5 Extremum Estimation

The previous chapter presented several (frequentist) estimation approaches for which no closed form solution for the point estimator can be obtained, such as the probit model or Lasso regression. These estimators $\hat{\theta}_n$ are defined implicitly by the minimization (or maximization) of some objective function $Q_n(\theta, Y^n)$, e.g. the likelihood. This chapter discusses asymptotic results for such estimators. It starts with the baseline case where an estimator is uniquely point-identified and lies in the interior of the parameter space, before proceeding to non-standard asymptotics. The results presented in this chapter are important over and beyond assessing the asymptotic behavior of an estimator because the asymptotic distribution is standardly used to approximate the finite sample distribution of an estimator (and therefore to obtain standard errors and enable hypothesis testing and the construction of confidence sets). The numerical optimization methods necessary for obtaining the estimators themselves are outlined in Section 7.1.

5.1 Standard Asymptotics

Let

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\theta, Y^n) \ ,$$

where Y^n contains all data, e.g. Y and X in the linear regression. For example, for the probit model from Section 4.4, we have

$$Q_n(\beta, Y^n) = -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i, \beta) = -\frac{1}{n}\sum_{i=1}^{n} \{y_i \ln(\Phi(x_i'\beta) + (1-y_i)\ln(\Phi(-x_i'\beta))\} \ .$$

For the Lasso regression from Section 4.5, we have $Q_n(\beta, Y^n) = \frac{1}{n}\left[(Y - X\beta)'(Y - X\beta) + \lambda\|\beta\|\right]$.

Proposition 25 (Extremum Estimation: Consistency)**.**
Assume

1. $\Theta \subset \mathbb{R}$ *is compact;*
2. $Q_n(\theta, Y^n)$ *converges uniformly in probability to a limit objective function* $Q(\theta)$*, i.e.* $\forall\, \epsilon > 0$, $\mathbb{P}\left[\sup_{\theta\in\Theta} |Q_n(\theta, Y^n) - Q(\theta)| < \epsilon\right] \to 1$*;*
3. $Q(\theta)$ *is uniquely minimized by* θ_0*, i.e.* $Q(\theta) > Q(\theta_0) \quad \forall\, \theta \in \Theta,\, \theta \neq \theta_0$*; and*
4. $Q(\theta)$ *is continuous on* Θ*.*

Then

$$\hat{\theta}_n \xrightarrow{p} \theta_0 \ .$$

A heuristic proof is given in the Appendix. As an example, consider the OLS or ML estimation of the simple model from Section 2.1, where

$$Q_n(\theta, y^n) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - 2\theta\sum_{i=1}^{n} y_i + \theta^2 \ .$$

Let's check the conditions one by one. First, to make Θ compact, we consider only parameter values in $\Theta = [-c, c] \subset \mathbb{R}$ for some c large. Second, we can see that $Q_n(\theta, Y^n)$ converges uniformly in probability to $Q(\theta) = \mathbb{E}[y_i^2] - 2\theta\mathbb{E}[y_i] + \theta^2 = \mathbb{V}[y_i] + (\mathbb{E}[y_i] - \theta)^2$:

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} y_i^2 - 2\theta\frac{1}{n}\sum_{i=1}^{n} y_i + \theta^2 - (\mathbb{E}[y_i^2] - 2\theta\mathbb{E}[y_i] + \theta^2)\right| \\
\leq &\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \mathbb{E}[y_i^2]\right| + \sup_{\theta\in\Theta} 2\,|\theta|\left|\frac{1}{n}\sum_{i=1}^{n} y_i - \mathbb{E}[y_i]\right| \\
= &\left|\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \mathbb{E}[y_i^2]\right| + 2c\left|\frac{1}{n}\sum_{i=1}^{n} y_i - \mathbb{E}[y_i]\right| \\
\xrightarrow{p} &\ 0 \ ,
\end{aligned}$$

where the last line uses the WLLN. Third, it is easy to see that $\theta_0 = \arg\min_{\theta\in\Theta} Q(\theta) = \mathbb{E}[y_i]$ is unique. Fourth, $Q(\theta)$ is clearly a continuous function of θ.

Often, uniform convergence in probability cannot be shown as easily as for the objective function above. In many of these cases, we can instead make use of the fact that Q_n is of the form $Q_n(\theta, y^n) = \frac{1}{n}\sum_{i=1}^{n} m(x_i; \theta)$ for some $m(x_i; \theta)$ and rely on the Uniform Law of Large Numbers (ULLN). An exmple application of it is discussed further below.

