Asymptotic Efficiency for Two-Stage Conditional M-estimators

Aristide Houndetoungan [1] Abdoul Haki Maoude [2]

[1] Cy Cergy Paris Université

[2] Concordia University

ESSEC RESEARCH DAY

April 13, 2023

Sequential Estimation Methods

- Models that cannot be estimated using a single-step approach.

  Examples include endogeneity issues, missing data, unobserved regressors, or when many DGPs are involved.

- Asymptotic properties of the estimator at the final stage.

  Estimator may not be normally distributed.

- Solutions

  1. Bootstrap approach.

     Time-consuming and sometimes infeasible for complex models.

  2. Orthogonal or "immunized" equations at the final stage that are locally insensitive to small mistakes in the prior estimates (Chernozhukov, Hansen, and Spindler 2015, Annu. Rev.).

Sequential Estimation Methods

- Models that cannot be estimated using a single-step approach.

  Examples include endogeneity issues, missing data, unobserved regressors, or when many DGPs are involved.

- Asymptotic properties of the estimator at the final stage.

  Estimator may not be normally distributed.

- Solutions

  1. Bootstrap approach.

     Time-consuming and sometimes infeasible for complex models.

  2. Orthogonal or "immunized" equations at the final stage that are locally insensitive to small mistakes in the prior estimates (Chernozhukov, Hansen, and Spindler 2015, Annu. Rev.).

This Paper

- Two-stage estimation strategy where the second stage leads to an M-estimator (conditional M-estimator).
- Objective function at the second stage

$$Q_n(\theta, \mathbf{y}_n, \mathbf{X}_n, \hat{\boldsymbol{\beta}}_n) = \frac{1}{n}\sum_{i=1}^{n} q_{n,i}(\theta, \hat{\boldsymbol{\beta}}_n). \tag{1}$$

- The estimator $\hat{\boldsymbol{\beta}}_n$ is general but consistent (e.g., posterior mean) and the practitioner should be able to simulate proposals from the distribution of $\hat{\boldsymbol{\beta}}_n$.
- A straightforward approach to estimate the distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$.
  - Take into account the uncertainty at the first stage (relevant for small samples).
  - Computationally more attractive than the Bootstrap method.

This Paper

- Two-stage estimation strategy where the second stage leads to an M-estimator (conditional M-estimator).
- Objective function at the second stage

$$Q_n(\theta, \mathbf{y}_n, \mathbf{X}_n, \hat{\boldsymbol{\beta}}_n) = \frac{1}{n}\sum_{i=1}^{n} q_{n,i}(\theta, \hat{\boldsymbol{\beta}}_n). \tag{1}$$

- The estimator $\hat{\boldsymbol{\beta}}_n$ is general but consistent (e.g., posterior mean) and the practitioner should be able to simulate proposals from the distribution of $\hat{\boldsymbol{\beta}}_n$.
- A straightforward approach to estimate the distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$.
  - Take into account the uncertainty at the first stage (relevant for small samples).
  - Computationally more attractive than the Bootstrap method.

This Paper

- Two-stage estimation strategy where the second stage leads to an M-estimator (conditional M-estimator).
- Objective function at the second stage

$$Q_n(\theta, \mathbf{y}_n, \mathbf{X}_n, \hat{\boldsymbol{\beta}}_n) = \frac{1}{n}\sum_{i=1}^{n} q_{n,i}(\theta, \hat{\boldsymbol{\beta}}_n). \tag{1}$$

- The estimator $\hat{\boldsymbol{\beta}}_n$ is general but consistent (e.g., posterior mean) and the practitioner should be able to simulate proposals from the distribution of $\hat{\boldsymbol{\beta}}_n$.
- A straightforward approach to estimate the distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$.
  - Take into account the uncertainty at the first stage (relevant for small samples).
  - Computationally more attractive than the Bootstrap method.

This Paper

- Two-stage estimation strategy where the second stage leads to an M-estimator (conditional M-estimator).
- Objective function at the second stage

$$Q_n(\boldsymbol{\theta}, \mathbf{y}_n, \mathbf{X}_n, \hat{\boldsymbol{\beta}}_n) = \frac{1}{n}\sum_{i=1}^{n} q_{n,i}(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_n). \tag{1}$$

- The estimator $\hat{\boldsymbol{\beta}}_n$ is general but consistent (e.g., posterior mean) and the practitioner should be able to simulate proposals from the distribution of $\hat{\boldsymbol{\beta}}_n$.
- A straightforward approach to estimate the distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$.
  - Take into account the uncertainty at the first stage (relevant for small samples).
  - Computationally more attractive than the Bootstrap method.

