Chapter 36

LARGE SAMPLE ESTIMATION AND HYPOTHESIS TESTING*

WHITNEY K. NEWEY
Massachusetts Institute of Technology

DANIEL McFADDEN
University of California, Berkeley

Contents

*We are grateful to the NSF for financial support and to Y. Ait-Sahalia, J. Porter, J. Powell, J. Robins, P. Ruud, and T. Stoker for helpful comments.

Handbook of Econometrics, Volume IV, Edited by R.F. Engle and D.L. McFadden
© 1994 Elsevier Science B.V. All rights reserved

Abstract

Asymptotic distribution theory is the primary method used to examine the properties of econometric estimators and tests. We present conditions for obtaining consistency and asymptotic normality of a very general class of estimators (extremum estimators). Consistent asymptotic variance estimators are given to enable approximation of the asymptotic distribution. Asymptotic efficiency is another desirable property then considered. Throughout the chapter, the general results are also specialized to common econometric estimators (e.g. MLE and GMM), and in specific examples we work through the conditions for the various results in detail. The results are also extended to two-step estimators (with finite-dimensional parameter estimation in the first step), estimators derived from nonsmooth objective functions, and semiparametric two-step estimators (with nonparametric estimation of an infinite-dimensional parameter in the first step). Finally, the trinity of test statistics is considered within the quite general setting of GMM estimation, and numerous examples are given.

1. Introduction

Large sample distribution theory is the cornerstone of statistical inference for econometric models. The limiting distribution of a statistic gives approximate distributional results that are often straightforward to derive, even in complicated econometric models. These distributions are useful for approximate inference, including constructing approximate confidence intervals and test statistics. Also, the location and dispersion of the limiting distribution provides criteria for choosing between different estimators. Of course, asymptotic results are sensitive to the accuracy of the large sample approximation, but the approximation has been found to be quite good in many cases and asymptotic distribution results are an important starting point for further improvements, such as the bootstrap. Also, exact distribution theory is often difficult to derive in econometric models, and may not apply to models with unspecified distributions, which are important in econometrics. Because asymptotic theory is so useful for econometric models, it is important to have general results with conditions that can be interpreted and applied to particular estimators as easily as possible. The purpose of this chapter is the presentation of such results.

Consistency and asymptotic normality are the two fundamental large sample properties of estimators considered in this chapter. A *consistent* estimator $\hat{\theta}$ is one that converges in probability to the true value θ_0, i.e. $\hat{\theta} \xrightarrow{p} \theta_0$, as the sample size n goes to infinity, for all possible true values.[1] This is a mild property, only requiring

[1] This property is sometimes referred to as weak consistency, with strong consistency holding when $\hat{\theta}$ converges almost surely to the true value. Throughout the chapter we focus on weak consistency, although we also show how strong consistency can be proven.

that the estimator is close to the truth when the number of observations is nearly infinite. Thus, an estimator that is not even consistent is usually considered inadequate. Also, consistency is useful because it means that the asymptotic distribution of an estimator is determined by its limiting behavior near the true parameter.

An *asymptotically normal* estimator $\hat{\theta}$ is one where there is an increasing function $v(n)$ such that the distribution function of $v(n)(\hat{\theta} - \theta_0)$ converges to the Gaussian distribution function with mean zero and variance V, i.e. $v(n)(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$. The variance V of the limiting distribution is referred to as the asymptotic variance of $\hat{\theta}$. The estimator is *$\sqrt{n}$-consistent* if $v(n) = \sqrt{n}$. This chapter focuses on the $\sqrt{n}$-consistent case, so that unless otherwise noted, asymptotic normality will be taken to include $\sqrt{n}$-consistency.

Asymptotic normality and a consistent estimator of the asymptotic variance can be used to construct approximate confidence intervals. In particular, for an estimator $\hat{V}$ of V and for $g_{\alpha/2}$ satisfying $\text{Prob}[N(0,1) > g_{\alpha/2}] = \alpha/2$, an asymptotic $1 - \alpha$ confidence interval is

$$\mathscr{I}_{1-\alpha} = [\hat{\theta} - g_{\alpha/2}(\hat{V}/n)^{1/2}, \hat{\theta} + g_{\alpha/2}(\hat{V}/n)^{1/2}].$$

If $\hat{V}$ is a consistent estimator of V and $V > 0$, then asymptotic normality of $\hat{\theta}$ will imply that $\text{Prob}(\theta_0 \in \mathscr{I}_{1-\alpha}) \to 1 - \alpha$ as $n \to \infty$.[2] Here asymptotic theory is important for econometric practice, where consistent standard errors can be used for approximate confidence interval construction. Thus, it is useful to know that estimators are asymptotically normal and to know how to form consistent standard errors in applications. In addition, the magnitude of asymptotic variances for different estimators helps choose between estimators in practice. If one estimator has a smaller asymptotic variance, then an asymptotic confidence interval, as above, will be shorter for that estimator in large samples, suggesting preference for its use in applications. A prime example is generalized least squares with estimated disturbance variance matrix, which has smaller asymptotic variance than ordinary least squares, and is often used in practice.

Many estimators share a common structure that is useful in showing consistency and asymptotic normality, and in deriving the asymptotic variance. The benefit of using this structure is that it distills the asymptotic theory to a few essential ingredients. The cost is that applying general results to particular estimators often requires thought and calculation. In our opinion, the benefits outweigh the costs, and so in these notes we focus on general structures, illustrating their application with examples.

One general structure, or framework, is the class of estimators that maximize some objective function that depends on data and sample size, referred to as *extremum* estimators. An estimator $\hat{\theta}$ is an extremum estimator if there is an

[2] The proof of this result is an exercise in convergence in distribution and the Slutzky theorem, which states that $Y_n \xrightarrow{d} Y_0$ and $Z_n \xrightarrow{p} c$ implies $Z_n Y_n \xrightarrow{d} c Y_0$.

objective function $\hat{Q}_n(\theta)$ such that

$$\hat{\theta} \text{ maximizes } \hat{Q}_n(\theta) \text{ subject to } \theta \in \Theta, \tag{1.1}$$

where Θ is the set of possible parameter values. In the notation, dependence of $\hat{\theta}$ on n and of $\hat{\theta}$ and $\hat{Q}_n(\theta)$ on the data is suppressed for convenience. This estimator is the maximizer of some objective function that depends on the data, hence the term "extremum estimator".[3] R.A. Fisher (1921, 1925), Wald (1949), Huber (1967), Jennrich (1969), and Malinvaud (1970) developed consistency and asymptotic normality results for various special cases of extremum estimators, and Amemiya (1973, 1985) formulated the general class of estimators and gave some useful results.

A prime example of an extremum estimator is the maximum likelihood (MLE). Let the data $(z_1, \ldots, z_n)$ be i.i.d. with p.d.f. $f(z|\theta_0)$ equal to some member of a family of p.d.f.'s $f(z|\theta)$. Throughout, we will take the p.d.f. $f(z|\theta)$ to mean a probability function where z is discrete, and to possibly be conditioned on part of the observation z.[4] The MLE satisfies eq. (1.1) with

$$\hat{Q}_n(\theta) = n^{-1} \sum_{i=1}^{n} \ln f(z_i|\theta). \tag{1.2}$$

Here $\hat{Q}_n(\theta)$ is the normalized log-likelihood. Of course, the monotonic transformation of taking the log of the likelihood and normalizing by n will not typically affect the estimator, but it is a convenient normalization in the theory. Asymptotic theory for the MLE was outlined by R.A. Fisher (1921, 1925), and Wald's (1949) consistency theorem is the prototype result for extremum estimators. Also, Huber (1967) gave weak conditions for consistency and asymptotic normality of the MLE and other extremum estimators that maximize a sample average.[5]

A second example is the nonlinear least squares (NLS), where for data $z_i = (y_i, x_i)$ with $E[y|x] = h(x, \theta_0)$, the estimator solves eq. (1.1) with

$$\hat{Q}_n(\theta) = -n^{-1} \sum_{i=1}^{n} [y_i - h(x_i, \theta)]^2. \tag{1.3}$$

Here maximizing $\hat{Q}_n(\theta)$ is the same as minimizing the sum of squared residuals. The asymptotic normality theorem of Jennrich (1969) is the prototype for many modern results on asymptotic normality of extremum estimators.

[3] "Extremum" rather than "maximum" appears here because minimizers are also special cases, with objective function equal to the negative of the minimand.

[4] More precisely, $f(z|\theta)$ is the density (Radon–Nikodym derivative) of the probability measure for z with respect to some measure that may assign measure 1 to some singleton's, allowing for discrete variables, and for $z = (y, x)$ may be the product of some measure for y with the marginal distribution of x, allowing $f(z|\theta)$ to be a conditional density given x.

[5] Estimators that maximize a sample average, i.e. where $\hat{Q}_n(\theta) = n^{-1} \sum_{i=1}^{n} q(z_i, \theta)$, are often referred to as *m-estimators*, where the "*m*" means "maximum-likelihood-like".

A third example is the generalized method of moments (GMM). Suppose that there is a "moment function" vector $g(z, \theta)$ such that the population moments satisfy $E[g(z, \theta_0)] = 0$. A GMM estimator is one that minimizes a squared Euclidean distance of sample moments from their population counterpart of zero. Let $\hat{W}$ be a positive semi-definite matrix, so that $(m'\hat{W}m)^{1/2}$ is a measure of the distance of m from zero. A GMM estimator is one that solves eq. (1.1) with

$$\hat{Q}_n(\theta) = -\left[n^{-1} \sum_{i=1}^{n} g(z_i, \theta) \right]' \hat{W} \left[n^{-1} \sum_{i=1}^{n} g(z_i, \theta) \right]. \tag{1.4}$$

This class includes linear instrumental variables estimators, where $g(z, \theta) = x \cdot (y - Y'\theta)$, x is a vector of instrumental variables, y is a left-hand-side dependent variable, and Y are right-hand-side variables. In this case the population moment condition $E[g(z, \theta_0)] = 0$ is the same as the product of instrumental variables x and the disturbance $y - Y'\theta_0$ having mean zero. By varying $\hat{W}$ one can construct a variety of instrumental variables estimators, including two-stage least squares for $\hat{W} = (n^{-1}\sum_{i=1}^{n} x_i x_i')^{-1}$.[6] The GMM class also includes nonlinear instrumental variables estimators, where $g(z, \theta) = x \cdot \rho(z, \theta)$ for a residual $\rho(z, \theta)$, satisfying $E[x \cdot \rho(z, \theta_0)] = 0$. Nonlinear instrumental variable estimators were developed and analyzed by Sargan (1959) and Amemiya (1974). Also, the GMM class was formulated and general results on asymptotic properties given in Burguete et al. (1982) and Hansen (1982).

The GMM class is general enough to also include MLE and NLS when those estimators are viewed as solutions to their first-order conditions. In this case the derivatives of $\ln f(z|\theta)$ or $-[y - h(x, \theta)]^2$ become the moment functions, and there are exactly as many moment functions as parameters. Thinking of GMM as including MLE, NLS, and many other estimators is quite useful for analyzing their asymptotic distribution, but not for showing consistency, as further discussed below.

A fourth example is classical minimum distance estimation (CMD). Suppose that there is a vector of estimators $\hat{\pi} \xrightarrow{p} \pi_0$ and a vector of functions $h(\theta)$ with $\pi_0 = h(\theta_0)$. The idea is that π consists of "reduced form" parameters, θ consists of "structural" parameters, and $h(\theta)$ gives the mapping from structure to reduced form. An estimator of θ can be constructed by solving eq. (1.1) with

$$\hat{Q}_n(\theta) = -[\hat{\pi} - h(\theta)]'\hat{W}[\hat{\pi} - h(\theta)], \tag{1.5}$$

where $\hat{W}$ is a positive semi-definite matrix. This class of estimators includes classical minimum chi-square methods for discrete data, as well as estimators for simultaneous equations models in Rothenberg (1973) and panel data in Chamberlain (1982). Its asymptotic properties were developed by Chiang (1956) and Ferguson (1958).

A different framework that is sometimes useful is minimum distance estimation,

[6] The $1/n$ normalization in $\hat{W}$ does not affect the estimator, but, by the law of large numbers, will imply that $\hat{W}$ converges in probability to a constant matrix, a condition imposed below.

a class of estimators that solve eq. (1.1) for $\hat{Q}_n(\theta) = -\hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta)$, where $\hat{g}_n(\theta)$ is a vector of the data and parameters such that $\hat{g}_n(\theta_0) \xrightarrow{p} 0$ and $\hat{W}$ is positive semi-definite. Both GMM and CMD are special cases of minimum distance, with $\hat{g}_n(\theta) = n^{-1}\sum_{i=1}^n g(z_i, \theta)$ for GMM and $\hat{g}_n(\theta) = \hat{\pi} - h(\theta)$ for CMD.[7] This framework is useful for analyzing asymptotic normality of GMM and CMD, because (once) differentiability of $\hat{g}_n(\theta)$ is a sufficient smoothness condition, while twice differentiability is often assumed for the objective function of an extremum estimator [see, e.g. Amemiya (1985)]. Indeed, as discussed in Section 3, asymptotic normality of an extremum estimator with a twice differentiable objective function $\hat{Q}_n(\theta)$ is actually a special case of asymptotic normality of a minimum distance estimator, with $\hat{g}_n(\theta) = \nabla_\theta \hat{Q}_n(\theta)$ and $\hat{W}$ equal to an identity matrix, where ∇_θ denotes the partial derivative. The idea here is that when analyzing asymptotic normality, an extremum estimator can be viewed as a solution to the first-order conditions $\nabla_\theta \hat{Q}_n(\hat{\theta}) = 0$, and in this form is a minimum distance estimator.

For consistency, it can be a bad idea to treat an extremum estimator as a solution to first-order conditions rather than a global maximum of an objective function, because the first-order condition can have multiple roots even when the objective function has a unique maximum. Thus, the first-order conditions may not identify the parameters, even when there is a unique maximum to the objective function. Also, it is often easier to specify primitive conditions for a unique maximum than for a unique root of the first-order conditions. A classic example is the MLE for the Cauchy location–scale model, where z is a scalar, μ is a location parameter, σ a scale parameter, and $f(z|\theta) = C\sigma^{-1}(1 + [(z-\mu)/\sigma]^2)^{-1}$ for a constant C. It is well known that, even in large samples, there are many roots to the first-order conditions for the location parameter μ, although there is a global maximum to the likelihood function; see Example 1 below. Econometric examples tend to be somewhat less extreme, but can still have multiple roots. An example is the censored least absolute deviations estimator of Powell (1984). This estimator solves eq. (1.1) for $\hat{Q}_n(\theta) = -n^{-1}\sum_{i=1}^n |y_i - \max\{0, x_i'\theta\}|$, where $y_i = \max\{0, x_i'\theta_0 + \varepsilon_i\}$, and ε_i has conditional median zero. A global maximum of this function over any compact set containing the true parameter will be consistent, under certain conditions, but the gradient has extraneous roots at any point where $x_i'\theta < 0$ for all i (e.g. which can occur if x_i is bounded).

