BIOST/STAT 578B

Modern inference in infinite-dimensional models

Chapter 2: Asymptotic linearity and calculus of influence functions

Marco Carone
Department of Biostatistics
School of Public Health, University of Washington

Winter 2015

Contents of this chapter

- Definition and general properties
- Examples of asymptotically linear estimators and their influence function
- Transformations of asymptotically linear estimators
- Example: IPTW estimator of the mean counterfactual outcome

Definition and general properties

Suppose we ask individual 1 to estimate a feature $\psi_{01} := \Psi_1(P_0)$ of P_0 based on observations $O_1, O_2, \ldots, O_n$ independently drawn from P_0, and we ask individual 2 to estimate $\psi_{02} := \Psi_2(P_0)$ of P_0 based on this same data.

Our goal is to produce an estimator of $\psi_0 := \psi_{01} + \psi_{02}$.

The two individuals each propose an estimator, say ψ_{1n} and ψ_{2n}, which they have shown to be asymptotically normal. Specifically, they claim that

$$n^{1/2}(\psi_{1n} - \psi_{01}) \xrightarrow{d} N(0, \sigma_1^2) \quad \text{and} \quad n^{1/2}(\psi_{2n} - \psi_{02}) \xrightarrow{d} N(0, \sigma_2^2) \ .$$

We then take as estimator of ψ_0 the sum $\psi_n := \psi_{1n} + \psi_{2n}$.

Can we readily obtain distributional results for ψ_n from the provided info?

Definition and general properties

Marginal distributional results are quite limiting, because they do not help us understand the joint behavior of two given estimators.

Suppose that we can write

$$\psi_{1n} = \frac{1}{n}\sum_{i=1}^{n}\phi_1(O_i) \quad \text{and} \quad \psi_{2n} = \frac{1}{n}\sum_{i=1}^{n}\phi_2(O_i)$$

for fixed functions ϕ_1 and ϕ_2. If these estimators are consistent, it must be that $P_0\phi_j = \psi_{0j}$ for $j = 1, 2$. Then, setting $\phi_3(o) := \phi_1(o) + \phi_2(o)$, we have that

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n}[\phi_3(O_i) - P_0\phi_3]$$

and so, by the CLT, $n^{1/2}(\psi_n - \psi_0) \xrightarrow{d} N(0, \sigma^2)$ with $\sigma^2 := P_0(\phi_3 - P_0\phi_3)^2$ provided the latter is finite.

Definition and general properties

Describing the joint behavior of sample means is incredibly easy. What about estimators that behave almost like sample means?

Suppose that ψ_{1n} and ψ_{2n} are *asymptotically linear* in that we can write

$$\psi_{1n} = \frac{1}{n}\sum_{i=1}^{n}\phi_1(O_i) + o_P(n^{-1/2}) \quad \text{and} \quad \psi_{2n} = \frac{1}{n}\sum_{i=1}^{n}\phi_2(O_i) + o_P(n^{-1/2})$$

for fixed ϕ_1 and ϕ_2 such that $P_0\phi_j = \psi_{0j}$ for $j = 1, 2$. Then, we can write

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n}[\phi_3(O_i) - P_0\phi_3] + o_P(n^{-1/2})$$

and the CLT again suffices to obtain distributional result for ψ_n.

The asymptotic distribution of ψ_n depends on ψ_{1n} and ψ_{2n} only through the functions ϕ_1 and ϕ_2, or equivalently, through $\phi_1 - P_0\phi_1$ and $\phi_2 - P_0\phi_2$. The latter are called the **influence functions** of ψ_{1n} and ψ_{2n}, respectively.