Examples

- Models with latent variables

$$\mathbf{E}(y|u, \mathbf{x}) = f(\theta_0 + \theta_1 u + \mathbf{x}'\boldsymbol{\theta}_2), \tag{2}$$

Popular in IO literature, where u is estimated using nonparametric methods (see Bajari, Hong, and Nekipelov 2013, Book).

Related to the Literature

- Case of M-estimators at both steps (e.g., Cameron and Trivedi 2005, Book)
- Orthogonality or "immunity" condition (Chernozhukov, Hansen, and Spindler 2015, Annu. Rev.).

  MVT at the second stage

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathbf{A}_n (1/\sqrt{n}) \nabla_\theta \sum_{i=1}^n q_{n,i}(\theta_0, \hat{\beta}_n) \quad (3)$$

  Classical CLT cannot be applied to $(1/\sqrt{n}) \nabla_\theta \sum_{i=1}^n q_{n,i}(\theta_0, \hat{\beta}_n)$.

  Implies that $(1/\sqrt{n}) \nabla_\theta \sum_{i=1}^n q_{n,i}(\theta_0, \hat{\beta}_n)$ and $(1/\sqrt{n}) \nabla_\theta \sum_{i=1}^n q_{n,i}(\theta_0, \beta_0)$ have the same distribution asymptotically.

  The CLT can be applied.

Related to the Literature

- Case of M-estimators at both steps (e.g., Cameron and Trivedi 2005, Book)
- Orthogonality or "immunity" condition (Chernozhukov, Hansen, and Spindler 2015, Annu. Rev.).

  MVT at the second stage

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \mathbf{A}_n(1/\sqrt{n})\nabla_{\boldsymbol{\theta}} \textstyle\sum_{i=1}^n q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n) \tag{3}$$

  Classical CLT cannot be applied to $(1/\sqrt{n})\nabla_{\boldsymbol{\theta}} \sum_{i=1}^n q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)$.

  Implies that $(1/\sqrt{n})\nabla_{\boldsymbol{\theta}} \sum_{i=1}^n q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)$ and $(1/\sqrt{n})\nabla_{\boldsymbol{\theta}} \sum_{i=1}^n q_{n,i}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)$ have the same distribution asymptotically.

  The CLT can be applied.

Related to the Literature

- Case of M-estimators at both steps (e.g., Cameron and Trivedi 2005, Book)
- Orthogonality or "immunity" condition (Chernozhukov, Hansen, and Spindler 2015, Annu. Rev.).

  MVT at the second stage

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \mathbf{A}_n(1/\sqrt{n})\nabla_{\boldsymbol{\theta}} \sum_{i=1}^n q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n) \tag{3}$$

  Classical CLT cannot be applied to $(1/\sqrt{n})\nabla_{\boldsymbol{\theta}} \sum_{i=1}^n q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)$.

  Implies that $(1/\sqrt{n})\nabla_{\boldsymbol{\theta}} \sum_{i=1}^n q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)$ and $(1/\sqrt{n})\nabla_{\boldsymbol{\theta}} \sum_{i=1}^n q_{n,i}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)$ have the same distribution asymptotically.

  The CLT can be applied.

Variance Estimation

- MVT at the second stage

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \mathbf{A}_n \underbrace{(1/\sqrt{n})\nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)}_{C_n}$$

- Classical CLT cannot be applied to C_n.
- Assumptions: $C_n = O_p(1)$ and $\text{plim}\,\mathbf{Var}\left(C_n\right) = \text{plim}\,\boldsymbol{\Sigma}_n$ exists.
- We show that

$$\boldsymbol{\Sigma}_n = \mathbf{E}\left\{(1/n)\sum_{i=1}^{n}\mathbf{Var}\left(\nabla_{\boldsymbol{\theta}} q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)|\hat{\boldsymbol{\beta}}_n\right)\right\} + (1/n)\,\mathbf{Var}\left\{\sum_{i=1}^{n}\mathbf{E}\left(\nabla_{\boldsymbol{\theta}} q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)|\hat{\boldsymbol{\beta}}_n\right)\right\}$$