The importance for consistency of an extremum estimator being a global maximum has practical implications. Many iterative maximization procedures (e.g. Newton–Raphson) may converge only to a local maximum, but consistency results only apply to the global maximum. Thus, it is often important to search for a global maximum. One approach to this problem is to try different starting values for iterative procedures, and pick the estimator that maximizes the objective from among the converged values. As long as the extremum estimator is consistent and the true parameter is an element of the interior of the parameter set Θ, an extremum estimator will be

[7] For GMM, the law of large numbers implies $\hat{g}_n(\theta_0) \xrightarrow{p} 0$.

a root of the first-order conditions asymptotically, and hence will be included among the local maxima. Also, this procedure can avoid extraneous boundary maxima, e.g. those that can occur in maximum likelihood estimation of mixture models.

Figure 1 shows a schematic, illustrating the relationships between the various types of estimators introduced so far. The name or mnemonic for each type of estimator (e.g. MLE for maximum likelihood) is given, along with objective function being maximized, except for GMM and CMD where the form of $\hat{g}_n(\theta)$ is given. The solid arrows indicate inclusion in a class of estimators. For example, MLE is included in the class of extremum estimators and GMM is a minimum distance estimator. The broken arrows indicate inclusion in the class when the estimator is viewed as a solution to first-order conditions. In particular, the first-order conditions for an extremum estimator are $\nabla_\theta \hat{Q}_n(\hat{\theta}) = 0$, making it a minimum distance estimator with $\hat{g}_n(\theta) = \nabla_\theta \hat{Q}_n(\theta)$ and $\hat{W} = I$. Similarly, the first-order conditions for MLE make it a GMM estimator with $g(z, \theta) = \nabla_\theta \ln f(z|\theta)$ and those for NLS a GMM estimator with $g(z, \theta) = -2[y - h(x, \theta)]\nabla_\theta h(x, \theta)$. As discussed above, these broken arrows are useful for analyzing the asymptotic distribution, but not for consistency. Also, as further discussed in Section 7, the broken arrows are not very useful when the objective function $\hat{Q}_n(\theta)$ is not smooth.

The broad outline of the chapter is to treat consistency, asymptotic normality, consistent asymptotic variance estimation, and asymptotic efficiency in that order. The general results will be organized hierarchically across sections, with the asymptotic normality results assuming consistency and the asymptotic efficiency results assuming asymptotic normality. In each section, some illustrative, self-contained examples will be given. Two-step estimators will be discussed in a separate section, partly as an illustration of how the general frameworks discussed here can be applied and partly because of their intrinsic importance in econometric applications. Two later sections deal with more advanced topics. Section 7 considers asymptotic normality when the objective function $\hat{Q}_n(\theta)$ is not smooth. Section 8 develops some asymptotic theory when $\hat{\theta}$ depends on a nonparametric estimator (e.g. a kernel regression, see Chapter 39).

This chapter is designed to provide an introduction to asymptotic theory for nonlinear models, as well as a guide to recent developments. For this purpose,

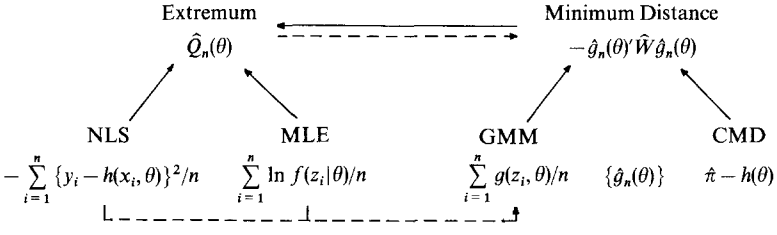

Figure 1.

Sections 2–6 have been organized in such a way that the more basic material is collected in the first part of each section. In particular, Sections 2.1–2.5, 3.1–3.4, 4.1–4.3, 5.1, and 5.2, might be used as text for part of a second-year graduate econometrics course, possibly also including some examples from the other parts of this chapter.

The results for extremum and minimum distance estimators are general enough to cover data that is a stationary stochastic process, but the regularity conditions for GMM, MLE, and the more specific examples are restricted to i.i.d. data. Modeling data as i.i.d. is satisfactory in many cross-section and panel data applications. Chapter 37 gives results for dependent observations.

This chapter assumes some familiarity with elementary concepts from analysis (e.g. compact sets, continuous functions, etc.) and with probability theory. More detailed familiarity with convergence concepts, laws of large numbers, and central limit theorems is assumed, e.g. as in Chapter 3 of Amemiya (1985), although some particularly important or potentially unfamiliar results will be cited in footnotes. The most technical explanations, including measurability concerns, will be reserved to footnotes.

Three basic examples will be used to illustrate the general results of this chapter.

Example 1.1 (Cauchy location–scale)

In this example z is a scalar random variable, $\theta = (\mu, \sigma)'$ is a two-dimensional vector, and z is continuously distributed with p.d.f. $f(z|\theta_0)$, where $f(z|\theta) = C \cdot \sigma^{-1}\{1 + [(z-\mu)/\sigma]^2\}^{-1}$ and C is a constant. In this example μ is a location parameter and σ a scale parameter. This example is interesting because the MLE will be consistent, in spite of the first-order conditions having many roots and the nonexistence of moments of z (e.g. so the sample mean is not a consistent estimator of θ_0).

Example 1.2 (Probit)

Probit is an MLE example where $z = (y, x')$ for a binary variable y, $y \in \{0, 1\}$, and a $q \times 1$ vector of regressors x, and the conditional probability of y given x is $f(z|\theta_0)$ for $f(z|\theta) = \Phi(x'\theta)^y[1 - \Phi(x'\theta)]^{1-y}$. Here $f(z|\theta_0)$ is a p.d.f. with respect to integration that sums over the two different values of y and integrates over the distribution of x, i.e. where the integral of any function $a(y, x)$ is $\int a(y, x)\,dz = E[a(1, x)] + E[a(0, x)]$. This example illustrates how regressors can be allowed for, and is a model that is often applied.

Example 1.3 (Hansen–Singleton)

This is a GMM (nonlinear instrumental variables) example, where $g(z, \theta) = x \cdot \rho(z, \theta)$ for $\rho(z, \theta) = \beta \cdot w \cdot y^\gamma - 1$. The functional form here is from Hansen and Singleton (1982), where β is a rate of time preference, γ a risk aversion parameter, w an asset return, y a consumption ratio for adjacent time periods, and x consists of variables

in the information set, of an agent maximizing expected constant relative risk aversion utility. This example is interesting because it illustrates the difficulty of specifying primitive identification conditions for GMM and the type of moment existence assumptions that are often useful.

2. Consistency

To motivate the precise conditions for consistency it is helpful to sketch the ideas on which the result is based. The basic idea is that if $\hat{Q}_n(\theta)$ converges in probability to $Q_0(\theta)$ for every θ, and $Q_0(\theta)$ is maximized at the true parameter θ_0, then the limit of the maximum $\hat{\theta}$ should be the maximum θ_0 of the limit, under conditions for interchanging the maximization and limiting operations. For example, consider the MLE. The law of large numbers suggests $\hat{Q}_n(\theta) \xrightarrow{p} Q_0(\theta) = E[\ln f(z|\theta)]$. By the well known information inequality, $Q_0(\theta)$ has a unique maximum at the true parameter when θ_0 is identified, as further discussed below. Then under technical conditions for the limit of the maximum to be the maximum of the limit, $\hat{\theta}$ should converge in probability to θ_0. Sufficient conditions for the maximum of the limit to be the limit of the maximum are that the convergence in probability is uniform and that the parameter set is compact.[8]

These ideas are illustrated in Figure 2. Let ε be a small positive number. If $\hat{Q}_n(\theta)$ lies in the "sleeve" $[Q_0(\theta) - \varepsilon, Q_0(\theta) + \varepsilon]$, for all θ, then $\hat{\theta}$ must lie in $[\theta_l, \theta_u]$, i.e. must be "close" to the value θ_0 that maximizes $Q_0(\theta)$. The estimator should then be consistent as long as θ_0 is the true parameter value.

It is essential for consistency that the limit $Q_0(\theta)$ have a unique maximum at the true parameter value. If there are multiple maxima, then this argument will only

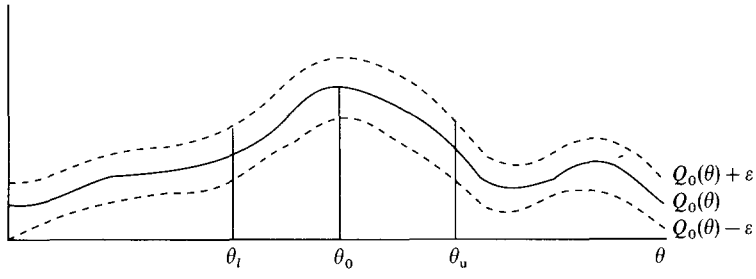

Figure 2.

[8] These ideas are also related to the result that the probability limit of a continuous function is the function of the probability limit. The maximum is a continuous function of $\{Q(\theta)\}$ where the maximum is unique, in the metric of uniform convergence on a compact set. Thus, if the probability limit, in this metric, of $\hat{Q}(\theta)$ is $Q(\theta)$, and the maximum of $Q(\theta)$ is unique, then the probability limit of $\hat{\theta}$ is the maximum of the limit $Q(\theta)$.

lead to the estimator being close to one of the maxima, which does not give consistency (because one of the maxima will not be the true value of the parameter). The condition that $Q_0(\theta)$ have a unique maximum at the true parameter is related to identification.

The discussion so far only allows for a compact parameter set. In theory compactness requires that one know bounds on the true parameter value, although this constraint is often ignored in practice. It is possible to drop this assumption if the function $\hat{Q}_n(\theta)$ cannot rise "too much" as θ becomes unbounded, as further discussed below.

Uniform convergence and continuity of the limiting function are also important. Uniform convergence corresponds to the feature of the graph that $\hat{Q}_n(\theta)$ was in the "sleeve" for *all* values of $\theta \in \Theta$. Conditions for uniform convergence are given below.

The rest of this section develops this descriptive discussion into precise results on consistency of extremum estimators. Section 2.1 presents the basic consistency theorem. Sections 2.2–2.5 give simple but general sufficient conditions for consistency, including results for MLE and GMM. More advanced and/or technical material is contained in Sections 2.6–2.8.

2.1. *The basic consistency theorem*

To state a theorem it is necessary to define precisely uniform convergence in probability, as follows:

Uniform convergence in probability: $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$ means $\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$.

The following is the fundamental consistency result for extremum estimators, and is similar to Lemma 3 of Amemiya (1973).

Theorem 2.1

If there is a function $Q_0(\theta)$ such that (i) $Q_0(\theta)$ is uniquely maximized at θ_0; (ii) Θ is compact; (iii) $Q_0(\theta)$ is continuous; (iv) $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$, then $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof

For any $\varepsilon > 0$ we have with probability approaching one (w.p.a.1) (a) $\hat{Q}_n(\hat{\theta}) > \hat{Q}_n(\theta_0) - \varepsilon/3$ by eq. (1.1); (b) $Q_0(\hat{\theta}) > \hat{Q}_n(\hat{\theta}) - \varepsilon/3$ by (iv); (c) $\hat{Q}_n(\theta_0) > Q_0(\theta_0) - \varepsilon/3$ by (iv).[9]

[9] The probability statements in this proof are only well defined if each of $\hat{\theta}, \hat{Q}_n(\hat{\theta})$, and $\hat{Q}_n(\theta_0)$ are measurable. The measurability issue can be bypassed by defining consistency and uniform convergence in terms of outer measure. The outer measure of a (possibly nonmeasurable) event $\mathscr{E}$ is the infimum of $E[Y]$ over all random variables Y with $Y \geqslant 1(\mathscr{E})$, where $1(\mathscr{E})$ is the indicator function for the event $\mathscr{E}$.

Therefore, w.p.a.1,

$$Q_0(\hat{\theta}) \overset{\text{(b)}}{>} \hat{Q}_n(\hat{\theta}) - \varepsilon/3 \overset{\text{(a)}}{>} \hat{Q}_n(\theta_0) - 2\varepsilon/3 \overset{\text{(c)}}{>} Q_0(\theta_0) - \varepsilon.$$

Thus, for any $\varepsilon > 0$, $Q_0(\hat{\theta}) > Q_0(\theta_0) - \varepsilon$ w.p.a.1. Let $\mathcal{N}$ be any open subset of Θ containing θ_0. By $\Theta \cap \mathcal{N}^c$ compact, (i), and (iii), $\sup_{\theta \in \Theta \cap \mathcal{N}^c} Q_0(\theta) = Q_0(\theta^*) < Q_0(\theta_0)$ for some $\theta^* \in \Theta \cap \mathcal{N}^c$. Thus, choosing $\varepsilon = Q_0(\theta_0) - \sup_{\theta \in \Theta \cap \mathcal{N}^c} Q_0(\theta)$, it follows that w.p.a.1 $Q_0(\hat{\theta}) > \sup_{\theta \in \Theta \cap \mathcal{N}^c} Q_0(\theta)$, and hence $\hat{\theta} \in \mathcal{N}$. Q.E.D.

The conditions of this theorem are slightly stronger than necessary. It is not necessary to assume that $\hat{\theta}$ actually maximizes the objective function. This assumption can be replaced by the hypothesis that $\hat{Q}_n(\hat{\theta}) \geqslant \sup_{\theta \in \Theta} \hat{Q}_n(\theta) + o_p(1)$. This replacement has no effect on the proof, in particular on part (a), so that the conclusion remains true. These modifications are useful for analyzing some estimators in econometrics, such as the maximum score estimator of Manski (1975) and the simulated moment estimators of Pakes (1986) and McFadden (1989). These modifications are not given in the statement of the consistency result in order to keep that result simple, but will be used later.

Some of the other conditions can also be weakened. Assumption (iii) can be changed to upper semi-continuity of $Q_0(\theta)$ and (iv) to $\hat{Q}_n(\theta_0) \xrightarrow{p} Q_0(\theta_0)$ and for all $\varepsilon > 0$, $\hat{Q}_n(\theta) < Q_0(\theta) + \varepsilon$ for all $\theta \in \Theta$ with probability approaching one.[10] Under these weaker conditions the conclusion still is satisfied, with exactly the same proof.

Theorem 2.1 is a weak consistency result, i.e. it shows $\hat{\theta} \xrightarrow{p} \theta_0$. A corresponding strong consistency result, i.e. $\hat{\theta} \xrightarrow{\text{a.s.}} \theta_0$, can be obtained by assuming that $\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| \xrightarrow{\text{a.s.}} 0$ holds in place of uniform convergence in probability. The proof is exactly the same as that above, except that "a.s. for large enough n" replaces "with probability approaching one". This and other results are stated here for convergence in probability because it suffices for the asymptotic distribution theory.

