Econ 2110, fall 2016, Part IVb
Asymptotic Theory:
δ-method and M-estimation

Maximilian Kasy

Department of Economics, Harvard University

Example

- Suppose we estimate the average effect of class size on student exam grades, using the project STAR data.
- What is the variance of our estimator?
- Can we form a confidence set for the size of the effect?
- Can we reject the null hypothesis of a zero average effect?
- Also if exam scores are not normally distributed?

Example

- Suppose we estimate the top 1% income share using data on the number of individuals in different tax brackets,
- assuming that top incomes are Pareto distributed.
- Suppose we calculate the implied optimal top tax rate.
- Can we form a 95% confidence interval for this optimal tax rate?

Takeaways for this part of class

- How we get our formulas for standard deviations in many settings.
- When and why we can expect asymptotic normality for many estimators (and what that means).
- When we might expect problems to arise for asymptotic approximations.

Roadmap

- IVa
  - Types of convergence
  - Laws of large numbers (LLN)
    and central limit theorems (CLT)
- IVb
  - The delta method
  - M- and Z-Estimators
  - Special M-Estimators
    - Ordinary least squares (OLS)
    - Maximum likelihood estimation (MLE)
  - Confidence sets

Part IVb

The delta method

- Suppose we know the asymptotic behavior of sequence X_n,
- we are interested in $Y_n = g(X_n)$, and
- g is "smooth."
- Often a Taylor expansion of g around the probability limit of X_n yields the answer,
- where we can ignore higher order terms in the limit.

$$Y_n = g(\beta) + g'(\beta) \cdot (X_n - \beta) + o(\|X_n - \beta\|).$$

- This idea is called the delta method.

Theorem (Delta method)

Assume that

$$r_n \cdot (X_n - \beta) \to^d X$$

for some sequence $r_n \to \infty$ and some random variable X.
Let $Y_n = g(X_n)$ for a function g which is differentiable at β.
Then

$$r_n \cdot (Y_n - g(\beta)) \to^d g'(\beta) \cdot X.$$

Proof:

- By differentiability of g,

$$Y_n = g(\beta) + g'(\beta) \cdot (X_n - \beta) + o(\|X_n - \beta\|).$$

- Rearranging gives

$$r_n \cdot (Y_n - g(\beta)) = r_n \cdot g'(\beta) \cdot (X_n - \beta) + r_n \cdot o(\|X_n - \beta\|).$$

- The second term vanishes asymptotically, since $r_n \cdot (X_n - \beta)$ converges in distribution.
- The continuous mapping theorem applied to matrix multiplication by $g'(\beta)$ now yields the claim.

Leading special case

- Let X_n be a sequence of random variables such that

$$\sqrt{n}(X_n - b) \to^d \mathscr{N}(0, \sigma^2).$$

- Let $g : \mathbb{R} \mapsto \mathbb{R}$ be continuously differentiable at a.
- Then

$$\sqrt{n}(g(X_n) - g(b)) \to^d \mathscr{N}(0, (g'(b))^2\sigma^2).$$

Attention

- There are important cases where the delta method provides poor approximations
- Examples: near $\beta = 0$, for
  1. $g(X) = |X|$
  2. $g(X) = 1/X$
  3. $g(X) = \sqrt{X}$
- Relevant for:
  1. weak instruments
  2. inference under partial identification / moment inequalities

Practice problem

- Suppose X_i are iid with mean 1 and variance 2, and $n = 25$.
- Let $Y = \overline{X}^2$.
- Provide an approximation for the distribution of Y.
- Now suppose X_i has mean 0 and variance 2.
- Provide an approximation for the distribution of Y.

M- and Z-Estimators

- Many interesting objects β can be written in the form

$$\beta_0 = \underset{\beta}{\operatorname{argmax}}\ E[m(\beta, X)]. \tag{1}$$

- This defines a mapping
  from the probability distribution of X
  to a parameter β.
- In our decision theory notation:

$$\beta_0 = \beta(\theta)$$

Example - Least squares

- The coefficients β_0
- of the best linear predictor

$$\widehat{Y} = X \cdot \beta_0$$

- minimize the average squared prediction error,

$$\beta_0 = \underset{\beta}{\operatorname{argmin}} \; E[(Y - X \cdot \beta)^2].$$

- Thus

$$m(\beta, X, Y) = (Y - X \cdot \beta)^2.$$

Example - Maximum likelihood

- Suppose Y is distributed according to the density

$$Y \sim f(Y, \beta_0).$$

- Then β_0 maximizes the expected log likelihood,

$$\beta_0 = \underset{\beta}{\operatorname{argmax}} \; E[\log(f(Y, \beta))].$$

- We will show this later.
- Thus

$$m(\beta, X) = \log(f(Y, \beta)).$$

M-Estimator

- Use E_n to denote the sample average, e.g.

$$E_n[X] = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- We can define an **estimator** for β which solves the **analogous** conditions
- replacing the population expectation by a sample average,
- that is

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmax}}\ E_n[m(\beta, X)]. \tag{2}$$

- Such an estimator is called an M-estimator (for “maximizer”).

