{0.5}{1-\epsilon} & \text{if } 0.5 < (1-\epsilon)x \\ x & \text{if } 0.5-\epsilon < (1-\epsilon)x < 0.5 \\ \frac{0.5-\epsilon}{1-\epsilon} & \text{if } (1-\epsilon)x < 0.5-\epsilon \end{cases}$$

Taking the derivative, and letting ϵ shrinks to zero, the influence function for the median of a uniform variable becomes:

$$\psi_{\tilde{M_X}}(x) = \begin{cases} 0.5 & \text{if } 0.5 < x \\ X & \text{if } x = 0.5 \\ -0.5 & \text{if } x < 0.5 \end{cases}$$

Now imagine the general random variable X with distribution function F_X as simply a transformation of a uniform random random variable, U: $X = F_X^{-1}(U)$ (this is an example of the inverse or Smirnov transform). The median of X, M_X can be represented as a transformation of the median of U, $M_X = F_X^{-1}(M_U)$. The derivative of F_X^{-1} is just $\frac{1}{f_X(M_X)}$, where f_X is the density function of X. Applying the chain rule for influence functions:

$$\psi_{\hat{M_X}}(x) = \begin{cases} \frac{0.5}{f_X(M_X)} & \text{if } M_X < x \\ 0 & \text{if } x = M_X \\ -\frac{0.5}{f_X(M_X)} & \text{if } x < M_X \end{cases}$$

It's difficult to find a sample counterpart for this influence function, as it requires density function estimation, but its equation is now known. In practice I've found that standard Kernel estimation works well for calculating the variance of a median, but for functions of multiple medians the error involved with density estimation starts to create problems.

3 M-estimators and influence functions

Most estimators we use in econometrics are actually solutions to maximizing or minimizing a criterion function. In general, an M-estimator is an estimator for a parameter, θ that is the solution to some maximization problem over the data:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathrm{E}\left[G(X, \theta)\right]$$

This concept nests maximum-likelihood, least-squares and generalized method of moments estimators. We're going to deal with a class of M-estimators for which G is differentiable and can therefore the estimate can written as the solution to a set of equations:

$$\mathrm{E}\left[g(X, \theta)\right] = 0$$

where $g(X, \theta) = \nabla_\theta G(X, \theta)$. For the contaminated distribution, θ must solve:

$$(1 - \epsilon)\,\mathrm{E}\left[g(X, \theta)\right] + \epsilon g(x, \theta) = 0$$

We can use total differentiation to figure out the effect of changing ϵ:

$$\begin{aligned} \frac{d}{d\epsilon}(1 - \epsilon)\,\mathrm{E}\left[g(X, \theta)\right] &= -\frac{d}{d\epsilon}\epsilon g(x, \theta) \\ -\mathrm{E}\left[g(X, \theta)\right] + (1 - \epsilon)\,\mathrm{E}\left[\nabla_\theta g(X, \theta)\right]\frac{d\theta}{d\epsilon} &= -g(x, \theta) - \epsilon \nabla_\theta g(x, \theta)\frac{d\theta}{d\epsilon} \end{aligned}$$

Around $\epsilon = 0$, the estimator must solve $\mathrm{E}\left[g(X, \theta)\right] = 0$, so:

$$\begin{aligned} \mathrm{E}\left[\nabla_\theta g(X, \theta)\right]\frac{d\theta}{d\epsilon} &= -g(x, \theta) \\ \frac{d\theta}{d\epsilon} &= -\mathrm{E}\left[\nabla_\theta g(X, \theta)\right]^{-1} g(x, \theta) = \psi_\theta(x) \end{aligned} \tag{10}$$

Here we already have one useful piece of information: take the expectation of this influence function over the distribution. You'll find you get (nearly) the same equation you would use for **Newton iterations** to find the maximum of the estimator. So you can find the actual M-estimator simply by using influence functions to update your guess (provided that estimator is suitably well behaved).

3.1 Interpretation for MLE

For MLE, $G(X, \theta)$ is the log-likelihood: $\ln f(X, \theta)$. This means that:

$$\begin{aligned} g(X, \hat{\theta}) &= \frac{1}{f(X, \hat{\theta})}\nabla_\theta f(X, \hat{\theta}) \\ \nabla_\theta g(X, \hat{\theta}) &= \left(\frac{1}{f(X, \hat{\theta})}\right)^2 \nabla_\theta f(X, \hat{\theta}) + \frac{1}{f(X, \hat{\theta})}\mathrm{H}_\theta f(X, \hat{\theta}) \\ \mathrm{E}\left[\nabla_\theta g(X, \hat{\theta})\right] &= \mathrm{E}\left[\frac{1}{f(X, \hat{\theta})}\mathrm{H}_\theta f(X, \hat{\theta})\right] \end{aligned}$$

So:

$$\psi_\theta(x) = \mathrm{E}\left[\frac{1}{f(X,\hat{\theta})}\mathrm{H}_\theta f(X,\hat{\theta})\right]^{-1}\frac{1}{f(x,\hat{\theta})}\nabla_\theta f(x,\hat{\theta}) \tag{11}$$