This result is quite general, applying to any topological space. Hence, it allows for θ to be infinite-dimensional, i.e. for θ to be a function, as would be of interest for nonparametric estimation of (say) a density or regression function. However, the compactness of the parameter space is difficult to check or implausible in many cases where θ is infinite-dimensional.

To use this result to show consistency of a particular estimator it must be possible to check the conditions. For this purpose it is important to have primitive conditions, where the word "primitive" here is used synonymously with the phrase "easy to interpret". The compactness condition is primitive but the others are not, so that it is important to discuss more primitive conditions, as will be done in the following subsections.

[10] Upper semi-continuity means that for any $\theta \in \Theta$ and $\varepsilon > 0$ there is an open subset $\mathcal{N}$ of Θ containing θ such that $Q_0(\theta') < Q_0(\theta) + \varepsilon$ for all $\theta' \in \mathcal{N}$.

Condition (i) is the identification condition discussed above, (ii) the boundedness condition on the parameter set, and (iii) and (iv) the continuity and uniform convergence conditions. These can be loosely grouped into "substantive" and "regularity" conditions. The identification condition (i) is substantive. There are well known examples where this condition fails, e.g. linear instrumental variables estimation with fewer instruments than parameters. Thus, it is particularly important to be able to specify primitive hypotheses for $Q_0(\theta)$ to have a unique maximum. The compactness condition (ii) is also substantive, with $\theta_0 \in \Theta$ requiring that bounds on the parameters be known. However, in applications the compactness restriction is often ignored. This practice is justified for estimators where compactness can be dropped without affecting consistency of estimators. Some of these estimators are discussed in Section 2.6.

Uniform convergence and continuity are the hypotheses that are often referred to as "the standard regularity conditions" for consistency. They will typically be satisfied when moments of certain functions exist and there is some continuity in $\hat{Q}_n(\theta)$ or in the distribution of the data. Moment existence assumptions are needed to use the law of large numbers to show convergence of $\hat{Q}_n(\theta)$ to its limit $Q_0(\theta)$. Continuity of the limit $Q_0(\theta)$ is quite a weak condition. It can even be true when $\hat{Q}_n(\theta)$ is not continuous, because continuity of the distribution of the data can "smooth out" the discontinuities in the sample objective function. Primitive regularity conditions for uniform convergence and continuity are given in Section 2.3. Also, Section 2.7 relates uniform convergence to stochastic equicontinuity, a property that is necessary and sufficient for uniform convergence, and gives more sufficient conditions for uniform convergence.

To formulate primitive conditions for consistency of an extremum estimator, it is necessary to first find $Q_0(\theta)$. Usually it is straightforward to calculate $Q_0(\theta)$ as the probability limit of $\hat{Q}_n(\theta)$ for any θ, a necessary condition for (iii) to be satisfied. This calculation can be accomplished by applying the law of large numbers, or hypotheses about convergence of certain components. For example, the law of large numbers implies that for MLE the limit of $\hat{Q}_n(\theta)$ is $Q_0(\theta) = E[\ln f(z|\theta)]$ and for NLS $Q_0(\theta) = -E[\{y - h(x,\theta)\}^2]$. Note the role played here by the normalization of the log-likelihood and sum of squared residuals, that leads to the objective function converging to a nonzero limit. Similar calculations give the limit for GMM and CMD, as further discussed below. Once this limit has been found, the consistency will follow from the conditions of Theorem 2.1.

One device that may allow for consistency under weaker conditions is to treat $\hat{\theta}$ as a maximum of $\hat{Q}_n(\theta) - \hat{Q}_n(\theta_0)$ rather than just $\hat{Q}_n(\theta)$. This is a magnitude normalization that sometimes makes it possible to weaken hypotheses on existence of moments. In the censored least absolute deviations example, where $\hat{Q}_n(\theta) = -n^{-1}\sum_{i=1}^n |y_i - \max\{0, x_i'\theta\}|$, an assumption on existence of the expectation of y is useful for applying a law of large numbers to show convergence of $\hat{Q}_n(\theta)$. In contrast $\hat{Q}_n(\theta) - \hat{Q}_n(\theta_0) = -n^{-1}\sum_{i=1}^n [|y_i - \max\{0, x_i'\theta\}| - |y_i - \max\{0, x_i'\theta_0\}|]$ is a bounded function of y_i, so that no such assumption is needed.

2.2. *Identification*

The identification condition for consistency of an extremum estimator is that the limit of the objective function has a unique maximum at the truth.[11] This condition is related to identification in the usual sense, which is that the distribution of the data at the true parameter is different than that at any other possible parameter value. To be precise, identification is a necessary condition for the limiting objective function to have a unique maximum, but it is not in general sufficient.[12] This section focuses on identification conditions for MLE, NLS, GMM, and CMD, in order to illustrate the kinds of results that are available.

2.2.1. *The maximum likelihood estimator*

An important feature of maximum likelihood is that identification is also sufficient for a unique maximum. Let $Y_1 \neq Y_2$ for random variables mean $\text{Prob}(\{Y_1 \neq Y_2\}) > 0$.

Lemma 2.2 (Information inequality)

If θ_0 is identified [$\theta \neq \theta_0$ and $\theta \in \Theta$ implies $f(z|\theta) \neq f(z|\theta_0)$] and $E[|\ln f(z|\theta)|] < \infty$ for all θ then $Q_0(\theta) = E[\ln f(z|\theta)]$ has a unique maximum at θ_0.

Proof

By the strict version of Jensen's inequality, for any nonconstant, positive random variable Y, $-\ln(E[Y]) < E[-\ln(Y)]$.[13] Then for $a = f(z|\theta)/f(z|\theta_0)$ and $\theta \neq \theta_0$, $Q_0(\theta_0) - Q_0(\theta) = E[\{-\ln[f(z|\theta)/f(z|\theta_0)]\}] > -\ln E[\{f(z|\theta)/f(z|\theta_0)\}] = -\ln[\int f(z|\theta)\mathrm{d}z] = 0$. Q.E.D.

The term "information inequality" refers to an interpretation of $Q_0(\theta)$ as an information measure. This result means that MLE has the very nice feature that uniqueness of the maximum of the limiting objective function occurs under the very weakest possible condition of identification of θ_0.

Conditions for identification in particular models are specific to those models. It

[11] If the set of maximands $\mathcal{M}$ of the objective function has more than one element, then this set does not distinguish between the true parameter and other values. In this case further restrictions are needed for identification. These restrictions are sometimes referred to as normalizations. Alternatively, one could work with convergence in probability to a set $\mathcal{M}$, but imposing normalization restrictions is more practical, and is needed for asymptotic normality.

[12] If θ_0 is not identified, then there will be some $\bar{\theta} \neq \theta_0$ such that the distribution of the data is the same when $\bar{\theta}$ is the true parameter value as when θ_0 is the true parameter value. Therefore, $Q_0(\theta)$ will also be limiting objective function when $\bar{\theta}$ is the true parameter, and hence the requirement that $Q_0(\theta)$ be maximized at the true parameter implies that $Q_0(\theta)$ has at least two maxima, θ_0 and $\bar{\theta}$.

[13] The strict version of Jensen's inequality states that if $a(y)$ is a strictly concave function [e.g. $a(y) = \ln(y)$] and Y is a nonconstant random variable, then $a(E[Y]) > E[a(Y)]$.

is often possible to specify them in a way that is easy to interpret (i.e. in a "primitive" way), as in the Cauchy example.

Example 1.1 continued

It will follow from Lemma 2.2 that $E[\ln f(z|\theta)]$ has a unique maximum at the true parameter. Existence of $E[|\ln f(z|\theta)|]$ for all θ follows from $|\ln f(z|\theta)| \leqslant C_1 + \ln(1+\sigma^{-2}|z-\mu|^2) \leqslant C_1 + \ln(C_2 + C_3|z|^2)$ for positive constants C_1, C_2, and C_3, and existence of $E[\ln(C_2 + C_3|z|^2)]$. Identification follows from $f(z|\theta)$ being one-to-one in the quadratic function $(1 + [(z-\mu)/\sigma]^2)$, the fact that quadratic functions intersect at no more than two points, and the fact that the probability of any two points is zero, so that $\text{Prob}(\{z: f(z|\theta) \neq f(z|\theta_0)\}) = 1 > 0$. Thus, by the information inequality, $E[\ln f(z|\theta)]$ has a unique maximum at θ_0. This example illustrates that it can be quite easy to show that the expected log-likelihood has a unique maximum, even when the first-order conditions for the MLE do not have unique roots.

Example 1.2 continued

Throughout the probit example, the identification and regularity conditions will be combined in the assumption that the second-moment matrix $E[xx']$ exists and is nonsingular. This assumption implies identification. To see why, note that nonsingularity of $E[xx']$ implies that it is positive definite. Let $\theta \neq \theta_0$, so that $E[\{x'(\theta-\theta_0)\}^2] = (\theta-\theta_0)'E[xx'](\theta-\theta_0) > 0$, implying that $x'(\theta-\theta_0) \neq 0$, and hence $x'\theta \neq x'\theta_0$, where as before "not equals" means "not equal on a set of positive probability". Both $\Phi(v)$ and $\Phi(-v)$ are strictly monotonic, so that $x'\theta \neq x'\theta_0$ implies both $\Phi(x'\theta) \neq \Phi(x'\theta_0)$ and $1 - \Phi(x'\theta) \neq 1 - \Phi(x'\theta_0)$, and hence that $f(z|\theta) = \Phi(x'\theta)^y[1 - \Phi(x'\theta)]^{1-y} \neq f(z|\theta_0)$.

Existence of $E[xx']$ also implies that $E[|\ln f(z|\theta)|] < \infty$. It is well known that the derivative $\mathrm{d}\ln\Phi(v)/\mathrm{d}v = \lambda(v) = \phi(v)/\Phi(v)$ [for $\phi(v) = \nabla_v\Phi(v)$], is convex and asymptotes to $-v$ as $v \to -\infty$ and to zero as $v \to \infty$. Therefore, a mean-value expansion around $\theta = 0$ gives

$$|\ln\Phi(x'\theta)| = |\ln\Phi(0) + \lambda(x'\tilde{\theta})x'\theta| \leqslant |\ln\Phi(0)| + \lambda(x'\tilde{\theta})|x'\theta|$$
$$\leqslant |\ln\Phi(0)| + C(1+|x'\tilde{\theta}|)|x'\theta| \leqslant |\ln\Phi(0)| + C(1+\|x\|\,\|\theta\|)\|x\|\,\|\theta\|.$$

Since $1 - \Phi(v) = \Phi(-v)$ and y is bounded, $|\ln f(z|\theta)| \leqslant 2[|\ln\Phi(0)| + C(1 + \|x\| \times \|\theta\|)\|x\|\,\|\theta\|]$, so existence of second moments of x implies that $E[|\ln f(z|\theta)|]$ is finite. This part of the probit example illustrates the detailed work that may be needed to verify that moment existence assumptions like that of Lemma 2.2 are satisfied.

2.2.2. *Nonlinear least squares*

The identification condition for NLS is that the mean square error $E[\{y - h(x,\theta)\}^2] = -Q_0(\theta)$ have a unique minimum at θ_0. As is easily shown, the mean square error

has a unique minimum at the conditional mean.[14] Since $h(x, \theta_0) = E[y|x]$ is the conditional mean, the identification condition for NLS is that $h(x, \theta) \neq h(x, \theta_0)$ if $\theta \neq \theta_0$, i.e. that $h(x, \theta)$ is not the conditional mean when $\theta \neq \theta_0$. This is a natural "conditional mean" identification condition for NLS.

In some cases identification will not be sufficient for conditional mean identification. Intuitively, only parameters that affect the first conditional moment of y given x can be identified by NLS. For example, if θ includes conditional variance parameters, or parameters of other higher-order moments, then these parameters may not be identified from the conditional mean.

As for identification, it is often easy to give primitive hypotheses for conditional mean identification. For example, in the linear model $h(x, \theta) = x'\theta$ conditional mean identification holds if $E[xx']$ is nonsingular, for then $\theta \neq \theta_0$ implies $x'\theta \neq x'\theta_0$, as shown in the probit example. For another example, suppose x is a positive scalar and $h(x, \theta) = \alpha + \beta x^{\gamma}$. As long as both β_0 and γ_0 are nonzero, the regression curve for a different value of θ intersects the true curve at most at three x points. Thus, for identification it is sufficient that x have positive density over any interval, or that x have more than three points that have positive probability.

2.2.3. *Generalized method of moments*

For generalized method of moments the limit function $Q_0(\theta)$ is a little more complicated than for MLE or NLS, but is still easy to find. By the law of large numbers, $\hat{g}_n(\theta) \xrightarrow{p} g_0(\theta) = E[g(z, \theta)]$, so that if $\hat{W} \xrightarrow{p} W$ for some positive semi-definite matrix W, then by continuity of multiplication, $\hat{Q}_n(\theta) \xrightarrow{p} Q_0(\theta) = -g_0(\theta)'Wg_0(\theta)$. This function has a maximum of zero at θ_0, so θ_0 will be identified if it is less than zero for $\theta \neq \theta_0$.

Lemma 2.3 (GMM identification)

If W is positive semi-definite and, for $g_0(\theta) = E[g(z, \theta)]$, $g_0(\theta_0) = 0$ and $Wg_0(\theta) \neq 0$ for $\theta \neq \theta_0$ then $Q_0(\theta) = -g_0(\theta)'Wg_0(\theta)$ has a unique maximum at θ_0.

Proof

Let R be such that $R'R = W$. If $\theta \neq \theta_0$, then $0 \neq Wg_0(\theta) = R'Rg_0(\theta)$ implies $Rg_0(\theta) \neq 0$ and hence $Q_0(\theta) = -[Rg_0(\theta)]'[Rg_0(\theta)] < Q_0(\theta_0) = 0$ for $\theta \neq \theta_0$. Q.E.D.

The GMM identification condition is that if $\theta \neq \theta_0$ then $g_0(\theta)$ is not in the null space of W, which for nonsingular W reduces to $g_0(\theta)$ being nonzero if $\theta \neq \theta_0$. A necessary order condition for GMM identification is that there be at least as many moment

[14] For $m(x) = E[y|x]$ and $a(x)$ any function with finite variance, iterated expectations gives $E[\{y - a(x)\}^2] = E[\{y - m(x)\}^2] + 2E[\{y - m(x)\}\{m(x) - a(x)\}] + E[\{m(x) - a(x)\}^2] \geq E[\{y - m(x)\}^2]$, with strict inequality if $a(x) \neq m(x)$.

functions as parameters. If there are fewer moments than parameters, then there will typically be many solutions to $g_0(\theta) = 0$.