Examples continued

1. **Least squares:**
   ordinary least squares (OLS) estimator

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}}\ E_n[(Y - X \cdot \beta)^2]$$

2. **Maximum likelihood:**
   maximum likelihood estimator (MLE)

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmax}}\ E_n[\log(f(Y, \beta))]$$

Z-Estimator

- If m is differentiable and β is an interior maximizer equation (1) implies the first order conditions

$$\frac{\partial}{\partial \beta} E[m(\beta, X)] = E[m'(\beta_0, X)] = 0.$$

- If we directly define the estimator via

$$E_n[m'(\widehat{\beta}, X_i)] = 0, \tag{3}$$

then $\widehat{\beta}$ is called a Z-estimator (for "zero").

Practice problem

Find the first order conditions for MLE and for OLS

Solution:

1. **Least squares:**

$$E_n[\widehat{e} \cdot X] = 0$$

where

$$\widehat{e} = Y - X \cdot \widehat{\beta}$$

is the regression residual.

2. **Maximum likelihood:**

$$E_n\left[S(Y, \widehat{\beta})\right] = 0$$

where

$$S(Y, \beta) := \frac{\partial}{\partial \beta} \log(f(Y, \beta))$$

is called the score.

Consistency

- Basic requirement for good estimators:
- That they are close to the population estimand
  with large probability
  as sample sizes get large:

$$P(\|\widehat{\beta} - \beta_0\| < \varepsilon) \to 1 \quad \forall \varepsilon.$$

- Thus:

$$\widehat{\beta} \to^p \beta_0$$

- This property is called **consistency**.

Theorem (Consistency of M-Estimators)

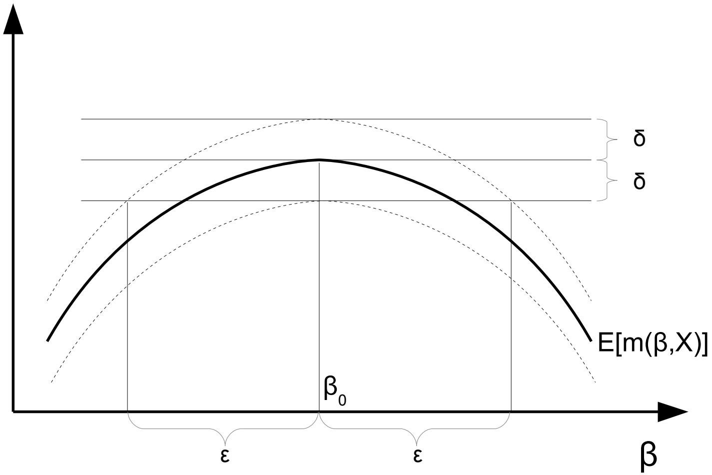
M-estimators are consistent if

1. $$\sup_{\beta} \| E_n[m(\beta, X)] - E[m(\beta, X)] \| \to^p 0$$

2. $$\sup_{\beta:\, \|\beta - \beta_0\| > \varepsilon} E[m(\beta, X)] < E[m(\beta_0, X)].$$

- The first condition holds in many case by some "uniform law of large numbers."
- The second condition states that the maximum is "well separated."

Figure: Proof of consistency



Sketch of proof:

- By assumption (2), for every ε there is a δ, such that if

$$\sup_{\beta} \|E_n[m(\beta,X)] - E[m(\beta,X)]\| < \delta$$

then $\|\widehat{\beta} - \beta_0\| < \varepsilon$.

- By assumption (1),

$$\sup_{\beta} \|E_n[m(\beta,X)] - E[m(\beta,X)]\| < \delta$$

happens with probability going to 1 as $n \to \infty$.

Asymptotic normality

- What is the (approximate) distribution of M-estimators?
- Consistency just states that they converge to a point.
- But if we “blow up” the scale appropriately?
- For instance by $\sqrt{n}$?
- Then we get convergence to a normal distribution!