Proposition 26 (Uniform Law of Large Numbers (ULLN))**.**
Assume

1. x_i *is i.i.d.;*
2. Θ *is compact;*
3. $m(x_i;\theta)$ *is continuous on* Θ*; and*
4. $\mathbb{E}\left[\sup_{\theta\in\Theta}||m(x_i;\theta)||\right] < \infty$

Then

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} m(x_i;\theta) - \mathbb{E}[m(x_i;\theta)]\right| \xrightarrow{p} 0 \ ,$$

i.e. $\frac{1}{n}\sum_{i=1}^{n} m(x_i;\theta)$ *converges uniformly in probability to* $\mathbb{E}[m(x_i;\theta)]$*.*

To illustrate Propositions 25 and 26, consider the nonlinear least squares (NLS) estimation of the regression model

$$y_i = (x_i'\beta)^3 + u_i \ , \quad \mathbb{E}[u_i|x_i] = 0 \ .$$

Define $\mathcal{B} = \{\beta \in \mathbb{R}^k : \|\beta\| \leq c\}$ for some $c > 0$ large and let

$$\hat{\beta} = \arg\min_{\beta\in\mathcal{B}} Q_n(\beta, Y^n) = \arg\min_{\beta\in\mathcal{B}} \frac{1}{2n}\sum_{i=1}^{n}(y_i - (x_i'\beta)^3)^2 \ .$$

To show $\hat{\beta} \xrightarrow{p} \beta_0$, we verify the conditions in Proposition 25, whereby we use Proposition 26 to show uniform convergence in probability of Q_n:

1. $\mathcal{B}$ is compact in $\mathbb{R}^k$.
2. $Q_n(\beta) = \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} m((x_i,y_i),\beta)$ converges uniformly in probability to $Q(\beta) = \frac{1}{2}\mathbb{E}[m(x;\theta)] = \mathbb{E}\left[(y_i - (x_i'\beta)^3)^2\right]$ because $m((x_i,y_i),\beta) = (y_i - (x_i'\beta)^3)^2$ satisfies the conditions for the ULLN; the first three are obvious, and for the fourth it is sufficient to assume $\mathbb{E}[||u_i||^2] < \infty$ and $\mathbb{E}[||x_i||^6] < \infty$, along with $\|\beta\| \leq c$:

$$\mathbb{E}\left[\sup_{\beta\in\mathcal{B}}||m((x_i,y_i),\beta)||\right] \leqslant \mathbb{E}\left[|y_i|^2\right] + \sup_{\beta\in\mathcal{B}} 2\mathbb{E}\left[|y_i||x_i|^3\|\beta\|^3\right] + \sup_{\beta\in\mathcal{B}}\mathbb{E}\left[|x_i|^6\|\beta\|^6\right] < \infty \ .$$

3. We know $\mathbb{E}[(y_i - h(x_i))^2]$ is uniquely minimized at $h(x_i) = \mathbb{E}[y_i|x_i] = (x_i'\beta_0)^3$. Thus, $Q_n(\beta) = \frac{1}{2}\mathbb{E}[(y_i - (x_i'\beta)^3)^2]$ is uniqely minimized at $\beta = \beta_0$.
4. $Q(\theta)$ is continuous.

In general, there are three ways to show that θ_0 is the unique minimizer of $Q(\theta)$. First, one can write out $Q(\theta)$ to see it explicitly by looking at FOCs (and SOCs) as in the first example above. Second, one can use the conditional-expectation-argument as in the second example above. Third, one can show that $Q(\tilde{\theta}) - Q(\theta_0) > 0 \quad \forall\, \tilde{\theta} \neq \theta_0$.