If the moment functions are linear, say $g(z, \theta) = g(z) + G(z)\theta$, then the necessary and sufficient rank condition for GMM identification is that the rank of $WE[G(z)]$ is equal to the number of columns. For example, consider a linear instrumental variables estimator, where $g(z, \theta) = x \cdot (y - Y'\theta)$ for a residual $y - Y'\theta$ and a vector of instrumental variables x. The two-stage least squares estimator of θ is a GMM estimator with $\hat{W} = (\sum_{i=1}^{n} x_i x_i'/n)^{-1}$. Suppose that $E[xx']$ exists and is nonsingular, so that $W = (E[xx'])^{-1}$ by the law of large numbers. Then the rank condition for GMM identification is $E[xY']$ has full column rank, the well known instrumental variables identification condition. If $E[Y'|x] = x'\pi$ then this condition reduces to π having full column rank, a version of the single equation identification condition [see F.M. Fisher (1976) Theorem 2.7.1]. More generally, $E[xY'] = E[xE[Y'|x]]$, so that GMM identification is the same as x having "full rank covariance" with $E[Y|x]$.

If $E[g(z, \theta)]$ is nonlinear in θ, then specifying primitive conditions for identification becomes quite difficult. Here conditions for identification are like conditions for unique solutions of nonlinear equations (as in $E[g(z, \theta)] = 0$), which are known to be difficult. This difficulty is another reason to avoid formulating $\hat{\theta}$ as the solution to the first-order condition when analyzing consistency, e.g. to avoid interpreting MLE as a GMM estimator with $g(z, \theta) = \nabla_\theta \ln f(z|\theta)$. In some cases this difficulty is unavoidable, as for instrumental variables estimators of nonlinear simultaneous equations models.[15]

Local identification analysis may be useful when it is difficult to find primitive conditions for (global) identification. If $g(z, \theta)$ is continuously differentiable and $\nabla_\theta E[g(z, \theta)] = E[\nabla_\theta g(z, \theta)]$, then by Rothenberg (1971), a sufficient condition for a unique solution of $WE[g(z, \theta)] = 0$ in a (small enough) neighborhood of θ_0 is that $WE[\nabla_\theta g(z, \theta_0)]$ have full column rank. This condition is also necessary for local identification, and hence provides a necessary condition for global identification, when $E[\nabla_\theta g(z, \theta)]$ has constant rank in a neighborhood of θ_0 [i.e. in Rothenberg's (1971) "regular" case]. For example, for nonlinear 2SLS, where $\rho(z, \theta)$ is a residual and $g(z, \theta) = x \cdot \rho(z, \theta)$, the rank condition for local identification is that $E[x \cdot \nabla_\theta \rho(z, \theta_0)']$ has rank equal to its number of columns.

A practical "solution" to the problem of global GMM identification, that has often been adopted, is to simply assume identification. This practice is reasonable, given the difficulty of formulating primitive conditions, but it is important to check that it is not a vacuous assumption whenever possible, by showing identification in some special cases. In simple models it may be possible to show identification under particular forms for conditional distributions. The Hansen–Singleton model provides one example.

[15] There are some useful results on identification of nonlinear simultaneous equations models in Brown (1983) and Roehrig (1989), although global identification analysis of instrumental variables estimators remains difficult.

Example 1.3 continued

Suppose that $\hat{W} = (n^{-1}\sum_{i=1}^{n} x_i x_i')$, so that the GMM estimator is nonlinear two-stage least squares. By the law of large numbers, if $E[xx']$ exists and is nonsingular, $\hat{W}$ will converge in probability to $W = (E[xx'])^{-1}$, which is nonsingular. Then the GMM identification condition is that there is a unique solution to $E[x\rho(z,\theta)] = 0$ at $\theta = \theta_0$, where $\rho(z,\theta) = \{\beta w y^{\gamma} - 1\}$. Quite primitive conditions for identification can be formulated in a special log-linear case. Suppose that $w = \exp[a(x) + u]$ and $y = \exp[b(x) + v]$, where (u, v) is independent of x, that $a(x) + \gamma_0 b(x)$ is constant, and that $\eta(\theta_0) = 1$ for $\eta(\theta) = \exp[a(x) + \gamma_0 b(x)]\beta E[\exp(u + \gamma v)]$. Suppose also that the first element is a constant, so that the other elements can be assumed to have mean zero (by "demeaning" if necessary, which is a nonsingular linear transformation, and so does not affect the identification analysis). Let $\alpha(x,\gamma) = \exp[(\gamma - \gamma_0)b(x)]$. Then $E[\rho(z,\theta)|x] = \alpha(x,\gamma)\eta(\theta) - 1$, which is zero for $\theta = \theta_0$, and hence $E[g(z,\theta_0)] = 0$. For $\theta \neq \theta_0$, $E[g(z,\theta)] = \{E[\alpha(x,\gamma)]\eta(\theta) - 1, \mathrm{Cov}[x', \alpha(x,\gamma)]\eta(\theta)\}'$. This expression is nonzero if $\mathrm{Cov}[x, \alpha(x,\gamma)]$ is nonzero, because then the second term is nonzero if $\eta(\theta)$ is nonzero and the first term is nonzero if $\eta(\theta) = 0$. Furthermore, if $\mathrm{Cov}[x, \alpha(x,\gamma)] = 0$ for some γ, then all of the elements of $E[g(z,\theta)]$ are zero for all β, and one can choose $\beta > 0$ so the first element is zero. Thus, $\mathrm{Cov}[x, \alpha(x,\gamma)] \neq 0$ for $\gamma \neq \gamma_0$ is a necessary and sufficient condition for identification. In other words, the identification condition is that for all γ in the parameter set, some coefficient of a nonconstant variable in the regression of $\alpha(x,\gamma)$ on x is nonzero. This is a relatively primitive condition, because we have some intuition about when regression coefficients are zero, although it does depend on the form of $b(x)$ and the distribution of x in a complicated way. If $b(x)$ is a nonconstant, monotonic function of a linear combination of x, then this covariance will be nonzero.[16] Thus, in this example it is found that the assumption of GMM identification is not vacuous, that there are some nice special cases where identification does hold.

2.2.4. *Classical minimum distance*

The analysis of CMD identification is very similar to that for GMM. If $\hat{\pi} \xrightarrow{p} \pi_0$ and $\hat{W} \xrightarrow{p} W$, W positive semi-definite, then $\hat{Q}(\theta) = -[\hat{\pi} - h(\theta)]'\hat{W}[\hat{\pi} - h(\theta)] \xrightarrow{p} -[\pi_0 - h(\theta)]'W[\pi_0 - h(\theta)] = Q_0(\theta)$. The condition for $Q_0(\theta)$ to have a unique maximum (of zero) at θ_0 is that $h(\theta_0) = \pi_0$ and $h(\theta) - h(\theta_0)$ is not in the null space of W if $\theta \neq \theta_0$, which reduces to $h(\theta) \neq h(\theta_0)$ if W is nonsingular. If $h(\theta)$ is linear in θ then there is a readily interpretable rank condition for identification, but otherwise the analysis of global identification is difficult. A rank condition for local identification is that the rank of $W \cdot \nabla_\theta h(\theta_0)$ equals the number of components of θ.

[16] It is well known that $\mathrm{Cov}[x, f(x)] \neq 0$ for any monotonic, nonconstant function $f(x)$ of a random variable x.

2.3. Uniform convergence and continuity

Once conditions for identification have been found and compactness of the parameter set has been assumed, the only other primitive conditions for consistency required by Theorem 2.1 are those for uniform convergence in probability and continuity of the limiting objective function. This subsection gives primitive hypotheses for these conditions that, when combined with identification, lead to primitive conditions for consistency of particular estimators.

For many estimators, results on uniform convergence of sample averages, known as *uniform laws of large numbers*, can be used to specify primitive regularity conditions. Examples include MLE, NLS, and GMM, each of which depends on sample averages. The following uniform law of large numbers is useful for these estimators. Let $a(z, \theta)$ be a matrix of functions of an observation z and the parameter θ, and for a matrix $A = [a_{jk}]$, let $\| A \| = (\sum_{j,k} a_{jk}^2)^{1/2}$ be the Euclidean norm.

Lemma 2.4

If the data are i.i.d., Θ is compact, $a(z_i, \theta)$ is continuous at each $\theta \in \Theta$ with probability one, and there is $d(z)$ with $\| a(z, \theta) \| \leq d(z)$ for all $\theta \in \Theta$ and $E[d(z)] < \infty$, then $E[a(z, \theta)]$ is continuous and $\sup_{\theta \in \Theta} \| n^{-1} \sum_{i=1}^{n} a(z_i, \theta) - E[a(z, \theta)] \| \xrightarrow{p} 0$.

The conditions of this result are similar to assumptions of Wald's (1949) consistency proof, and it is implied by Lemma 1 of Tauchen (1985).

The conditions of this result are quite weak. In particular, they allow for $a(z, \theta)$ to not be continuous on all of Θ for given z.[17] Consequently, this result is useful even when the objective function is not continuous, as for Manski's (1975) maximum score estimator and the simulation-based estimators of Pakes (1986) and McFadden (1989). Also, this result can be extended to dependent data. The conclusion remains true if the i.i.d. hypothesis is changed to strict stationarity and ergodicity of z_i.[18]

The two conditions imposed on $a(z, \theta)$ are a continuity condition and a moment existence condition. These conditions are very primitive. The continuity condition can often be verified by inspection. The moment existence hypothesis just requires a data-dependent upper bound on $\| a(z, \theta) \|$ that has finite expectation. This condition is sometimes referred to as a "dominance condition", where $d(z)$ is the dominating function. Because it only requires that certain moments exist, it is a "regularity condition" rather than a "substantive restriction".

It is often quite easy to see that the continuity condition is satisfied and to specify moment hypotheses for the dominance condition, as in the examples.

[17] The conditions of Lemma 2.4 are not sufficient for measurability of the supremum in the conclusion, but are sufficient for convergence of the supremum in outer measure. Convergence in outer measure is sufficient for consistency of the estimator in terms of outer measure, a result that is useful when the objective function is not continuous, as previously noted.

[18] Strict stationarity means that the distribution of $(z_i, z_{i+1}, \ldots, z_{i+m})$ does not depend on i for any m, and ergodicity implies that $n^{-1} \sum_{i=1}^{n} a(z_i) \longrightarrow E[a(z_i)]$ for (measurable) functions $a(z)$ with $E[|a(z)|] < \infty$.

Example 1.1 continued

For the Cauchy location-scale likelihood, continuity of $\ln f(z|\theta) = \ln C - \ln \sigma - \ln(1 + \{(z-\mu)/\sigma\}^2)$ is obvious. Also, as in the Example 1.1 discussion in Section 2.2.1, for any Θ where θ is bounded and $\sigma \geq 0$ is bounded away from zero, the dominance condition of Lemma 2.4 is satisfied for $a(z,\theta) = \ln f(z|\theta)$ and $d(z) = C_1 + \ln(C_2 + C_3|z|^2)$, for certain positive constants C_1, C_2, and C_3. Thus, by the conclusion of Lemma 2.4, $E[\ln f(z|\theta)]$ is continuous and the average log-likelihood converges uniformly in probability to the expected log-likelihood.

Example 1.2 continued

For the probit example, continuity of $\ln f(z|\theta) = y \ln \Phi(x'\theta) + (1-y)\ln \Phi(-x'\theta)$ is obvious, while the dominance condition of Lemma 2.4 follows as in Section 2.2.1, with $C(1 + \|x\|^2) = d(z)$. Then the conclusion of Lemma 2.4 applies to $a(z,\theta) = \ln f(z|\theta)$.

Example 1.3 continued

In the Hansen–Singleton example, the GMM objective function depends on θ through the average moment functions $\hat{g}_n(\theta) = n^{-1}\sum_{i=1}^n g(z_i,\theta) = n^{-1}\sum_{i=1}^n x_i \times (\beta w_i y_i^{\gamma} - 1)$. Consequently, as shown below for general GMM estimators, uniform convergence of the objective function and continuity of the limit will hold if the hypotheses of Lemma 2.4 are satisfied with $a(z,\theta)$ equal to each element of $g(z,\theta)$. By inspection, each element of $g(z,\theta)$ is continuous. Also, assuming Θ is specified so that β and γ are bounded, and letting β_ℓ, β_u, and γ_ℓ, γ_u denote upper and lower bounds, respectively,

$$\|g(z,\theta)\| \leq \|x\|[1 + (|\beta_\ell| + |\beta_u|)|w|(|y|^{\gamma_u} + |y|^{\gamma_\ell})], \quad \theta \in \Theta.$$

Thus, the dominance condition will be satisfied if each of $\|x\||w||y|^{\gamma_u}$, $\|x\||w||y|^{\gamma_\ell}$ and $\|x\|$ have finite expectations. In this example existence of $E[\|x\||w||y|^{\gamma_u}]$ and $E[\|x\||w||y|^{\gamma_\ell}]$ may place bounds on how large or small γ can be allowed to be.

Lemma 2.4 is useful, but it only applies to stationary data and to sample averages. There are many examples of models and estimators in econometrics where more general uniform convergence results are needed. It is possible to formulate necessary and sufficient conditions for uniform convergence using a *stochastic equicontinuity* condition. Stochastic equicontinuity is an important concept in recent developments in asymptotic theory, is used elsewhere in this chapter, and is fully discussed in Andrews' chapter in this volume. However, because this concept is somewhat more technical, and not needed for many results, we have placed the discussion of uniform convergence and stochastic equicontinuity in Section 2.7, and left all description of its other uses until needed in Section 7.

2.4. *Consistency of maximum likelihood*

The conditions for identification in Section 2.2 and the uniform convergence result of Lemma 2.4, allow specification of primitive regularity conditions for particular kinds of estimators. A consistency result for MLE can be formulated as follows:

Theorem 2.5

Suppose that z_i, $(i = 1, 2, \ldots)$, are i.i.d. with p.d.f. $f(z_i|\theta_0)$ and (i) if $\theta \neq \theta_0$ then $f(z_i|\theta) \neq f(z_i|\theta_0)$; (ii) $\theta_0 \in \Theta$, which is compact; (iii) $\ln f(z_i|\theta)$ is continuous at each $\theta \in \Theta$ with probability one; (iv) $E[\sup_{\theta \in \Theta} |\ln f(z|\theta)|] < \infty$. Then $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof

Proceed by verifying the conditions of Theorem 2.1. Condition 2.1(i) follows by 2.5(i) and (iv) and Lemma 2.2. Condition 2.1(ii) holds by 2.5(ii). Conditions 2.1(iii) and (iv) follow by Lemma 2.4. Q.E.D.

The conditions of this result are quite primitive and also quite weak. The conclusion is consistency of the MLE. Thus, a particular MLE can be shown to be consistent by checking the conditions of this result, which are identification, compactness, continuity of the log-likelihood at particular points, and a dominance condition for the log-likelihood. Often it is easy to specify conditions for identification, continuity holds by inspection, and the dominance condition can be shown to hold with a little algebra. The Cauchy location-scale model is an example.