Variance Estimation

- Because the second stage uses an M-estimator approach and $Q_n(\boldsymbol{\theta}, \mathbf{y}_n, \mathbf{X}_n, \hat{\boldsymbol{\beta}}_n)$ is known, we can compute $(1/n)\sum_{i=1}^{n} \mathbf{Var}\left(\nabla_{\boldsymbol{\theta}} q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)|\hat{\boldsymbol{\beta}}_n\right) = \mathbf{H}_v(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)$ and $(1/\sqrt{n})\sum_{i=1}^{n} \mathbf{E}\left(\nabla_{\boldsymbol{\theta}} q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)|\hat{\boldsymbol{\beta}}_n\right) = \mathbf{H}_e(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)$.
- $\boldsymbol{\Sigma}_n$ can be consistently estimated by

$$\hat{\boldsymbol{\Sigma}}_n = \frac{1}{B}\sum_{b=1}^{B} \mathbf{H}_v(\boldsymbol{\theta}_0, \boldsymbol{\beta}^{(b)}) + \frac{1}{B-1}\sum_{b=1}^{B}(\mathbf{H}_e(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}^{(b)}) - \boldsymbol{\Omega})(\mathbf{H}_e(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}^{(b)}) - \boldsymbol{\Omega})',$$

  where $\boldsymbol{\Omega} = (1/B)\sum_{b=1}^{B} \mathbf{H}_e(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}^{(b)})$ and $\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(B)}$ are draws from the distribution of $\hat{\boldsymbol{\beta}}_n$.
- $\mathbf{Var}\left(\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\right)$ can be estimated by $\hat{\mathbf{A}}_n\hat{\boldsymbol{\Sigma}}_n\hat{\mathbf{A}}_n'$.

Variance Estimation

- Because the second stage uses an M-estimator approach and $Q_n(\boldsymbol{\theta}, \mathbf{y}_n, \mathbf{X}_n, \hat{\boldsymbol{\beta}}_n)$ is known, we can compute $(1/n)\sum_{i=1}^n \mathbf{Var}\left(\nabla_{\boldsymbol{\theta}} q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n) | \hat{\boldsymbol{\beta}}_n\right) = \mathbf{H}_v(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)$ and $(1/\sqrt{n})\sum_{i=1}^n \mathbf{E}\left(\nabla_{\boldsymbol{\theta}} q_{n,i}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n) | \hat{\boldsymbol{\beta}}_n\right) = \mathbf{H}_e(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}_n)$.
- $\boldsymbol{\Sigma}_n$ can be consistently estimated by

$$\hat{\boldsymbol{\Sigma}}_n = \frac{1}{B}\sum_{b=1}^{B} \mathbf{H}_v(\boldsymbol{\theta}_0, \boldsymbol{\beta}^{(b)}) + \frac{1}{B-1}\sum_{b=1}^{B}(\mathbf{H}_e(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}^{(b)}) - \boldsymbol{\Omega})(\mathbf{H}_e(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}^{(b)}) - \boldsymbol{\Omega})',$$

  where $\boldsymbol{\Omega} = (1/B)\sum_{b=1}^B \mathbf{H}_e(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}^{(b)})$ and $\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(B)}$ are draws from the distribution of $\hat{\boldsymbol{\beta}}_n$.
- $\mathbf{Var}\left(\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\right)$ can be estimated by $\hat{\mathbf{A}}_n \hat{\boldsymbol{\Sigma}}_n \hat{\mathbf{A}}_n'$.

Confidence Intervals using the Variance

- A confidence interval for θ_0 can be obtained regardless of the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$.
- E.g., assume that θ_0 is a scalar. We are looking for $\mathbf{a}_n$ such that $\mathcal{P}\left\{\theta_0 \in (\hat{\theta}_n \pm \mathbf{a}_n)\right\} \geq 1 - \alpha$, which implies $\mathcal{P}\left(|\hat{\theta}_n - \theta_0| > \mathbf{a}_n\right) \leq \alpha$.
- Chebyshev's inequality: $\mathbf{a}_n = \dfrac{\sigma\left(\sqrt{n}(\hat{\theta}_n - \theta_0)\right)}{(n\alpha)^{1/2}}$.
- A weaker test. In most cases, $\mathbf{a}_n$ is higher than for the case of normal distribution where $\mathbf{a}_n = \dfrac{\sigma\left(\sqrt{n}(\hat{\theta}_n - \theta_0)\right)}{n^{1/2}}\Phi\left(1 - \frac{\alpha}{2}\right)$.
- For $\alpha = 5\%$, H_0 is rejected if $\hat{\theta}_n/\sigma(\sqrt{n}(\hat{\theta}_n) > 4.47$ against 1.96 for the normal distribution.

Confidence Intervals using the Variance

- A confidence interval for θ_0 can be obtained regardless of the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$.
- E.g., assume that θ_0 is a scalar. We are looking for $\mathbf{a}_n$ such that $\mathcal{P}\left\{\theta_0 \in (\hat{\theta}_n \pm \mathbf{a}_n)\right\} \geq 1 - \alpha$, which implies $\mathcal{P}\left(|\hat{\theta}_n - \theta_0| > \mathbf{a}_n\right) \leq \alpha$.
- Chebyshev's inequality: $\mathbf{a}_n = \dfrac{\sigma\left(\sqrt{n}(\hat{\theta}_n - \theta_0)\right)}{(n\alpha)^{1/2}}$.
- A weaker test. In most cases, $\mathbf{a}_n$ is higher than for the case of normal distribution where $\mathbf{a}_n = \dfrac{\sigma\left(\sqrt{n}(\hat{\theta}_n - \theta_0)\right)}{n^{1/2}}\Phi\left(1 - \frac{\alpha}{2}\right)$.
- For $\alpha = 5\%$, H_0 is rejected if $\hat{\theta}_n/\sigma(\sqrt{n}(\hat{\theta}_n) > 4.47$ against 1.96 for the normal distribution.