Proposition 27 (Extremum Estimation: Asymptotic Normality)**.**
In addition to the conditions in Proposition 25, assume

1. $\theta_0 \in int(\Theta)$*;*
2. $\sqrt{n}Q_n^{(1)}(\theta_0, Y^n) \xrightarrow{d} N(0, M)$*;*
3. $Q_n(\theta, Y^n)$ *is twice differentiable w.r.t.* $\theta \;\forall\, Y^n$*.*
   Also, $\exists H$*, p.d., s.t.* $Q_n^{(2)}(\theta_0, Y^n) \xrightarrow{p} H$*, and* $Q_n^{(2)}(\theta_n, Y^n) \xrightarrow{p} H \;\; \forall\, \theta_n \xrightarrow{p} \theta_0$*.*

Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, H^{-1}MH^{-1}) \; .$$

Again, a heuristic proof is given in the Appendix. Given consistent estimators $\hat{H}$ and $\hat{M}$, the asymptotic variance can be estimated as $\hat{H}^{-1}\hat{M}\hat{H}^{-1}$. The finite sample distribution of $\hat{\theta}$ is then often approximated as $\hat{\theta} \sim N(\theta_0, \frac{1}{n}\hat{H}^{-1}\hat{M}\hat{H}^{-1})$. Based on the result above, we know this approximation is good for large n, but how large n has to be depends on the application. In some cases, this approximation may not resemble at all the finite sample distribution.[1] Once asymptotic Normality is established, the Wald test becomes applicable.

Consider again the NLS example from above. To show $\sqrt{n}(\hat{\beta} - \beta) \sim N(0, H^{-1}MH^{-1})$ – with M and H defined below –, we verify the conditions in Proposition 27 in addition to the conditions in Proposition 25:

1. $\beta_0 \in int(\mathcal{B})$ for c large enough.
2. By CLT,

$$\begin{aligned}
\sqrt{n}Q_n^{(1)}(\beta_0, Y^n) &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(y_i - (x_i'\beta_0)^3)3(x_i'\beta_0)^2 x_i \\
&\xrightarrow{d} N(0, \mathbb{E}[9u_i^2(x_i'\beta_0)^4 x_i x_i']) \\
&\equiv N(0, M) \; .
\end{aligned}$$

[1] For example, this is the case under Instrumental Variable (IV) estimation with weak IVs (see Section 6.4).

3. $Q_n(\beta)$ is twice differentiable w.r.t. β. Also,

$$Q_n^{(2)}(\beta_0, Y^n) = \frac{1}{n}\sum_{i=1}^{n} -(y_i - (x_i'\beta_0)^3)6(x_i'\beta_0)x_i x_i' + 9(x_i'\beta_0)^4 x_i x_i' \xrightarrow{p} \mathbb{E}[9(x_i'\beta_0)^4 x_i x_i'] \equiv H \ .$$

In addition, the last part of this condition is satisfied: $Q_n^{(2)}(\beta_n, Y^n) \xrightarrow{p} H \ \forall \ \beta_n \xrightarrow{p} \beta_0$. This is because $Q_n^{(2)}(\beta_0, Y^n) = \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\beta\partial\beta'} m((x_i, y_i), \beta)$, and $\frac{\partial^2}{\partial\beta\partial\beta'} m((x_i, y_i), \beta)$ is a continuous function, which means that the plug-in property applies.

We can estimate the asymptotic variance in a heteroskedasticity-robust way as $\hat{H}^{-1}\hat{M}\hat{H}^{-1}$ with $\hat{M} = \frac{1}{n}\sum_{i=1}^{n} 9\hat{u}_i^2(x_i'\hat{\beta})^4 x_i x_i'$ and $\hat{H} = \frac{1}{n}\sum_{i=1}^{n} 9(x_i'\hat{\beta})^4 x_i x_i'$. Under homoskedasticity, $M = \sigma^2\mathbb{E}[9(x_i'\beta_0)^4 x_i x_i']$ and the asymptotic variance simplifies to $\sigma^2\hat{H}^{-1}$, which can be estimated using $\hat{H}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$.

Maximum Likelihood Estimation Under ML estimation, the above asymptotic variance simplifies. We have $Q_n(\theta, Y^n) = -\frac{1}{n}\ell(\theta|Y^n) = -\frac{1}{n}\sum_{i=1}^{n}\ln p(y_i|\theta)$. The information matrix inequality implies $M = H$. Therefore, $H^{-1}MH^{-1} = M^{-1} = H^{-1}$, and if the above conditions are satisfied for a particular applicaton, we get $\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, M^{-1})$.[2] More details are in the Appendix.