Example 1.1 continued

To show consistency of the Cauchy MLE, one can proceed to verify the hypotheses of Theorem 2.5. Condition (i) was shown in Section 2.2.1. Conditions (iii) and (iv) were shown in Section 2.3. Then the conditions of Theorem 2.5 imply that when Θ is any compact set containing θ_0, the Cauchy MLE is consistent.

A similar result can be stated for probit (i.e. Example 1.2). It is not given here because it is possible to drop the compactness hypothesis of Theorem 2.5. The probit log-likelihood turns out to be concave in parameters, leading to a simple consistency result without a compact parameter space. This result is discussed in Section 2.6.

Theorem 2.5 remains true if the i.i.d. assumption is replaced with the condition that $z_1, z_2, \ldots$ is stationary and ergodic with (marginal) p.d.f. of z_i given by $f(z|\theta_0)$. This relaxation of the i.i.d. assumption is possible because the limit function remains unchanged (so the information inequality still applies) and, as noted in Section 2.3, uniform convergence and continuity of the limit still hold.

A similar consistency result for NLS could be formulated by combining conditional mean identification, compactness of the parameter space, $h(x, \theta)$ being conti-

nuous at each θ with probability one, and a dominance condition. Formulating such a result is left as an exercise.

2.5. *Consistency of GMM*

A consistency result for GMM can be formulated as follows:

Theorem 2.6

Suppose that z_i, $(i = 1, 2, \ldots)$, are i.i.d., $\hat{W} \xrightarrow{p} W$, and (i) W is positive semi-definite and $WE[g(z, \theta)] = 0$ only if $\theta = \theta_0$; (ii) $\theta_0 \in \Theta$, which is compact; (iii) $g(z, \theta)$ is continuous at each $\theta \in \Theta$ with probability one; (iv) $E[\sup_{\theta \in \Theta} \| g(z, \theta) \|] < \infty$. Then $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof

Proceed by verifying the hypotheses of Theorem 2.1. Condition 2.1(i) follows by 2.6(i) and Lemma 2.3. Condition 2.1(ii) holds by 2.6(ii). By Lemma 2.4 applied to $a(z, \theta) = g(z, \theta)$, for $\hat{g}_n(\theta) = n^{-1} \sum_{i=1}^{n} g(z_i, \theta)$ and $g_0(\theta) = E[g(z, \theta)]$, one has $\sup_{\theta \in \Theta} \| \hat{g}_n(\theta) - g_0(\theta) \| \xrightarrow{p} 0$ and $g_0(\theta)$ is continuous. Thus, 2.1(iii) holds by $Q_0(\theta) = -g_0(\theta)' W g_0(\theta)$ continuous. By Θ compact, $g_0(\theta)$ is bounded on Θ, and by the triangle and Cauchy–Schwartz inequalities,

$$
\begin{aligned}
&|\hat{Q}_n(\theta) - Q_0(\theta)| \\
&\quad \leqslant |[\hat{g}_n(\theta) - g_0(\theta)]' \hat{W} [\hat{g}_n(\theta) - g_0(\theta)]| + |g_0(\theta)' (\hat{W} + \hat{W}')[\hat{g}_n(\theta) - g_0(\theta)]| \\
&\qquad + |g_0(\theta)' (\hat{W} - W) g_0(\theta)| \\
&\quad \leqslant \| \hat{g}_n(\theta) - g_0(\theta) \|^2 \| \hat{W} \| + 2 \| g_0(\theta) \| \| \hat{g}_n(\theta) - g_0(\theta) \| \| \hat{W} \| \\
&\qquad + \| g_0(\theta) \|^2 \| \hat{W} - W \|,
\end{aligned}
$$

so that $\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$, and 2.1(iv) holds. Q.E.D.

The conditions of this result are quite weak, allowing for discontinuity in the moment functions.[19] Consequently, this result is general enough to cover the simulated moment estimators of Pakes (1986) and McFadden (1989), or the interval moment estimator of Newey (1988).

To use this result to show consistency of a GMM estimator, one proceeds to check the conditions, as in the Hansen–Singleton example.

[19] Measurability of the estimator becomes an issue in this case, although this can be finessed by working with outer measure, as previously noted.

Example 1.3 continued

Assume that $E[xx'] < \infty$, so that $\hat{W} \xrightarrow{p} W = (E[xx'])^{-1}$. For hypothesis (i), simply assume that $E[g(z, \theta)] = 0$ has a unique solution at θ_0 among all $\theta \in \Theta$. Unfortunately, as discussed in Section 2.2, it is difficult to give more primitive assumptions for this identification condition. Also, assume that Θ is compact, so that (ii) holds. Then (iii) holds by inspection, and as discussed in Section 2.3, (iv) holds as long as the moment existence conditions given there are satisfied. Thus, under these assumptions, the estimator will be consistent.

Theorem 2.6 remains true if the i.i.d. assumption is replaced with the condition that $z_1, z_2, \ldots$ is stationary and ergodic. Also, a similar consistency result could be formulated for CMD, by combining uniqueness of the solution to $\pi_0 = h(\theta)$ with compactness of the parameter space and continuity of $h(\theta)$. Details are left as an exercise.

2.6. *Consistency without compactness*

The compactness assumption is restrictive, because it implicitly requires that there be known bounds on the true parameter value. It is useful in practice to be able to drop this restriction, so that conditions for consistency without compactness are of interest. One nice result is available when the objective function is concave. Intuitively, concavity prevents the objective function from "turning up" as the parameter moves far away from the truth. A precise result based on this intuition is the following one:

Theorem 2.7

If there is a function $Q_0(\theta)$ such that (i) $Q_0(\theta)$ is uniquely maximized at θ_0; (ii) θ_0 is an element of the interior of a convex set Θ and $\hat{Q}_n(\theta)$ is concave; and (iii) $\hat{Q}_n(\theta) \xrightarrow{p} Q_0(\theta)$ for all $\theta \in \Theta$, then $\hat{\theta}_n$ exists with probability approaching one and $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Proof

Let $\mathscr{C}$ be a closed sphere of radius 2ε around θ_0 that is contained in the interior of Θ and let $\mathscr{B}$ be its boundary. Concavity is preserved by pointwise limits, so that $Q_0(\theta)$ is also concave. A concave function is continuous on the interior of its domain, so that $Q_0(\theta)$ is continuous on $\mathscr{C}$. Also, by Theorem 10.8 of Rockafellar (1970), pointwise convergence of concave functions on a dense subset of an open set implies uniform convergence on any compact subset of the open set. It then follows as in Andersen and Gill (1982) that $\hat{Q}_n(\theta)$ converges to $Q_0(\theta)$ in probability uniformly on any compact subset of Θ, and in particular on $\mathscr{C}$. Hence, by Theorem 2.1, the maximand $\tilde{\theta}_n$ of $\hat{Q}_n(\theta)$ on $\mathscr{C}$ is consistent for θ_0. Then the event that $\tilde{\theta}_n$ is within ε of θ_0, so that $\hat{Q}_n(\tilde{\theta}_n) \geqslant \max_{\mathscr{B}} \hat{Q}_n(\theta)$, occurs with probability approaching one. In this event, for any θ outside $\mathscr{C}$, there is a linear convex combination $\lambda\tilde{\theta}_n + (1-\lambda)\theta$

that lies in $\mathscr{B}$ (with $\lambda < 1$), so that $\hat{Q}_n(\tilde{\theta}_n) \geq \hat{Q}_n[\lambda\tilde{\theta}_n + (1-\lambda)\theta]$. By concavity, $\hat{Q}_n[\lambda\tilde{\theta}_n + (1-\lambda)\theta] \geq \lambda\hat{Q}_n(\tilde{\theta}_n) + (1-\lambda)\hat{Q}_n(\theta)$. Putting these inequalities together, $(1-\lambda)\hat{Q}_n(\tilde{\theta}) \geq (1-\lambda)\hat{Q}_n(\theta)$, implying $\tilde{\theta}_n$ is the maximand over Θ. Q.E.D.

This theorem is similar to Corollary II.2 of Andersen and Gill (1982) and Lemma A of Newey and Powell (1987). In addition to allowing for noncompact Θ, it only requires pointwise convergence. This weaker hypothesis is possible because pointwise convergence of concave functions implies uniform convergence (see the proof). This result also contains the additional conclusion that $\hat{\theta}$ exists with probability approaching one, which is needed because of noncompactness of Θ.

This theorem leads to simple conditions for consistency without compactness for both MLE and GMM. For MLE, if in Theorem 2.5, (ii)–(iv) are replaced by Θ convex, $\ln f(z|\theta)$ concave in θ (with probability one), and $E[|\ln f(z|\theta)|] < \infty$ for all θ, then the law of large numbers and Theorem 2.7 give consistency. In other words, with concavity the conditions of Lemma 2.2 are sufficient for consistency of the MLE. Probit is an example.

Example 1.2 continued

It was shown in Section 2.2.1 that the conditions of Lemma 2.2 are satisfied. Thus, to show consistency of the probit MLE it suffices to show concavity of the log-likelihood, which will be implied by concavity of $\ln \Phi(x'\theta)$ and $\ln \Phi(-x'\theta)$. Since $x'\theta$ is linear in θ, it suffices to show concavity of $\ln \Phi(v)$ in v. This concavity follows from the well known fact that $\mathrm{d} \ln \Phi(v)/\mathrm{d}v = \phi(v)/\Phi(v)$ is monotonic decreasing [as well as the general Pratt (1981) result discussed below].

For GMM, if $g(z, \theta)$ is linear in θ and $\hat{W}$ is positive semi-definite then the objective function is concave, so if in Theorem 2.6, (ii)–(iv) are replaced by the requirement that $E[\|g(z, \theta)\|] < \infty$ for all $\theta \in \Theta$, the conclusion of Theorem 2.7 will give consistency of GMM. This linear moment function case includes linear instrumental variables estimators, where compactness is well known to not be essential.

This result can easily be generalized to estimators with objective functions that are concave after reparametrization. If conditions (i) and (iii) are satisfied and there is a one-to-one mapping $\tau(\theta)$ with continuous inverse such that $\hat{Q}_n[\tau^{-1}(\lambda)]$ is concave on $\tau(\Theta)$ and $\tau(\theta_0)$ is an element of the interior of $\tau(\Theta)$, then the maximizing value $\hat{\lambda}$ of $\hat{Q}_n[\tau^{-1}(\lambda)]$ will be consistent for $\lambda_0 = \tau(\theta_0)$ by Theorem 2.7 and invariance of a maxima to one-to-one reparametrization, and $\hat{\theta} = \tau^{-1}(\hat{\lambda})$ will be consistent for $\theta_0 = \tau^{-1}(\lambda_0)$ by continuity of the inverse.

An important class of estimators with objective functions that are concave after reparametrization are univariate continuous/discrete regression models with log-concave densities, as discussed in Olsen (1978) and Pratt (1981). To describe this class, first consider a continuous regression model $y = x'\beta_0 + \sigma_0\varepsilon$, where ε is independent of x with p.d.f. $g(\varepsilon)$. In this case the (conditional on x) log-likelihood is $-\ln\sigma + \ln g[\sigma^{-1}(y - x'\beta)]$ for $(\beta', \sigma) \in \Theta = \mathbb{R}^k x(0, \infty)$. If $\ln g(\varepsilon)$ is concave, then this

log-likelihood need not be concave, but the likelihood $\ln \gamma + \ln g(\gamma y - x'\delta)$ is concave in the one-to-one reparametrization $\gamma = \sigma^{-1}$ and $\delta = \beta/\sigma$. Thus, the average log-likelihood is also concave in these parameters, so that the above generalization of Theorem 2.7 implies consistency of the MLE estimators of β and σ when the maximization takes place over $\Theta = \mathbb{R}^k x(0, \infty)$, if $\ln g(\varepsilon)$ is concave. There are many log-concave densities, including those proportional to $\exp(-|x|^\alpha)$ for $\alpha \geqslant 1$ (including the Gaussian), logistic, and the gamma and beta when the p.d.f. is bounded, so this concavity property is shared by many models of interest.

The reparametrized log-likelihood is also concave when y is only partially observed. As shown by Pratt (1981), concavity of $\ln g(\varepsilon)$ also implies concavity of $\ln[G(v) - G(w)]$ in v and w, for the CDF $G(v) = \int_{-\infty}^{v} g(\varepsilon)\, d\varepsilon$.[20] That is, the log-probability of an interval will be concave in the endpoints. Consequently, the log-likelihood for partial observability will be concave in the parameters when each of the endpoints is a linear function of the parameters. Thus, the MLE will be consistent without compactness in partially observed regression models with log-concave densities, which includes probit, logit, Tobit, and ordered probit with unknown censoring points.

There are many other estimators with concave objective functions, where some version of Theorem 2.7 has been used to show consistency without compactness. These include the estimators in Andersen and Gill (1982), Newey and Powell (1987), and Honoré (1992).

It is also possible to relax compactness with some nonconcave objective functions. Indeed, the original Wald (1949) MLE consistency theorem allowed for noncompactness, and Huber (1967) has given similar results for other estimators. The basic idea is to bound the objective function above uniformly in parameters that are far enough away from the truth. For example, consider the MLE. Suppose that there is a compact set $\mathscr{C}$ such that $E[\sup_{\theta \in \Theta \cap \mathscr{C}^c} \ln f(z|\theta)] < E[\ln f(z|\theta_0)]$. Then by the law of large numbers, with probability approaching one, $\sup_{\theta \in \Theta \cap \mathscr{C}^c} \hat{Q}_n(\theta) \leqslant n^{-1} \times \sum_{i=1}^{n} \sup_{\theta \in \Theta \cap \mathscr{C}^c} \ln f(z_i|\theta) < n^{-1} \sum_{i=1}^{n} \ln f(z|\theta_0)$, and the maximum must lie in $\mathscr{C}$. Once the maximum is known to be in a compact set with probability approaching one, Theorem 2.1 applies to give consistency.

Unfortunately, the Wald idea does not work in regression models, which are quite common in econometrics. The problem is that the likelihood depends on regression parameters θ through linear combinations of the form $x'\theta$, so that for given x changing θ along the null-space of x' does not change the likelihood. Some results that do allow for regressors are given in McDonald and Newey (1988), where it is shown how compactness on Θ can be dropped when the objective takes the form $\hat{Q}_n(\theta) = n^{-1} \sum_{i=1}^{n} a(z_i, x_i'\theta)$ and $a(z, v)$ goes to $-\infty$ as v becomes unbounded. It would be useful to have other results that apply to regression models with nonconcave objective functions.

[20] Pratt (1981) also showed that concavity of $\ln g(\varepsilon)$ is necessary as well as sufficient for $\ln[G(v) - G(w)]$ to be concave over all v and w.