Definition and general properties

A general estimator ψ_n of $\psi_0 := \Psi(P_0)$ based on $O_1, O_2, ..., O_n \overset{iid}{\sim} P_0$ is said to be **asymptotically linear** if there exists a function $o \mapsto \mathrm{IC}_{P_0}(o)$ such that

1. $\mathrm{IC}_{P_0}(O)$ has mean zero and finite variance;
2. ψ_n admits the representation

$$\psi_n = \psi_0 + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_{P_0}(O_i) + o_P(n^{-1/2}) \ .$$

The function IC is called the **influence function** (or curve) of ψ_n.

The asymptotic study of ψ_n reduces simply to the study of sample means.

- Weak Law of Large Numbers $\longrightarrow$ consistency;
- Central Limit Theorem $\longrightarrow$ asymptotic normality.

As discussed so far, it is also useful for studying the joint behavior of estimators.

Suppose that ψ_n is an asymptotically linear estimator of ψ_0 with influence function IC_{P_0}. The multivariate CLT tells us that

$$n^{1/2}\left(\psi_n - \psi_0\right) \xrightarrow{d} MVN(0, \Sigma_0) \ ,$$

where $\Sigma_0 := P_0\left[\mathrm{IC}_{P_0}\mathrm{IC}_{P_0}^\top\right]$ is the asymptotic variance-covariance matrix.

If $\hat{P}$ is a consistent estimator of P_0, a consistent estimator of Σ_0 can often be constructed as

$$\Sigma_n := \frac{1}{n}\sum_{i=1}^{n}\mathrm{IC}_{\hat{P}}(O_i)\mathrm{IC}_{\hat{P}}^\top(O_i) \ .$$

An approximate $100 \times (1-\alpha)\%$ confidence interval for $\psi_{0,r}$, the r^{th} component of ψ_0, is then given by

$$\left(\psi_{n,r} - q_{1-\alpha/2}\Sigma_{n,rr}^{1/2}n^{-1/2},\ \psi_{n,r} + q_{1-\alpha/2}\Sigma_{n,rr}^{1/2}n^{-1/2}\right),$$

where q_β is the β-quantile of the standard normal distribution.

Definition and general properties

The nomenclature *influence function* seems to have been coined by F. Hampel in his 1968 PhD thesis focusing on the study of robust statistics.

- The influence function is called so because it allows us to identify the extent to which a single outlier can drive off our estimator.
- An estimator with a bounded influence function is generally considered robust and often preferred in practice.
- If the influence function is bounded, say by $M < +\infty$, a single outlier can only displace the estimator by about M/n in large samples.
- For both theoretical and practical purposes, it is useful to determine conditions on P_0 such that the influence curve is indeed bounded.
- For smooth functionals of the empirical distribution function, the influence function can be obtained as a Gâteaux derivative in the direction of a singleton distribution.

Sample mean

Suppose that $O_1, O_2, ..., O_n \overset{iid}{\sim} P_0$ and $\Psi(P_0) := \int u dP_0(u)$, denote by μ_0, is the true mean.

The sample mean $\mu_n := \Psi(P_n) = \frac{1}{n}\sum_{i=1}^{n} O_i$ may be written as

$$\mu_n = \mu_0 + \frac{1}{n}\sum_{i=1}^{n}(O_i - \mu_0) \; ,$$

so that its influence curve is $\mathrm{IC}_{P_0}(o) := o - \mu_0$. The sample mean is not only asymptotically linear: it is linear in finite samples.

Sample variance

Suppose that $O_1, O_2, ..., O_n \overset{iid}{\sim} P_0$, $\Psi_1(P_0) := \int u dP_0(u)$, denoted by μ_0, is the mean, and $\Psi_2(P_0) := \int (u - \Psi_1(P_0))^2 dP_0(u)$, denoted by σ_0^2, is the variance.