5.2 Non-Standard Asymptotics

5.2.1 Two-Step Estimation

Suppose we are interested in estimating

$$\hat{\theta} = \arg\min_{\theta\in\Theta} Q_n(\theta, \hat{\gamma}, Y^n) \ ,$$

where $\hat{\gamma}$ is a first-step estimate – often called a "nuisance parameter" – with properties $\hat{\gamma} \xrightarrow{p} \gamma^*$ and $\sqrt{n}(\hat{\gamma} - \gamma^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} r(y_i, \gamma^*) \xrightarrow{d} N(0, V)$. An example is the weighted nonlinear least squares estimator (WNLS)

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{2n}\sum_{i=1}^{n}[y_i - m(x_i, \theta)]^2/h(x_i, \hat{\gamma}) \ ,$$

[2] As the second-order derivatives required for H are tedious to compute for some more complicated models, typically M^{-1} is used to estimate the asymptotic variance. However, both expressions are correct.

different philosophies takn by the two approaches.

Let $\theta = \phi + \alpha$ for $\alpha \in [0, \lambda]$. We can rewrite the likelihood function (for ϕ) as $l(\phi, \hat{\phi}_n) = C - \frac{n}{2}(\phi - \hat{\phi}_n)^2$, where $\hat{\phi}_n = \frac{1}{n}\sum_{i=1}^n y_i$ is a sufficient statistic for the data Y^n. Replacing the parameter ϕ with the parameters θ and α, this yields $l(\theta, \alpha, \hat{\phi}_n) = C - \frac{n}{2}(\theta - \alpha - \hat{\phi}_n)^2$.

Under the frequentist paradigm, one would construct an objective function with θ as the only argument by "profiling" or "concentrating out" the objective function $l(\theta, \alpha, \hat{\phi}_n)$, i.e. taking the lowest possible α for any θ:

$$Q_n(\theta; \hat{\phi}_n) = \inf_{\alpha \in [0,\lambda]} \frac{n}{2}(\theta - \alpha - \hat{\phi}_n)^2 = \begin{cases} n(\hat{\phi}_n - \theta)^2 & \text{if } \theta \leqslant \hat{\phi}_n \\ 0 & \text{if } \hat{\phi}_n \leqslant \theta \leqslant \hat{\phi}_n + \lambda \\ (\hat{\phi}_n + \lambda - \theta)^2 & \text{if } \hat{\phi}_n + \lambda \leqslant \theta \end{cases} .$$

As a result, we can only say that $\hat{\theta} \in [\hat{\phi}, \hat{\phi} + \lambda]$, but we cannot provide a point-estimator for θ. In contrast, under Bayesian inference, we would integrate out α by specifying a prior distribution. As the likelihood does not depend on α, its posterior distribution is equal to its prior distribution. This prior together with the posterior for ϕ determine the posterior for θ, based on which we can compute the posterior mean, mode and any other statistic in the usual way. Thereby, the posterior for θ only differs from its prior to the extent that the posterior of ϕ differs from its prior, which moves the location of the identified set. Within the identified set (i.e. conditional on ϕ), the posterior of θ is proportional to its prior. In other words, for set-identified parameters, the data only tells us where the identified set lies, but not where θ lies within the identified set. This information can only come from the prior distribution. These insights do not change even if λ is estimated as long as $\hat{\lambda} \neq 0$. Also, all these insights hold in finite samples as well as asymptotically (except if $\hat{\lambda} \xrightarrow{p} 0$, in which case $\hat{\theta}$ is point-identified asymptotically).

Appendix

Heuristic proof of extremum estimation consistency Let $N_\delta(\theta_0) = \{\theta \in \Theta : |\theta - \theta_0| < \delta\}$ and let $N_\delta^c(\theta_0)$ denote its complement. WTS: $\mathbb{P}\{\hat{\theta} \in N_\delta(\theta_0)\} \to 1 \quad \forall\, \delta > 0$.

Partition the sample space; take the set of samples

$$F_n(\epsilon) = \{y^n : \sup_{\theta \in \Theta} |Q_n(\theta, Y^n) - Q(\theta)| < \frac{\epsilon}{2}\} , \quad \text{where } \epsilon = \min_{\theta \in N_\delta^c(\theta_0)} Q(\theta) - Q(\theta_0) ,$$

i.e. for every Y^n in $F_n(\epsilon)$, he know that $Q_n(\theta, Y^n)$ is always within $\frac{\epsilon}{2}$ distance from $Q(\theta)$,

$\forall\ \theta \in \Theta$.