3.2 Interpretation for GMM

For GMM, $G(X,\theta) = M(X,\theta)'WM(X,\theta)$, and therefore at the estimate:

$$g(X,\hat{\theta}) = m(X,\hat{\theta})'WM(X,\hat{\theta})$$
$$\mathrm{E}\left[\nabla_\theta g(X,\theta)\right] = \mathrm{E}\left[m(X,\theta)'Wm(X,\theta)\right]$$

where $m(X,\hat{\theta}) = \nabla_\theta M(X,\hat{\theta})$. So the influence function for a GMM estimate is:

$$\psi_\theta(x) = \mathrm{E}\left[m(X,\hat{\theta})'Wm(X,\hat{\theta})\right]^{-1} m(x,\hat{\theta})'WM(x,\hat{\theta}) \tag{12}$$

This clearly nests the influence functions for mean, variance, covariance and simple regression coefficients in the sections above. In addition, imagine that the moment condition is additively seperable from the parameter of interest, such that $M(x,\hat{\theta}) = h(x) - \hat{\theta}$. Then $m(x,\hat{\theta}) = 1$, so that $\psi_\theta(x) = M(x,\hat{\theta}) = h(x) - \hat{\theta}$. This simple equation explains the pattern we saw earlier in the influence functions for mean, variance and covariance, where the influence function for an observation was just the moment condition evaluated at that particular observation. This will hold true for all higher order moments of a distribution as well.

3.3 General influence function for OLS

Consider OLS as a subset of GMM. Moment conditions are of the form $[X_i'(Y_i - X_i\beta)] = 0$. We can think about a homoscedastic version of OLS, where $W = \frac{1}{\sigma}I$. So the influence function for an OLS estimate is:

$$\psi_\theta(x) = \mathrm{E}\left[X'X\right]^{-1} x_i'(y_i - x_i\beta)$$

Looking at the simple linear regression example above, it's clear that these two influence functions are consistent.

4 Influence functions and variance

Note for this class of estimators it must be the case $\mathrm{E}\left[\psi_\theta(x)\right] = 0$. This makes the computation of variances and covariances for influence functions extremely easy:

$$\mathrm{Cov}(\psi_{\theta_1}(x), \psi_{\theta_2}(x)) = \mathrm{E}\left[\psi_{\theta_1}(x)\psi_{\theta_2}(x)\right] - \mathrm{E}\left[\psi_{\theta_1}(x)\right]\mathrm{E}\left[\psi_{\theta_2}(x)\right] = \mathrm{E}\left[\psi_{\theta_1}(x)\psi_{\theta_2}(x)\right] \tag{13}$$

Why is this useful? Say we have an estimate $\hat{\theta}$ from a random sample. We can look at this sample as a series of ϵ-contaminations to the true distribution, each of which puts $1/n$ weight on the derivative. Then for large enough n we can represent the difference between our $\hat{\theta}$ and the true θ by use of a Taylor expansion:

$$\hat{\theta} = \theta + \sum_{i=1}^{n}\psi_\theta(x_i)\frac{1}{n} + \text{ many higher order terms}$$

These higher order terms will converge in probability to zero, even when multiplied by $\sqrt{n}$. This implies that:

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_\theta(x_i) + o_p(1) \tag{14}$$

So the asymptotic distribution of the estimator and the asymptotic distribution of the influence function are directly related. This means that we have now come up with a hugely convenient way to find

the variances and covariances of a variety of estimators: using the same properties of influence functions.

Why is this particularly useful? Imagine that you want to know the covariance between the coefficients of a linear regression and the autocorrelation of one of your variables. Without influence functions, you have three options available to you:

1. Bootstrap your estimates.
2. Make structural assumptions and derive the covariance.
3. Estimate all the above as a GMM system.

Option 1, bootstrapping, is probably what most researchers would think to do first. But bootstrapping is computationally intensive, you have to sample and resample hundreds of times and if you have large data or a complicated estimator this process can be prohibitively expensive. Bootstraps can also have poor finite sample performance. Option 2, making structural assumptions, usually requires a good deal of work on the derivation, and at the end of all that work your results will still depend on the parametric assumptions made. Option 3 is actually equivalent to using influence functions, but the way most researchers approach this method requires re-estimating the linear regression and autocorrelation. Again, for more complicated estimators this can be time consuming.

Using influence functions provides an easy way to sidestep these three time-consuming options. You calculate the empirical equivalents of the influence functions for each estimate, and stack them. If you have estimators $\theta_1, \ldots, \theta_M$ and observations $i = 1, \ldots, N$ you create a matrix:

$$\Psi = [\psi_{\theta_1}, \ldots, \psi_{\theta_N}]_{N \times M}$$

This matrix Ψ has columns corresponding to each estimator, and rows corresponding to each observation. Every element of Ψ, $\Psi_{i,j}$ is equal to $\psi_{\theta_j}(x_i)$. Since the distribution of each estimator is the same as $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_{\theta_j}(x_i)$, we can use these to easily calculate a consistent estimate of the variance and covariance of any set estimator of estimators by simply taking:

$$V = \frac{1}{N^2} \left(\Psi^T \Psi \right)$$

Moreover, this calculation of variance is computationally inexpensive, since influence functions are produced as a side product of many estimation routines (because of their application to Newton iterations) and are generally easy to calculate.