Compactness is essential for consistency of some extremum estimators. For example, consider the MLE in a model where z is a mixture of normals, having likelihood $f(z|\theta) = p\cdot\sigma^{-1}\phi[\sigma^{-1}(z-\mu)] + (1-p)\gamma^{-1}\phi[\gamma^{-1}(z-\alpha)]$ for $\theta = (\mu, \sigma, \alpha, \gamma)'$, some $0 < p < 1$, and the standard normal p.d.f. $\phi(\varepsilon) = (2\pi)^{-1/2}e^{-\varepsilon^2/2}$. An interpretation of this model is that z is drawn from $N(\mu, \sigma^2)$ with probability p and from $N(\alpha, \gamma^2)$ with probability $(1-p)$. The problem with noncompactness for the MLE in this model is that for certain μ (and α) values, the average log-likelihood becomes unbounded as σ (or γ) goes to zero. Thus, for existence and consistency of the MLE it is necessary to bound σ (and γ) away from zero. To be specific, suppose that $\mu = z_i$ for some i. Then $f(z_i|\theta) = p\cdot\sigma^{-1}\phi(0) + (1-p)\gamma^{-1}\phi[\gamma^{-1}(z_i-\alpha)] \to \infty$ as $\sigma \to 0$, and assuming that $z_j \neq z_i$ for all $j \neq i$, as occurs with probability one, $f(z_j|\theta) \to (1-p)\gamma^{-1}\phi[\gamma^{-1}(z_j-\alpha)] > 0$. Hence, $\hat{Q}_n(\theta) = n^{-1}\sum_{i=1}^n \ln f(z_i|\theta)$ becomes unbounded as $\sigma \to 0$ for $\mu = z_i$. In spite of this fact, if the parameter set is assumed to be compact, so that σ and γ are bounded away from zero, then Theorem 2.5 gives consistency of the MLE. In particular, it is straightforward to show that θ is identified, so that, by the information inequality, $E[\ln f(z|\theta)]$ has a unique maximum at θ_0. The problem here is that the convergence of the sample objective function is not uniform over small values of σ.

This example is extreme, but there are interesting econometric examples that have this feature. One of these is the disequilibrium model without observed regime of Fair and Jaffee (1972), where $y = \min\{x'\beta_0 + \sigma_0\varepsilon, w'\delta_0 + \gamma_0 u\}$, ε and u are standard normal and independent of each other and of x and w, and the regressors include constants. This model also has an unbounded average log-likelihood as $\sigma \to 0$ for a certain values of β, but the MLE over any compact set containing the truth will be consistent under the conditions of Theorem 2.5.

Unfortunately, as a practical matter one may not be sure about lower bounds on variances, and even if one were sure, extraneous maxima can appear at the lower bounds in small samples. An approach to this problem is to search among local maxima that satisfy the first-order conditions for the one that maximizes the likelihood. This approach may work in the normal mixture and disequilibrium models, but might not give a consistent estimator when the true value lies on the boundary (and the first-order conditions are not satisfied on the boundary).

2.7. *Stochastic equicontinuity and uniform convergence*

Stochastic equicontinuity is important in recent developments in asymptotic distribution theory, as described in the chapter by Andrews in this handbook. This concept is also important for uniform convergence, as can be illustrated by the nonstochastic case. Consider a sequence of continuous, nonstochastic functions $\{Q_n(\theta)\}_{n=1}^{\infty}$. For nonrandom functions, equicontinuity means that the "gap" between $Q_n(\tilde{\theta})$ and $Q_n(\theta)$ can be made small uniformly in n by making $\tilde{\theta}$ be close enough to θ, i.e. a sequence of functions is equicontinuous if they are continuous uniformly in

n. More precisely, equicontinuity holds if for each $\theta, \varepsilon > 0$ there exists $\delta > 0$ with $|Q_n(\tilde{\theta}) - Q_n(\theta)| < \varepsilon$ for all $\|\tilde{\theta} - \theta\| < \delta$ and all n.[21] It is well known that if $Q_n(\theta)$ converges to $Q_0(\theta)$ pointwise, i.e. for all $\theta \in \Theta$, and Θ is compact, then equicontinuity is a necessary and sufficient condition for uniform convergence [e.g. see Rudin (1976)]. The ideas behind it being a necessary and sufficient condition for uniform convergence is that pointwise convergence is the same as uniform covergence on any finite grid of points, and a finite grid of points can approximately cover a compact set, so that uniform convergence means that the functions cannot vary too much as θ moves off the grid.

To apply the same ideas to uniform convergence in probability it is necessary to define an "in probability" version of equicontinuity. The following version is formulated in Newey (1991a).

Stochastic equicontinuity: For every ε, $\eta > 0$ there exists a sequence of random variables $\hat{\Delta}_n$ and a sample size n_0 such that for $n \geqslant n_0$, $\text{Prob}(|\hat{\Delta}_n| > \varepsilon) < \eta$ and for each θ there is an open set $\mathcal{N}$ containing θ with

$$\sup_{\tilde{\theta} \in \mathcal{N}} |\hat{Q}_n(\tilde{\theta}) - \hat{Q}_n(\theta)| \leqslant \hat{\Delta}_n, \quad n \geqslant n_0.$$

Here the function $\hat{\Delta}_n$ acts like a "random epsilon", bounding the effect of changing θ on $\hat{Q}_n(\theta)$. Consequently, similar reasoning to the nonstochastic case can be used to show that stochastic equicontinuity is an essential condition for uniform convergence, as stated in the following result:

Lemma 2.8

Suppose Θ is compact and $Q_0(\theta)$ is continuous. Then $\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$ if and only if $\hat{Q}_n(\theta) \xrightarrow{p} Q_0(\theta)$ for all $\theta \in \Theta$ and $\hat{Q}_n(\theta)$ is stochastically equicontinuous.

The proof of this result is given in Newey (1991a). It is also possible to state an almost sure convergence version of this result, although this does not seem to produce the variety of conditions for uniform convergence that stochastic equicontinuity does; see Andrews (1992).

One useful sufficient condition for uniform convergence that is motivated by the form of the stochastic equicontinuity property is a global, "in probability" Lipschitz condition, as in the hypotheses of the following result. Let $O_p(1)$ denote a sequence of random variables that is bounded in probability.[22]

[21] One can allow for discontinuity in the functions by allowing the difference to be less than ε only for $n > \bar{n}$, where $\bar{n}$ depends on ε, but not on θ. This modification is closer to the stochastic equicontinuity condition given here, which does allow for discontinuity.

[22] Y_n is bounded in probability if for every $\varepsilon > 0$ there exists $\bar{n}$ and η such that $\text{Prob}(|Y_n| > \eta) < \varepsilon$ for $n > \bar{n}$.

Lemma 2.9

If Θ is compact, $Q_0(\theta)$ is continuous, $\hat{Q}_n(\theta) \xrightarrow{p} Q_0(\theta)$ for all $\theta \in \Theta$, and there is $\alpha > 0$ and $\hat{B}_n = O_p(1)$ such that for all $\tilde{\theta}, \theta \in \Theta$, $|\hat{Q}_n(\tilde{\theta}) - \hat{Q}_n(\theta)| \leq \hat{B}_n \| \tilde{\theta} - \theta \|^{\alpha}$, then $\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$.

Proof

By Lemma 2.8 it suffices to show stochastic equicontinuity. Pick ε, $\eta > 0$. By $\hat{B}_n = O_p(1)$ there is M such that $\text{Prob}(|\hat{B}_n| > M) < \eta$ for all n large enough. Let $\hat{\Delta}_n = \hat{B}_n \varepsilon / M$ and $\mathcal{N} = \{\tilde{\theta} : \| \tilde{\theta} - \theta \|^{\alpha} \leq \varepsilon / M\}$. Then $\text{Prob}(|\hat{\Delta}_n| > \varepsilon) = \text{Prob}(|\hat{B}_n| > M) < \eta$ and for all $\tilde{\theta}$, $\theta \in \mathcal{N}$, $|\hat{Q}_n(\tilde{\theta}) - \hat{Q}_n(\theta)| \leq \hat{B}_n \| \tilde{\theta} - \theta \|^{\alpha} \leq \hat{\Delta}_n$. Q.E.D.

This result is useful in formulating the uniform law of large numbers given in Wooldridge's chapter in this volume. It is also useful when the objective function $\hat{Q}_n(\theta)$ is not a simple function of sample averages (i.e. where uniform laws of large numbers do not apply). Further examples and discussion are given in Newey (1991a).

2.8. *Least absolute deviations examples*

Estimators that minimize a sum of absolute deviations provide interesting examples. The objective function that these estimators minimize is not differentiable, so that weak regularity conditions are needed for verifying consistency and asymptotic normality. Also, these estimators have certain robustness properties that make them interesting in their own right. In linear models the least absolute deviations estimator is known to be more asymptotically more efficient than least squares for thick-tailed distributions. In the binary choice and censored regression models the least absolute deviations estimator is consistent without any functional form assumptions on the distribution of the disturbance. The linear model has been much discussed in the statistics and economics literature [e.g. see Bloomfeld and Steiger (1983)], so it seems more interesting to consider here other cases. To this end two examples are given: maximum score, which applies to the binary choice model, and censored least absolute deviations.

2.8.1. *Maximum score*

The maximum score estimator of Manski (1975) is an interesting example because it has a noncontinuous objective function, where the weak regularity conditions of Lemma 2.4 are essential, and because it is a distribution-free estimator for binary choice. Maximum score is used to estimate θ_0 in the model $y = 1(x'\theta_0 + \varepsilon > 0)$, where $1(\mathcal{A})$ denotes the indicator for the event $\mathcal{A}$ (equal to one if $\mathcal{A}$ occurs and zero

otherwise), and ε is a disturbance term with a conditional median (given x) of zero.[23] The estimator solves eq. (1.1) for

$$\hat{Q}_n(\theta) = -n^{-1}\sum_{i=1}^{n} |y_i - 1(x_i'\theta > 0)|.$$

A scale normalization is necessary (as usual for binary choice), and a convenient one here is to restrict all elements of Θ to satisfy $\|\theta\| = 1$.

To show consistency of the maximum score estimator, one can use conditions for identification and Lemma 2.4 to directly verify all the hypotheses of Theorem 2.1. By the law of large numbers, $\hat{Q}_n(\theta)$ will have probability limit $Q_0(\theta) = -E[|y - 1(x'\theta > 0)|]$. To show that this limiting objective has a unique maximum at θ_0, one can use the well known result that for any random variable Y, the expected absolute deviation $E[|Y - a(x)|]$ is strictly minimized at any median of the conditional distribution of Y given x. For a binary variable such as y, the median is unique when $\text{Prob}(y = 1|x) \neq \frac{1}{2}$, equal to one when the conditional probability is more than $\frac{1}{2}$ and equal to zero when it is less than $\frac{1}{2}$. Assume that 0 is the unique conditional median of ε given x and that $\text{Prob}(x'\theta_0 = 0) = 0$. Then $\text{Prob}(y = 1|x) > (<)\ \frac{1}{2}$ if and only if $x'\theta_0 > (<)\ 0$, so $\text{Prob}(y = 1|x) = \frac{1}{2}$ occurs with probability zero, and hence $1(x'\theta_0 > 0)$ is the unique median of y given x. Thus, it suffices to show that $1(x'\theta > 0) \neq 1(x'\theta_0 > 0)$ if $\theta \neq \theta_0$. For this purpose, suppose that there are corresponding partitions $\theta = (\theta_1, \theta_2')'$ and $x = (x_1, x_2')'$ such that $x_2'\delta = 0$ only if $\delta = 0$; also assume that the conditional distribution of x_1 given x_2 is continuous with a p.d.f. that is positive on $\mathbb{R}$, and the coefficient θ_{01} of x_1 is nonzero. Under these conditions, if $\theta \neq \theta_0$ then $1(x'\theta > 0) \neq 1(x'\theta_0 > 0)$, the idea being that the continuous distribution of x_1 means that it is allowed that there is a region of x_1 values where the sign of $x'\theta$ is different. Also, under this condition, $x'\theta_0 = 0$ with zero probability, so y has a unique conditional median of $1(x'\theta_0 > 0)$ that differs from $1(x'\theta > 0)$ when $\theta \neq \theta_0$, so that $Q_0(\theta)$ has a unique maximum at θ_0.

For uniform convergence it is enough to assume that $x'\theta$ is continuously distributed for each θ. For example, if the coefficient of x_1 is nonzero for all $\theta \in \Theta$ then this condition will hold. Then, $1(x'\theta > 0)$ will be continuous at each θ with probability one, and by y and $1(x'\theta > 0)$ bounded, the dominance condition will be satisfied, so the conclusion of Lemma 2.4 gives continuity of $Q_0(\theta)$ and uniform convergence of $\hat{Q}_n(\theta)$ to $Q_0(\theta)$. The following result summarizes these conditions:

Theorem 2.10

If $y = 1(x'\theta_0 + \varepsilon > 0)$ and (i) the conditional distribution of ε given x has a unique median at $\varepsilon = 0$; (ii) there are corresponding partitions $x = (x_1, x_2')'$ and $\theta = (\theta_1, \theta_2')'$

[23] A median of the distribution of a random variable Y is the set of values m such that $\text{Prob}(Y \geq m) \geq \frac{1}{2}$ and $\text{Prob}(y \leq m) \geq \frac{1}{2}$.

such that $\text{Prob}(x_2'\delta \neq 0) > 0$ for $\delta \neq 0$ and the conditional distribution of x_1 given x_2 is continuous with support $\mathbb{R}$; and (iii) $x'\theta$ is continuously distributed for all $\theta \in \Theta = \{\theta: \|\theta\| = 1\}$; then $\hat{\theta} \xrightarrow{p} \theta_0$.

2.8.2. *Censored least absolute deviations*

Censored least absolute deviations is used to estimate θ_0 in the model $y = \max\{0, x'\theta_0 + \varepsilon\}$ where ε has a unique conditional median at zero. It is obtained by solving eq. (1.1) for $\hat{Q}_n(\theta) = -n^{-1}\sum_{i=1}^n (|y_i - \max\{0, x_i'\theta\}| - |y_i - \max\{0, x_i'\theta_0\}|) = \hat{Q}_n(\theta) - \hat{Q}_n(\theta_0)$. Consistency of $\hat{\theta}$ can be shown by using Lemma 2.4 to verify the conditions of Theorem 2.1. The function $|y_i - \max\{0, x_i'\theta\}| - |y_i - \max\{0, x_i'\theta_0\}|$ is continuous in θ by inspection, and by the triangle inequality its absolute value is bounded above by $|\max\{0, x_i'\theta\}| + |\max\{0, x_i'\theta_0\}| \leq \|x_i\|(\|\theta\| + \|\theta_0\|)$, so that if $E[\|x\|] < \infty$ the dominance condition is satisfied. Then by the conclusion of Lemma 2.4, $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta) = E[|y - \max\{0, x'\theta\}| - |y - \max\{0, x'\theta_0\}|]$. Thus, for the normalized objective function, uniform convergence does not require any moments of y to exist, as promised in Section 2.1.