For all $Y^n \in F_n(\epsilon)$, we have $Q_n(\theta, Y^n) > Q(\theta) - \frac{\epsilon}{2}$. Thus, $\min_{\theta \in N_\delta^c(\theta_0)} Q_n(\theta, Y^n) > \min_{\theta \in N_\delta^c(\theta_0)} Q(\theta) - \frac{\epsilon}{2} = Q(\theta_0) + \epsilon - \frac{\epsilon}{2} = Q(\theta_0) + \frac{\epsilon}{2}$ (by the definition of ϵ above).

Moreover, for all $Y^n \in F_n(\epsilon)$, we have $Q_n(\theta_0, Y^n) < Q(\theta_0) + \frac{\epsilon}{2}$. Because $\theta_0 \in N_\delta(\theta_0)$, we also have $\min_{\theta \in N_\delta(\theta_0)} Q_n(\theta, Y^n) \leqslant Q_n(\theta_0, Y^n)$. Putting these two pieces together, we have $\min_{\theta \in N_\delta(\theta_0)} Q_n(\theta, Y^n) < Q|\theta_0) + \frac{\epsilon}{2}$.

These two results above tell us that

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\theta, Y^n) \in N_\delta(\theta_0) \quad \forall\ Y^n \in F_n(\epsilon)\ .$$

This implies $\{Y^n : \hat{\theta}_n \in N_\delta(\theta_0)\} \supseteq F_n(\epsilon)$ and therefore, $\mathbb{P}\{\hat{\theta}_n \in N_\delta(\theta_0)\} \geqslant \mathbb{P}\{Y^n \in F_n(\epsilon)\}$. By assumption 2, $\mathbb{P}\{Y^n \in F_n(\epsilon)\} \to 1$, and so $\mathbb{P}\{\hat{\theta}_n \in N_\delta(\theta_0)\} \to 1$. The whole argument holds $\forall\ \delta > 0$.

Heuristic proof of extremum estimation asymptotic Normality Because of assumption 1, for large n, $\hat{\theta_n}$ is likely to be in int(Θ) and thus to satisfy the FOC $Q_n^{(1)}(\hat{\theta_n}, Y^n) = 0$.

By the mean value theorem $\exists\ \theta_n$ in-between $\hat{\theta}_n$ and θ_0 s.t. $Q_n^{(1)}(\hat{\theta_n}, y^n) = Q_n^{(1)}(\theta_0, Y^n) + Q_n^{(2)}(\theta_n, Y^n)(\hat{\theta}_n - \theta_0)$. Because this is equal to zero by FOC, we can rearrange it to yield

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -[Q_n^{(2)}(\theta_n, y^n)]^{-1}\sqrt{n}Q_n^{(1)}(\theta_0, y^n) \xrightarrow{d} N(0, H^{-1}MH^{-1})$$

by assumptions 2 and 3 and by Slutsky's theorem.

Asymptotic variance under ML estimation Under ML estimation, we have $Q_n(\theta, Y^n) = -\frac{1}{n}\ell(\theta|Y^n) = -\frac{1}{n}\sum_{i=1}^n \ln p(y_i|\theta)$. Assume all of the conditions for asymptotic Normality of $\hat{\theta}_{ML}$ are satisfied. Thanks to the information matrix equality, the asymptotic variance simplifies to $H^{-1}MH^{-1} = M^{-1} = H^{-1}$.

Let $s_i(\theta) = \frac{\partial}{\partial\theta}\ln p(y_i|\theta)$ and $H_i(\theta) = \frac{\partial^2}{\partial\theta\partial\theta'}\ln p(y_i|\theta)$. By CLT,

$$\sqrt{n}Q_n^{(1)}(\theta_0, Y^n) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n s_i(\theta_0) \xrightarrow{d} N(0, M)\ ,$$

with $M = \mathbb{E}\left[s_i(\theta_0)s_i(\theta_0)'\right]$. Also, note that we have $\mathbb{E}[s_i(\theta_0)] = 0$ provided that $\theta_0 \in int(\Theta)$.

This is because $\theta_0 = \arg\max_\theta \ Q(\theta) = \arg\max_\theta \ -\frac{1}{n}\mathbb{E}[\ln p(y_i|\theta)]$. By WLLN,

$$Q_n^{(2)}(\theta_0, Y^n) = -\frac{1}{n}\sum_{i=1}^{n} H_i(\theta_0) \to H \ ,$$

with $H = -\mathbb{E}[H_i(\theta_0)]$. By the information matrix inequality, $M = H$ (see Section 4.2 and Appendix to Chapter 4), and hence $\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, M^{-1}) = N(0, H^{-1})$.