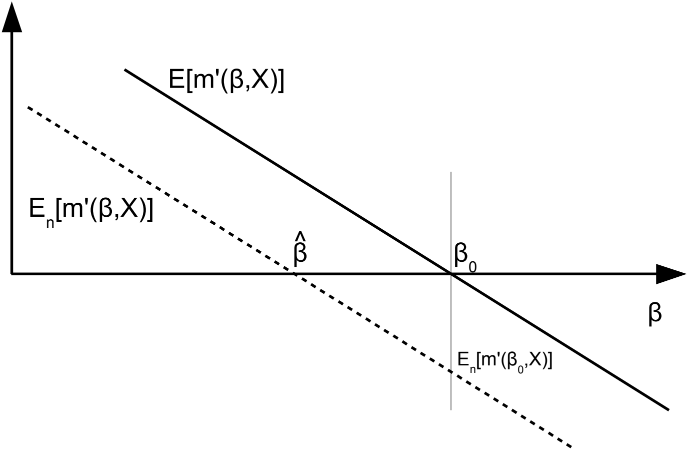
Theorem

Under suitable differentiability conditions, M-estimators and Z-estimators are asymptotically normal,

$$\sqrt{n}(\widehat{\beta} - \beta_0) \to^d N(0, V)$$

for some V.

Figure: Proof of asymptotic normality



Sketch of proof:

- follows by arguments similar to our derivation of the delta method.
- if m is twice differentiable, by the intermediate value theorem

$$0 = E_n[m'(\widehat{\beta}, X)] = E_n[m'(\beta_0, X)] + E_n[m''(\tilde{\beta}, X)] \cdot (\widehat{\beta} - \beta_0)$$

  for some $\tilde{\beta}$ between $\widehat{\beta}$ and β_0.
- Rearranging yields

$$\sqrt{n}(\widehat{\beta} - \beta_0) = -\left(E_n[m''(\tilde{\beta}, X)]\right)^{-1} \cdot \sqrt{n} E_n[m'(\beta_0, X)].$$

- Consistency of $\widehat{\beta}$ and a uniform law of large numbers for m'' imply

$$\left(E_n[m''(\tilde{\beta}, X)]\right)^{-1} \to^p \left(E[m''(\beta_0, X)]\right)^{-1}.$$

- The central limit theorem implies

$$\sqrt{n}E_n[m'(\beta_0, X)] \to^d N(0, \mathrm{Var}(m'(\beta_0, X))).$$

- Slutsky's theorem then yields the asymptotic distribution of $\widehat{\beta}$ as

$$\sqrt{n}(\widehat{\beta} - \beta_0) \to^d N(0, V)$$

where

$$V = \left(E[m''(\beta_0, X)]\right)^{-1} \cdot \mathrm{Var}(m'(\beta_0, X)) \cdot \left(E[m''(\beta_0, X)]\right)^{-1}. \quad (4)$$

Estimators of the asymptotic variance

- Asymptotic variance: “sandwich” form
- Estimators for this variance:
  sample analogs of both components
- For instance:

$$\widehat{V} = \left(E_n[m''(\widehat{\beta}, X)]\right)^{-1} \cdot E_n\left[(m'(\widehat{\beta}, X))^2\right] \cdot \left(E_n[m''(\widehat{\beta}, X)]\right)^{-1}$$

- This is the kind of variance estimator you get when you type

  `, robust`

  after some estimation commands in Stata.

Least squares

- Recall OLS:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}}\ E_n[(Y - X \cdot \beta)^2]$$

- First order condition:

$$E_n[e \cdot X] = 0$$

where

$$e := Y - X \cdot \widehat{\beta}$$

- In our general notation:

$$m(Y, X, \beta) = e^2 = (Y - X \cdot \beta)^2$$
$$m'(Y, X, \beta) = -2e \cdot X$$
$$m''(Y, X, \beta) = 2 \cdot XX^t$$

- Apply the asymptotic results for general M-estimators
- $\widehat{\beta}$ is consistent for β_0, the "best linear predictor,"

$$\beta_0 = \underset{\beta}{\operatorname{argmin}}\ E[(Y - X \cdot \beta)^2].$$

- $\widehat{\beta}$ is asymptotically normal

$$\sqrt{n} \cdot \left(\widehat{\beta} - \beta_0\right) \to^d N(0, V)$$

- Asymptotic variance

$$V = \left(E[m''(\beta_0, X)]\right)^{-1} \cdot \mathrm{Var}(m'(\beta_0, X)) \cdot \left(E[m''(\beta_0, X)]\right)^{-1}$$
$$= E[XX^t]^{-1} \cdot E[e^2 XX^t] \cdot E[XX^t]^{-1}$$