Identification will follow from the fact that the conditional median minimizes the expected absolute deviation. Suppose that $P(x'\theta_0 > 0)$ and $P(x'\delta \neq 0 | x'\theta_0 > 0) > 0$ if $\delta \neq 0$.[24] By ε having a unique conditional median at zero, y has a unique conditional median at $\max\{0, x'\theta_0\}$. Therefore, to show identification it suffices to show that $\max\{0, x'\theta\} \neq \max\{0, x'\theta_0\}$ if $\theta \neq \theta_0$. There are two cases to consider. In case one, $1(x'\theta > 0) \neq 1(x'\theta_0 > 0)$, implying $\max\{0, x'\theta_0\} \neq \max\{0, x'\theta\}$. In case two, $1(x'\theta > 0) = 1(x'\theta_0 > 0)$, so that $\max\{0, x'\theta\} - \max\{0, x'\theta_0\} = 1(x'\theta_0 > 0)x'(\theta - \theta_0) \neq 0$ by the identifying assumption. Thus, $Q_0(\theta)$ has a unique maximum over all of $\mathbb{R}^q$ at θ_0. Summarizing these conditions leads to the following result:

Theorem 2.11

If (i) $y = \max\{0, x'\theta_0 + \varepsilon\}$, the conditional distribution of ε given x has a unique median at $\varepsilon = 0$; (ii) $\text{Prob}(x'\theta_0 > 0) > 0$, $\text{Prob}(x'\delta \neq 0 | x'\theta_0 > 0) > 0$; (iii) $E[\|x\|] < \infty$; and (iv) Θ is any compact set containing θ_0, then $\hat{\theta} \xrightarrow{p} \theta_0$.

As previously promised, this result shows that no assumption on the existence of moments of y is needed for consistency of censored least absolute deviations. Also, it shows that in spite of the first-order conditions being identically zero over all θ where $x_i'\theta < 0$ for all the observations, the global maximum of the least absolute deviations estimator, over any compact set containing the true parameter, will be consistent. It is not known whether the compactness restriction can be relaxed for this estimator; the objective function is not concave, and it is not known whether some other approach can be used to get rid of compactness.

[24] It suffices for the second condition that $E[1(x'\theta_0 > 0)xx']$ is nonsingular.

3. Asymptotic normality

Before giving precise conditions for asymptotic normality, it is helpful to sketch the main ideas. The key idea is that in large samples estimators are approximately equal to linear combinations of sample averages, so that the central limit theorem gives asymptotic normality. This idea can be illustrated by describing the approximation for the MLE. When the log-likelihood is differentiable and $\hat{\theta}$ is in the interior of the parameter set Θ, the first-order condition $0 = n^{-1}\sum_{i=1}^{n} \nabla_\theta \ln f(z_i|\hat{\theta})$ will be satisfied. Assuming twice continuous differentiability of the log-likelihood, the mean-value theorem applied to each element of the right-hand side of this first-order condition gives

$$0 = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ln f(z_i|\theta_0) + \left[\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta} \ln f(z_i|\bar{\theta})\right](\hat{\theta} - \theta_0), \tag{3.1}$$

where $\bar{\theta}$ is a mean value on the line joining $\hat{\theta}$ and θ_0 and $\nabla_{\theta\theta}$ denotes the Hessian matrix of second derivatives.[25] Let $J = E[\nabla_\theta \ln f(z|\theta_0)\{\nabla_\theta \ln f(z|\theta_0)\}']$ be the information matrix and $H = E[\nabla_{\theta\theta} \ln f(z|\theta_0)]$ the expected Hessian. Multiplying through by $\sqrt{n}$ and solving for $\sqrt{n}(\hat{\theta} - \theta_0)$ gives

$$\begin{aligned}
&\sqrt{n}(\hat{\theta} - \theta_0) \\
&= -\underbrace{\left[\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta} \ln f(z_i|\bar{\theta})\right]^{-1}}_{\downarrow\ \text{p (Hessian Conv.) (Inverse Cont.)}\ H^{-1}} \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_\theta \ln f(z_i|\theta_0)}_{\downarrow\ \text{d (CLT)}\ N(0,J)} \xrightarrow[\text{(Slutzky theorem)}]{\text{d}} -H^{-1}N(0,J).
\end{aligned} \tag{3.2}$$

By the well known zero-mean property of the score $\nabla_\theta \ln f(z|\theta_0)$ and the central limit theorem, the second term will converge in distribution to $N(0, J)$. Also, since $\bar{\theta}$ is between $\hat{\theta}$ and θ_0, it will be consistent if $\hat{\theta}$ is, so that by a law of large numbers that is uniform in θ converging to θ, the Hessian term converges in probability to H. Then the inverse Hessian converges in probability to H^{-1} by continuity of the inverse at a nonsingular matrix. It then follows from the Slutzky theorem that $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\text{d}} N(0, H^{-1}JH^{-1})$.[26] Furthermore, by the information matrix equality

[25] The mean-value theorem only applies to individual elements of the partial derivatives, so that $\bar{\theta}$ actually differs from element to element of the vector equation (3.1). Measurability of these mean values holds because they minimize the absolute value of the remainder term, setting it equal to zero, and thus are extremum estimators; see Jennrich (1969).

[26] The Slutzky theorem is $Y_n \xrightarrow{\text{d}} Y_0$ and $Z_n \xrightarrow{\text{p}} c \Rightarrow Z_n Y_n \xrightarrow{\text{d}} cY_0$.

$H = -J$, the asymptotic variance will have the usual inverse information matrix form J^{-1}.

This expansion shows that the maximum likelihood estimator is approximately equal to a linear combination of the average score in large samples, so that asymptotic normality follows by the central limit theorem applied to the score. This result is the prototype for many other asymptotic normality results. It has several components, including a first-order condition that is expanded around the truth, convergence of an inverse Hessian, and a score that follows the central limit theorem. Each of these components is important to the result. The first-order condition is a consequence of the estimator being in the interior of the parameter space.[27] If the estimator remains on the boundary asymptotically, then it may not be asymptotically normal, as further discussed below. Also, if the inverse Hessian does not converge to a constant or the average score does not satisfy a central limit theorem, then the estimator may not be asymptotically normal. An example like this is least squares estimation of an autoregressive model with a unit root, as further discussed in Chapter 2.

One condition that is not essential to asymptotic normality is the information matrix equality. If the distribution is misspecified [i.e. is not $f(z|\theta_0)$] then the MLE may still be consistent and asymptotically normal. For example, for certain exponential family densities, such as the normal, conditional mean parameters will be consistently estimated even though the likelihood is misspecified; e.g. see Gourieroux et al. (1984). However, the distribution misspecification will result in a more complicated form $H^{-1}JH^{-1}$ for the asymptotic variance. This more complicated form must be allowed for to construct a consistent asymptotic variance estimator under misspecification.

As described above, asymptotic normality results from convergence in probability of the Hessian, convergence in distribution of the average score, and the Slutzky theorem. There is another way to describe the asymptotic normality results that is often used. Consider an estimator $\hat{\theta}$, and suppose that there is a function $\psi(z)$ such that

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sum_{i=1}^{n} \psi(z_i)/\sqrt{n} + o_p(1), \quad E[\psi(z)] = 0, \quad E[\psi(z)\psi(z)'] \text{ exists}, \tag{3.3}$$

where $o_p(1)$ denotes a random vector that converges in probability to zero. Asymptotic normality of $\hat{\theta}$ then results from the central limit theorem applied to $\sum_{i=1}^{n} \psi(z_i)/\sqrt{n}$, with asymptotic variance given by the variance of $\psi(z)$. An estimator satisfying this equation is referred to as *asymptotically linear*. The function $\psi(z)$ is referred to as the *influence function*, motivated by the fact that it gives the effect of a single

[27] It is sufficient that the estimator be in the "relative interior" of Θ, allowing for equality restrictions to be imposed on θ, such as $\theta = \tau(\gamma)$ for smooth $\tau(\gamma)$ and the true γ being in an open ball. The first-order condition does rule out inequality restrictions that are asymptotically binding.

observation on the estimator, up to the $o_p(1)$ remainder term. This description is useful because all the information about the asymptotic variance is summarized in the influence function. Also, the influence function is important in determining the robustness properties of the estimator; e.g. see Huber (1964).

The MLE is an example of an asymptotically linear estimator, with influence function $\psi(z) = -H^{-1}\nabla_\theta \ln f(z|\theta_0)$. In this example the remainder term is, for the mean value $\bar{\theta}$, $-[(n^{-1}\sum_{i=1}^n \nabla_{\theta\theta} \ln f(z_i|\bar{\theta}))^{-1} - H^{-1}]n^{-1/2}\sum_{i=1}^n \nabla_0 \ln f(z_i|\theta_0)$, which converges in probability to zero because the inverse Hessian converges in probability to H and the $\sqrt{n}$ times the average score converges in distribution. Each of NLS and GMM is also asymptotically linear, with influence functions that will be described below. In general the CMD estimator need not be asymptotically linear, because its asymptotic properties depend only on the reduced form estimator $\hat{\pi}$. However, if the reduced form estimator $\hat{\pi}$ is asymptotically linear the CMD will also be.

The idea of approximating an estimator by a sample average and applying the central limit theorem can be used to state rigorous asymptotic normality results for extremum estimators. In Section 3.1 precise results are given for cases where the objective function is "sufficiently smooth", allowing a Taylor expansion like that of eq. (3.1). Asymptotic normality for nonsmooth objective functions is discussed in Section 7.

3.1. *The basic results*

For asymptotic normality, two basic results are useful, one for an extremum estimator and one for a minimum distance estimator. The relationship between these results will be discussed below. The first theorem is for an extremum estimator.

Theorem 3.1

Suppose that $\hat{\theta}$ satisfies eq. (1.1), $\hat{\theta} \xrightarrow{p} \theta_0$, and (i) $\theta_0 \in \text{interior}(\Theta)$; (ii) $\hat{Q}_n(\theta)$ is twice continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0; (iii) $\sqrt{n}\nabla_0 \hat{Q}_n(\theta_0) \xrightarrow{d} N(0, \Sigma)$; (iv) there is $H(\theta)$ that is continuous at θ_0 and $\sup_{\theta \in \mathcal{N}} \| \nabla_{\theta\theta} \hat{Q}_n(\theta) - H(\theta) \| \xrightarrow{p} 0$; (v) $H = H(\theta_0)$ is nonsingular. Then $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1}\Sigma H^{-1})$.

Proof

A sketch of a proof is given here, with full details described in Section 3.5. Conditions (i)–(iii) imply that $\nabla_\theta \hat{Q}_n(\hat{\theta}) = 0$ with probability approaching one. Expanding around θ_0 and solving for $\sqrt{n}(\hat{\theta} - \theta_0) = -\hat{H}(\bar{\theta})^{-1}\sqrt{n}\nabla_\theta \hat{Q}_n(\theta_0)$, where $\hat{H}(\theta) = \nabla_{\theta\theta}\hat{Q}_n(\theta)$ and $\bar{\theta}$ is a mean value, located between $\hat{\theta}$ and θ_0. By $\bar{\theta} \xrightarrow{p} \theta_0$ and (iv), with probability approaching one, $\| \hat{H}(\bar{\theta}) - H \| \leqslant \| \hat{H}(\bar{\theta}) - H(\bar{\theta}) \| + \| H(\bar{\theta}) - H \| \leqslant \sup_{\theta \in \Theta} \| \hat{H}(\theta) - H(\theta) \| + \| H(\bar{\theta}) - H \| \xrightarrow{p} 0$. Then by continuity of matrix inversion, $-\hat{H}(\bar{\theta})^{-1} \xrightarrow{p} -H^{-1}$. The conclusion then follows by the Slutzky theorem. Q.E.D.

The asymptotic variance matrix in the conclusion of this result has a complicated form, being equal to the product $H^{-1}\Sigma H^{-1}$. In the case of maximum likelihood this form simplifies to J^{-1}, the inverse of the information matrix, because of the information matrix equality. An analogous simplification occurs for some other estimators, such as NLS where $\mathrm{Var}(y|x)$ is constant (i.e. under homoskedasticity). As further discussed in Section 5, a simplified asymptotic variance matrix is a feature of an efficient estimator in some class.

The true parameter being interior to the parameter set, condition (i), is essential to asymptotic normality. If Θ imposes inequality restrictions on θ that are asymptotically binding, then the estimator may not be asymptotically normal. For example, consider estimation of the mean of a normal distribution that is constrained to be nonnegative, i.e. $f(z|\theta) = (2\pi\sigma^2)^{-1}\exp[-(z-\mu)^2/2\sigma^2]$, $\theta = (\mu, \sigma^2)$, and $\Theta = [0, \infty) \times (0, \infty)$. It is straightforward to check that the MLE of μ is $\hat{\mu} = \bar{z}, \bar{z} > 0$, $\hat{\mu} = 0$ otherwise. If $\mu_0 = 0$, violating condition (ii), then $\mathrm{Prob}(\hat{\mu} = 0) = \frac{1}{2}$ and $\sqrt{n}\hat{\mu}$ is $N(0, \sigma^2)$ conditional on $\hat{\mu} > 0$. Therefore, for every n (and hence also asymptotically), the distribution of $\sqrt{n}(\hat{\mu} - \mu_0)$ is a mixture of a spike at zero with probability $\frac{1}{2}$ and the positive half normal distribution. Thus, the conclusion of Theorem 3.1 is not true. This example illustrates that asymptotic normality can fail when the maximum occurs on the boundary. The general theory for the boundary case is quite complicated, and an account will not be given in this chapter.

Condition (ii), on twice differentiability of $Q_n(\theta)$, can be considerably weakened without affecting the result. In particular, for GMM and CMD, asymptotic normality can easily be shown when the moment functions only have first derivatives. With considerably more work, it is possible to obtain asymptotic normality when $\hat{Q}_n(\theta)$ is not even once differentiable, as discussed in Section 7.

Condition (iii) is analogous to asymptotic normality of the scores. It will often follow from a central limit theorem for the sample averages that make up $\nabla_\theta \hat{Q}_n(\theta_0)$.

Condition (iv) is uniform convergence of the Hessian over a neighborhood of the true parameter and continuity of the limiting function. This same type of condition (on the objective function) is important for consistency of the estimator, and was discussed in Section 2. Consequently, the results of Section 2 can be applied to give primitive hypotheses for condition (iv). In particular, when the Hessian is a sample average, or depends on sample averages, Lemma 2.4 can be applied. If the average is continuous in the parameters, as will typically be implied by condition (iv), and a dominance condition is satisfied, then the conclusion of Lemma 2.4 will give uniform convergence. Using Lemma 2.4 in this way will be illustrated for MLE and GMM.