The sample variance $\sigma_n^2 := \Psi_2(P_n) = \frac{1}{n} \sum_{i=1}^n (O_i - \mu_n)^2$ may be written as

$$\sigma_n^2 = \sigma_0^2 + \frac{1}{n} \sum_{i=1}^n \left[(O_i - \mu_0)^2 - \sigma_0^2 \right] - \left[\frac{1}{n} \sum_{i=1}^n (O_i - \mu_0) \right]^2 .$$

The last summand is $o_P(1)$ by the CLT and the Continuous Mapping Theorem, and it follows that σ_n^2 has influence curve $\mathrm{IC}_{P_0}(o) := (o - \mu_0)^2 - \sigma_0^2$.

We note here that $\frac{1}{n} \sum_{i=1}^n (O_i - \mu_0)^2$ has the same influence function as σ_n^2: having had to estimate the nuisance μ_0 is without consequence asymptotically.

This is an example of an *adaptive estimator* – see, e.g., Bickel (1982).

Examples of asymptotically linear estimators and their influence function

Estimating equation-based estimator

Suppose that $O_1, O_2, ..., O_n \overset{iid}{\sim} P_0 \in \mathcal{M}$ and we wish to estimate a feature $\psi_0 := \Psi(P_0)$ using these data.

Suppose that $o \mapsto U(\psi)(o)$ is such that $P_0 U(\psi) = 0$ has unique solution ψ_0. Then, U is said to be an **unbiased estimating equation** for ψ_0, and we consider the estimator ψ_n of ψ_0 defined as the solution of

$$0 = P_n U(\psi) = \frac{1}{n} \sum_{i=1}^{n} U(\psi)(O_i) \ .$$

Can we derive the asymptotic behavior of ψ_n despite its implicit definition?

Examples of asymptotically linear estimators and their influence function

Estimating equation-based estimator

Under sufficient regularity conditions (i.e., sweeping details under the carpet), we can use a first-order Taylor approximation to show that

$$\psi_n = \psi_0 - \frac{1}{n} \sum_{i=1}^{n} \left[P_0 \dot{U}(\psi_0) \right]^{-1} U(\psi_0)(O_i) + o_P(n^{-1/2})$$

with $\dot{U}(\psi) := \frac{\partial}{\partial \psi} U(\psi)$, so that ψ_n has influence curve

$$\mathrm{IC}_{P_0}(o) := -[P_0 \dot{U}(\psi_0)]^{-1} U(\psi_0)(o) \ .$$

This implies that $n^{1/2} (\psi_n - \psi_0) \xrightarrow{d} MVN(0, \Sigma)$, where $\Sigma := RSR^\top$ with $R := [P_0 \dot{U}(\psi_0)]^{-1}$ and $S := U(\psi_0) U(\psi_0)^\top$.

What is the natural parallel between the above and our knowledge of maximum likelihood estimators in parametric models?

Examples of asymptotically linear estimators and their influence function

Sample quantile

Suppose that $O_1, O_2, ..., O_n \stackrel{iid}{\sim} P_0$ have cdf F_0 and density f_0. Denote the p^{th} quantile of P_0 by $Q_0(p)$, and define $Q_n(p) := \inf\{y : F_n(y) \geq p\}$ to be the sample p^{th} quantile, with F_n the empirical cdf. Suppose that $f_0(Q_0(p)) > 0$.

It is possible to show that

$$Q_n(p) = Q_0(p) + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{F_0(Q_0(p)) - I(O_i \leq Q_0(p))}{f_0(Q_0(p))}\right] + o_P(n^{-1/2}) \ ,$$

so that $Q_n(p)$ has influence curve

$$\mathrm{IC}_{P_0}(o) = \frac{F_0(Q_0(p)) - I(o \leq Q_0(p))}{f_0(Q_0(p))} \ .$$

Sample quantile

Suppose that $f_0 > 0$ on its support, so that $F_0(Q_0(p)) = p$ for each $p \in (0, 1)$. Write $\mathbf{Q}_n = (Q_n(p_1), Q_n(p_2), ..., Q_n(p_K))$ and $\mathbf{Q}_0 = (Q_0(p_1), Q_0(p_2), ..., Q_0(p_K))$ for given $0 < p_1 < p_2 < ... < p_K < 1$.