- "heteroskedasticity robust variance estimator for ordinary least squares:"

$$\frac{1}{n} \cdot E_n[XX^t]^{-1} \cdot E_n[\widehat{e}^2 XX^t] \cdot E_n[XX^t]^{-1} \tag{5}$$

- Factor of $1/n$ to get variance of $\widehat{\beta}$ rather than $\sqrt{n} \cdot \widehat{\beta}$

Maximum likelihood

Lemma

- Suppose $Y \sim f(y, \beta_0)$,
- where f denotes a family of densities indexed by β.
- Then
$$E[\log(f(Y, \beta_0))] \geq E[\log(f(Y, \beta))]. \tag{6}$$
- The inequality is strict if $f(Y, \beta_0) \neq f(Y, \beta)$ with positive probability.

Sketch of proof:

- Want to show:

$$0 \geq \int \log\left(f(y,\beta)\right) f(y,\beta_0)dy - \int \log\left(f(y,\beta_0)\right) f(y,\beta_0)dy$$
$$= \int \log\left(\frac{f(y,\beta)}{f(y,\beta_0)}\right) f(y,\beta_0)dy.$$

- Jensen's inequality, applied to the concave function log:

$$\int \log\left(\frac{f(y,\beta)}{f(y,\beta_0)}\right) f(y,\beta_0)dy$$
$$\leq \log\left(\int \frac{f(y,\beta)}{f(y,\beta_0)} f(y,\beta_0)dy\right)$$
$$= \log(1) = 0.$$

Terminology for maximum likelihood

- Log likelihood:

$$L_n(\beta) = n \cdot E_n[m(Y,\beta)] = \sum_i \log(f(Y_i,\beta))$$

- Score:

$$S_i(\beta) = m'(Y_i,\beta) = \frac{\partial}{\partial\beta}\log(f(Y_i,\beta))$$

- Information:

$$I(\beta) = -E[m''(Y,\beta)] = -E[\partial S/\partial\beta]$$

Lemma

If $Y_i \sim f(y, \beta_0)$, then

$$\mathrm{Var}(S(\beta_0)) = I(\beta_0) = -E[\partial S(\beta_0)/\partial \beta].$$

- **Proof:**
  Differentiate $0 = E[S] = \int S(y, \beta_0) f(y, \beta_0) dy$ with respect to β_0 to get
  $$0 = \int S' f dy + \int S f' dy = E[S'] + E[S^2].$$
- **But:**
  Parametric models are usually wrong.
  So don't trust this equality.
- If it holds, the asymptotic variance for the MLE simplifies to
  $$V = E[S']^{-1} \cdot E[S^2] \cdot E[S']^{-1} = I(\beta_0)^{-1}.$$

Confidence sets

- **Confidence set** C:
  a set of βs,
  which is calculated as a function of data Y
- Confidence set C for β **of level** α:

$$P(\beta_0 \in C) \geq 1 - \alpha. \tag{7}$$

  for all distributions of Y (i.e., all θ)
  and corresponding β_0.
- In this expression β_0 is fixed and C is random.
- Confidence set C_n for β of **asymptotic level** α:

$$\lim_{n \to \infty} P(\beta \in C_n) \geq 1 - \alpha. \tag{8}$$

Confidence sets for M-estimators

- can use asymptotic normality to get asymptotic confidence sets
- Suppose

$$\sqrt{n}(\widehat{\beta} - \beta_0) \to^d N(0, V)$$
$$\widehat{V} \to^p V$$

- Define

$$\tilde{\beta} := \sqrt{n} \cdot \widehat{V}^{-1/2} \cdot (\widehat{\beta} - \beta_0).$$

- Slutsky's theorem ⇒

$$\tilde{\beta} \to^d N(0, I),$$

and therefore

$$\|\tilde{\beta}\|^2 \to^d \chi^2_k,$$

where $k = \dim(\beta)$.

- Let $\chi^2_{k,1-\alpha}$ be the $1-\alpha$ quantile of the χ^2_k distribution.
- Define

$$C_n = \left\{\beta : \|\sqrt{n} \cdot V^{-1/2} \cdot (\widehat{\beta} - \beta)\|^2 \leq \chi^2_{k,1-\alpha}\right\}. \tag{9}$$

- We get

$$P(\beta_0 \in C_n) \to 1 - \alpha.$$

- C_n is a confidence set for β of asymptotic level α.
- C_n is an ellipsoid.