Distribution Approximation

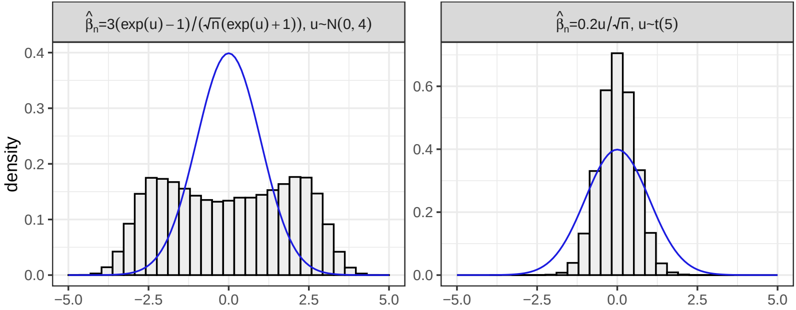

Figure: Distriution of the plug-in estimator

Distribution Approximation

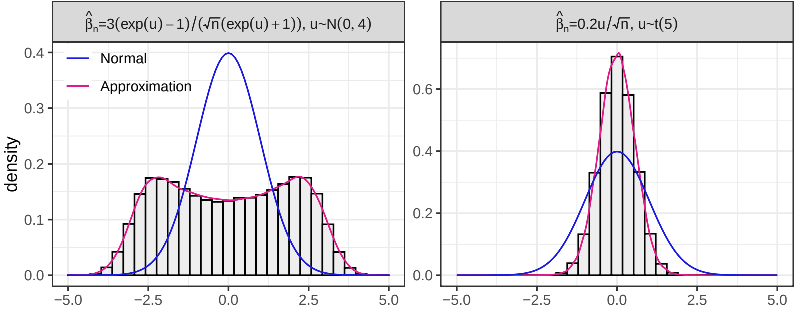

Figure: Distriution of the plug-in estimator and distribution approximation

Distribution Approximation

- Poisson model: $\lambda_i = \exp(\theta_{0,1} + \theta_{0,2} p_i)$, where p_i is not observed.

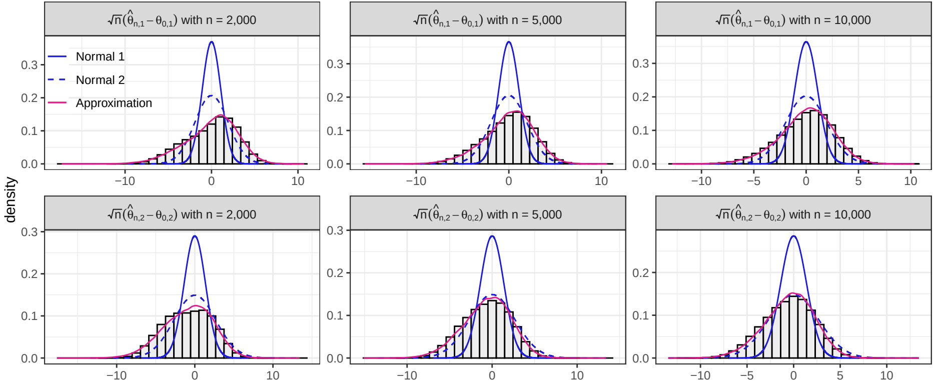

Figure: Distriution of the plug-in estimator and distribution approximation

Conclusion

- Asymptotic analysis of conditional M-estimators.
- First stage is general and includes non-Gaussian distributions (e.g., Bayesian estimators).
- Estimation of the variance at the second stage taking into account the uncertainty of the first stage.
- The asymptotic normality is not necessarily guaranteed at the second stage.
- Weak statistical tests at the second stage and distribution approximation.
- To do
  1. Application on real data.

Conclusion

- Asymptotic analysis of conditional M-estimators.
- First stage is general and includes non-Gaussian distributions (e.g., Bayesian estimators).
- Estimation of the variance at the second stage taking into account the uncertainty of the first stage.
- The asymptotic normality is not necessarily guaranteed at the second stage.
- Weak statistical tests at the second stage and distribution approximation.
- To do
  1. Application on real data.

THANK YOU