As a consequence of the asymptotic representation of the sample quantile, we may conclude that $\sqrt{n}\,(\mathbf{Q}_n - \mathbf{Q}_0)$ is asymptotically normal with zero mean and covariance matrix Σ with ijth element

$$\Sigma_{ij} \;=\; \frac{\max(p_i, p_j) - p_i p_j}{f_0(Q_0(p_i))f_0(Q_0(p_j))} \;.$$

Transformations of asymptotically linear estimators

Many estimators can be written as functions of simpler, asymptotically linear estimators. This can be leveraged in our asymptotic study.

Suppose that ψ_n is an asymptotically linear estimator of $\psi_0 := \Psi(P_0) \in \mathbb{R}^p$ with influence curve IC_{P_0} based on independent draws $O_1, O_2, ..., O_n$ from P_0.

Suppose also that h is differentiable at ψ_0 with nonzero derivative $h'(\psi_0)$. Under regularity conditions, we have that

$$h(\psi_n) = h(\psi_0) + \frac{1}{n}\sum_{i=1}^{n} h'(\psi_0)^\top \mathrm{IC}_{P_0}(O_i) + o_P(n^{-1/2}) \ ,$$

implying that $h(\psi_n)$ is asymptotically linear with influence function $h'(\psi_0)\mathrm{IC}_{P_0}$.

This is the **delta method for influence curves**.

Sample coefficient of variation:

Suppose that $O_1, O_2, ..., O_n \overset{iid}{\sim} P_0$, and denote the mean and variance of P_0 by $\mu_0 := \int u dP_0(u)$ and $\sigma_0^2 := \int (u-\mu_0)^2 dP_0(u)$, respectively. We consider $C_n := \sigma_n/\mu_n$ as a natural estimator of the coefficient of variation $C_0 := \sigma_0/\mu_0$.

Let $h(x,y) := x^{1/2}y^{-1}$ and note that $C_0 := h(\sigma_0^2, \mu_0)$ and $C_n := h(\sigma_n^2, \mu_n)$. Using that $h'(x,y) = (x^{-1/2}y^{-1}/2, -x^{1/2}y^{-2})^\top$, we get that

$$C_n \;=\; C_0 + \frac{1}{n}\sum_{i=1}^{n} h'(\sigma_0^2, \mu_0)^\top \begin{bmatrix} (O_i-\mu_0)^2 - \sigma_0^2 \\ O_i - \mu_0 \end{bmatrix} + o_P(n^{-1/2})$$

and upon simplification, we find that C_n has influence curve given by

$$\mathrm{IC}_{P_0}(o) := C_0\left[\frac{1}{2}\left(\frac{o-\mu_0}{\sigma_0}\right)^2 - \frac{o}{\mu_0} + \frac{1}{2}\right].$$

Example: IPTW estimator of the mean counterfactual outcome

Suppose that $O := (W, A, Y)$ and we observe $O_1, O_2, \ldots, O_n \overset{iid}{\sim} P_0$.

Under causal assumptions, the mean counterfactual outcome corresponding to treatment level $A = 1$ is given by $\Psi(P) := E_P\left[E_P\left(Y|A = 1, W\right)\right]$.

It will be useful to define the following quantities:

$$\begin{aligned} \bar{Q}(w,a) &:= E_P(Y|A = a, W = w)\ , \\ g(w) &:= P(A = 1|W = w)\ , \\ Q_W(w) &:= P(W \leq w)\ . \end{aligned}$$

Writing $\psi_0 := \Psi(P_0)$, we observe first that

$$E_0\left[\frac{AY}{g_0(W)}\right] = E_0\left[\bar{Q}_0(W, 1)\right] = \psi_0\ .$$

Example: IPTW estimator of the mean counterfactual outcome

If g_0 is known, this motivates the use of the IPTW estimator

$$\psi_{0n} := \frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{g_0(W_i)} =: P_n f_0 \ ,$$

where $f_0(o) := ay/g_0(w)$.

We see that ψ_{0n} is linear with influence function $\mathrm{IC}_{P_0}(o) := f_0(o) - \psi_0$.

Suppose now that g_0 is only known to lie in a parametric model $\{g_\theta : \theta \in \Theta\}$, i.e., there exists some $\theta_0 \in \Theta$ such that $g_0 = g_{\theta_0}$. Suppose also that θ_n is an asymptotically linear estimator of θ_0 with influence function ϕ_{θ_0}.

Writing $g_n := g_{\theta_n}$ and $f_n(o) := ay/g_n(w)$, we may now consider the estimator

$$\psi_n := \frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{g_n(W_i)} = P_n f_n \ .$$

Example: IPTW estimator of the mean counterfactual outcome

We wish to study the asymptotic behavior of the IPTW estimator ψ_n.

We observe that we can write

$$\begin{aligned}\psi_n - \psi_0 \;=\; P_n f_n - P_0 f_0 \;&=\; (P_n - P_0) f_0 + P_0(f_n - f_0) + (P_n - P_0)(f_n - f_0)\\ &=\; P_n(f_0 - P_0 f_0) + P_0(f_n - f_0) + (P_n - P_0)(f_n - f_0)\;.\end{aligned}$$

The blue term is already linear.

We will ignore the orange term for now – it will generally be negligible under some conditions. Empirical process theory is useful to establish this.

The green term requires more work, as will generally be the case!

Example: IPTW estimator of the mean counterfactual outcome

First, we note that

$$g_n(w) - g_0(w) \;=\; g_{\theta_n}(w) - g_{\theta_0}(w) \;=\; \left.\frac{\partial}{\partial\theta} g_\theta(w)\right|_{\theta=\theta_0} (\theta_n - \theta_0) + o_P(n^{-1/2}) \;.$$

Under certain regularity conditions, we can then show that

$$\begin{aligned} P_0(f_n - f_0) \;&=\; \int \bar{Q}_0(w,1) g_0(w) \left[\frac{1}{g_n(w)} - \frac{1}{g_0(w)}\right] dQ_{W,0}(w) \\ &=\; -\int \frac{\bar{Q}_0(w,1)}{g_0(w)} \left[g_n(w) - g_0(w)\right] dQ_{W,0}(w) + o_P(n^{-1/2}) \\ &=\; \frac{1}{n}\sum_{i=1}^{n} \gamma_0 \phi_{\theta_0}(O_i) + o_P(n^{-1/2}) \;, \end{aligned}$$

where $\gamma_0 := -\int \bar{Q}_0(w,1) g_0(w)^{-1} \left.\frac{\partial}{\partial\theta} g_\theta(w)\right|_{\theta=\theta_0} dQ_{W,0}(w)$.

Example: IPTW estimator of the mean counterfactual outcome

We conclude that ψ_n is an asymptotically linear estimator of ψ_0 with influence function given by

$$\mathrm{IC}^*_{P_0}(o) := \mathrm{IC}_{P_0}(o) + \gamma_0 \phi_{\theta_0}(o) \ .$$

If θ_n is an efficient estimator of θ_0, we can indeed verify that

$$\mathrm{var}_{P_0}\left[\mathrm{IC}^*_{P_0}(O)\right] \ \leq \ \mathrm{var}_{P_0}\left[\mathrm{IC}_{P_0}(O)\right].$$

This result can be very counterintuitive: we can **increase estimation efficiency by ignoring certain knowledge!!!**

This follows from a general result: it is generally recommended to ignore information about so-called *orthogonal nuisance parameters*. More on this later!

Relevant references

- Bickel, PJ (1982). On adaptive estimation. The Annals of Statistics.
- Hampel FR (1968). Contribution to the theory of robust estimation. PhD Thesis, UC Berkeley.