Condition (v) can be interpreted as a strict local identification condition, because $H = \nabla_{\theta\theta} Q_0(\theta_0)$ (under regularity conditions that allow interchange of the limiting and differentiation operations.) Thus, nonsingularity of H is the sufficient (second-order) condition for there to be a unique local maximum at θ_0. Furthermore, if $\nabla_{\theta\theta} Q_0(\theta)$ is "regular", in the sense of Rothenberg (1971) that it has constant rank in a neighborhood of θ_0, then nonsingularity of H follows from $Q_0(\theta)$ having a unique

maximum at θ_0. A local identification condition in these cases is that H is nonsingular.

As stated above, asymptotic normality of GMM and CMD can be shown under once differentiability, rather than twice differentiability. The following asymptotic normality result for general minimum distance estimators is useful for this purpose.

Theorem 3.2

Suppose that $\hat{\theta}$ satisfies eq. (1.1) for $\hat{Q}_n(\theta) = -\hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta)$ where $\hat{W} \xrightarrow{p} W$, W is positive semi-definite, $\hat{\theta} \xrightarrow{p} \theta_0$, and (i) $\theta_0 \in \text{interior}(\Theta)$; (ii) $\hat{g}_n(\theta)$ is continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0; (iii) $\sqrt{n}\hat{g}_n(\theta_0) \xrightarrow{d} N(0, \Omega)$; (iv) there is $G(\theta)$ that is continuous at θ_0 and $\sup_{\theta \in \mathcal{N}} \| \nabla_\theta \hat{g}_n(\theta) - G(\theta) \| \xrightarrow{p} 0$; (v) for $G = G(\theta_0)$, $G'WG$ is nonsingular. Then $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N[0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}]$.

Proof

The argument is similar to the proof of Theorem 3.1. By (i) and (ii), with probability approaching one the first-order conditions $\hat{G}(\hat{\theta})'\hat{W}\hat{g}_n(\hat{\theta}) = 0$ are satisfied, for $\hat{G}(\theta) = \nabla_\theta \hat{g}_n(\theta)$. Expanding $\hat{g}_n(\hat{\theta})$ around θ_0 and solving gives $\sqrt{n}(\hat{\theta} - \theta_0) = -[\hat{G}(\hat{\theta})' \times \hat{W}\hat{G}(\bar{\theta})]^{-1}\hat{G}(\hat{\theta})'\hat{W}\sqrt{n}\hat{g}_n(\theta_0)$, where $\bar{\theta}$ is a mean value. By (iv) and similar reasoning as for Theorem 3.1, $\hat{G}(\hat{\theta}) \xrightarrow{p} G$ and $\hat{G}(\bar{\theta}) \xrightarrow{p} G$. Then by (v), $-[\hat{G}(\hat{\theta})'\hat{W}\hat{G}(\bar{\theta})]^{-1}\hat{G}(\hat{\theta})'\hat{W} \xrightarrow{p} -(G'WG)^{-1}G'W$, so the conclusion follows by (iii) and the Slutzky theorem.

Q.E.D.

When $W = \Omega^{-1}$, the asymptotic variance of a minimum distance estimator simplifies to $(G'\Omega^{-1}G)^{-1}$. As is discussed in Section 5, the value $W = \Omega^{-1}$ corresponds to an efficient weighting matrix, so as for the MLE the simpler asymptotic variance matrix is associated with an efficient estimator.

Conditions (i)–(v) of Theorem 3.2 are analogous to the corresponding conditions of Theorem 3.1, and most of the discussion given there also applies in the minimum distance case. In particular, the differentiability condition for $\hat{g}_n(\theta)$ can be weakened, as discussed in Section 7.

For analyzing asymptotic normality, extremum estimators can be thought of as a special case of minimum distance estimators, with $\nabla_\theta \hat{Q}_n(\theta) = \hat{g}_n(\theta)$ and $\hat{W} = I = W$. The first-order conditions for extremum estimators imply that $\hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta) = \nabla_\theta \hat{Q}_n(\theta)'\nabla_\theta \hat{Q}_n(\theta)$ has a minimum (of zero) at $\theta = \hat{\theta}$. Then the G and Ω of Theorem 3.2 are the H and Σ of Theorem 3.1, respectively, and the asymptotic variance of the extremum estimator is that of the minimum distance estimator, with $(G'WG)^{-1} \times G'W\Omega WG(G'WG)^{-1} = (H'H)^{-1}H'\Sigma H(H'H)^{-1} = H^{-1}\Sigma H^{-1}$. Thus, minimum distance estimation provides a general framework for analyzing asymptotic normality, although, as previously discussed, it is better to work directly with the maximum, rather than the first-order conditions, when analyzing consistency.[28]

[28] This generality suggests that Theorem 3.1 could be formulated as a special case of Theorem 3.2. The results are not organized in this way because it seems easier to apply Theorem 3.1 directly to particular extremum estimators.

3.2. *Asymptotic normality for MLE*

The conditions for asymptotic normality of an extremum estimator can be specialized to give a result for MLE.

Theorem 3.3

Suppose that $z_1, \ldots, z_n$ are i.i.d., the hypotheses of Theorem 2.5 are satisfied and (i) $\theta_0 \in \text{interior}(\Theta)$; (ii) $f(z|\theta)$ is twice continuously differentiable and $f(z|\theta) > 0$ in a neighborhood $\mathcal{N}$ of θ_0; (iii) $\int \sup_{\theta \in \mathcal{N}} \|\nabla_\theta f(z|\theta)\| \, dz < \infty$, $\int \sup_{\theta \in \mathcal{N}} \|\nabla_{\theta\theta} f(z|\theta)\| \, dz < \infty$; (iv) $J = E[\nabla_\theta \ln f(z|\theta_0)\{\nabla_\theta \ln f(z|\theta_0)\}']$ exists and is nonsingular; (v) $E[\sup_{\theta \in \mathcal{N}} \|\nabla_{\theta\theta} \times \ln f(z|\theta)\|] < \infty$. Then $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, J^{-1})$.

Proof

The proof proceeds by verifying the hypotheses of Theorem 3.1. By Theorem 2.5, $\hat{\theta} \xrightarrow{p} \theta_0$. Important intermediate results are that the score $s(z) = \nabla_\theta \ln f(z|\theta_0)$ has mean zero and the information matrix equality $J = -E[\nabla_{\theta\theta} \ln f(z|\theta_0)]$. These results follow by differentiating the identity $\int f(z|\theta) dz$ twice, and interchanging the order of differentiation and integration, as allowed by (iii) and Lemma 3.6 in Section 3.5. Then conditions 3.1(i), (ii) hold by 3.3(i), (ii). Also, 3.1(iii) holds, with $\Sigma = J$, by $E[s(z)] = 0$, existence of J, and the Lindberg–Levy central limit theorem. To show 3.1(iv) with $H = -J$, let Θ be a compact set contained in $\mathcal{N}$ and containing θ_0 in its interior, so that the hypotheses of Lemma 2.4 are satisfied for $a(z, \theta) = \nabla_{\theta\theta} \ln f(z|\theta)$ by (ii) and (v). Condition 3.1(v) then follows by nonsingularity of J. Now $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1} J H^{-1}) = N(0, J^{-1})$ follows by the conclusion of Theorem 3.1 and $H = -J$. Q.E.D.

The hypotheses of Theorem 2.5 are only used to make sure that $\hat{\theta} \xrightarrow{p} \theta_0$, so that they can be replaced by any other conditions that imply consistency. For example, the conditions that θ_0 is identified, $\ln f(z|\theta)$ is concave in θ, and $E[|\ln f(z|\theta)|] < \infty$ for all θ can be used as replacements for Theorem 2.5, because Theorem 2.7 then gives $\hat{\theta} \xrightarrow{p} \theta_0$. More generally, the MLE will be asymptotically normal if it is consistent and the other conditions (i)–(v) of Theorem 3.3 are satisfied.

It is straightforward to derive a corresponding result for nonlinear least squares, by using Lemma 2.4, the law of large numbers, and the Lindberg–Levy central limit theorem to provide primitive conditions for Theorem 3.1. The statement of a theorem is left as an exercise for the interested reader. The resulting asymptotic variance for NLS will be $H^{-1} \Sigma H^{-1}$, for $E[y|x] = h(x, \theta_0)$, $h_\theta(x, \theta) = \nabla_\theta h(x, \theta)$, $H = -E[h_\theta(x, \theta_0) h_\theta(x, \theta_0)']$ and $\Sigma = E[\{y - h(x, \theta_0)\}^2 h_\theta(x, \theta_0) h_\theta(x, \theta_0)']$. The variance matrix simplifies to $\sigma^2 H^{-1}$ when $E[\{y - h(x, \theta_0)\}^2 | x]$ is a constant σ^2, a well known efficiency condition for NLS.

As previously stated, MLE and NLS will be asymptotically linear, with the MLE influence function given by $J^{-1}\nabla_\theta \ln f(z|\theta_0)$. The NLS influence function will have a similar form,

$$\psi(z) = \{E[h_\theta(x,\theta_0)h_\theta(x,\theta_0)']\}^{-1}h_\theta(x,\theta_0)[y - h(x,\theta_0)], \tag{3.4}$$

as can be shown by expanding the first-order conditions for NLS.

The previous examples provide useful illustrations of how the regularity conditions can be verified.

Example 1.1 continued

In the Cauchy location and scale case, $f(z|\theta) = \sigma^{-1}g[\sigma^{-1}(z-\mu)]$ for $g(\varepsilon) = 1/[\pi(1+\varepsilon^2)]$. To show asymptotic normality of the MLE, the conditions of Theorem 3.3 can be verified. The hypotheses of Theorem 2.5 were shown in Section 2. For the parameter set previously specified for this example, condition (i) requires that μ_0 and σ_0 are interior points of the allowed intervals. Condition (ii) holds by inspection. It is straightforward to verify the dominance conditions for (iii) and (v). For example, (v) follows by noting that $\nabla_{\theta\theta}\ln f(z|\theta)$ is bounded, uniformly in bounded μ and σ, and σ bounded away from zero. To show condition (iv), consider $\alpha = (\alpha_1,\alpha_2)' \neq 0$. Note that $\sigma_0(1+z^2)[\alpha'\nabla_\theta \ln f(z|\theta_0)] = \alpha_1 2z + \alpha_2(1+z^2) + \alpha_2 2z^2 = \alpha_2 + (2\alpha_1)z + (3\alpha_2)z^2$ is a polynomial and hence is nonzero on an interval. Therefore, $E[\{\alpha'\nabla_\theta \ln f(z|\theta_0)\}^2] = \alpha' J\alpha > 0$. Since this conclusion is true for any $\alpha \neq 0$, J must be nonsingular.

Example 1.2 continued

Existence and nonsingularity of $E[xx']$ are sufficient for asymptotic normality of the probit MLE. Consistency of $\hat{\theta}$ was shown in Section 2.6, so that only conditions (i)–(v) of Theorem 3.3 are needed (as noted following Theorem 3.3). Condition (i) holds because $\Theta = \mathbb{R}^q$ is an open set. Condition (ii) holds by inspection of $f(z|\theta) = y\Phi(x'\theta) + (1-y)\Phi(-x'\theta)$. For condition (iii), it is well known that $\phi(v)$ and $\phi_v(v)$ are uniformly bounded, implying $\nabla_\theta f(z|\theta) = (1-2y)\phi(x'\theta)x$ and $\nabla_{\theta\theta} f(z|\theta) = (1-2y)\times \phi_v(x'\theta)xx'$ are bounded by $C(1+\|x\|^2)$ for some constant C. Also, integration over dz is the sum over y and the expectation over x {i.e. $\int a(y,x)dz = E[a(0,x) + a(1,x)]$}, so that $\int(1+\|x\|^2)dz = 2 + 2E[\|x\|^2] < \infty$. For (iv), it can be shown that $J = E[\lambda(x'\theta_0)\lambda(-x'\theta_0)xx']$, for $\lambda(v) = \phi(v)/\Phi(v)$. Existence of J follows by $\lambda(v)\lambda(-v)$ bounded, and nonsingularity by $\lambda(v)\lambda(-v)$ bounded away from zero on any open interval.[29] Condition (v) follows from $\nabla_{\theta\theta}\ln f(z|\theta_0) = [\lambda_v(x'\theta_0)y + \lambda_v(-x'\theta_0)(1-y)]xx'$

[29] It can be shown that $\lambda(v)\lambda(-v)$ is bounded using l'Hôpital's rule. Also, for any $\bar{v} > 0$, $J \geq E[1(|x'\theta_0| \leq \bar{v})\lambda(x'\theta_0)\lambda(-x'\theta_0)xx'] \geq CE[1(|x'\theta_0| \leq \bar{v})xx']$ in the positive semi-definite sense, the last term is positive definite for large enough $\bar{v}$ by nonsingularity of $E[xx']$.

and boundedness of $\lambda_v(v)$. This example illustrates how conditions on existence of moments may be useful regularity conditions for consistency and asymptotic normality of an MLE, and how detailed work may be needed to check the conditions.

3.3. *Asymptotic normality for GMM*

The conditions on asymptotic normality of minimum distance estimators can be specialized to give a result for GMM.

Theorem 3.4

Suppose that the hypotheses of Theorem 2.6 are satisfied, $\hat{W} \xrightarrow{p} W$, and (i) $\theta_0 \in$ interior of Θ; (ii) $g(z,\theta)$ is continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0, with probability approaching one; (iii) $E[g(z,\theta_0)] = 0$ and $E[\|g(z,\theta_0)\|^2]$ is finite; (iv) $E[\sup_{\theta\in\mathcal{N}} \|\nabla_\theta g(z,\theta)\|] < \infty$; (v) $G'WG$ is nonsingular for $G = E[\nabla_\theta g(z,\theta_0)]$. Then for $\Omega = E[g(z,\theta_0)g(z,\theta_0)']$, $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N[0, (G'WG)G'W\Omega WG(G'WG)^{-1}]$.

Proof

The proof will be sketched, although a complete proof like that of Theorem 3.1 given in Section 3.5 could be given. By (i), (ii), and (iii), the first-order condition $2\hat{G}_n(\hat{\theta})'\hat{W}\hat{g}_n(\hat{\theta}) = 0$ is satisfied with probability approaching one, for $\hat{G}_n(\theta) = \nabla_\theta \hat{g}_n(\theta)$. Expanding $\hat{g}_n(\hat{\theta})$ around θ_0, multiplying through by $\sqrt{n}$, and solving gives

$$\sqrt{n}(\hat{\theta} - \theta_0) = -[\hat{G}_n(\hat{\theta})'\hat{W}\hat{G}_n(\bar{\theta})]^{-1}\hat{G}_n(\hat{\theta})'\hat{W}\sqrt{n}g_n(\theta_0), \tag{3.5}$$

where $\bar{\theta}$ is the mean value. By (iv), $\hat{G}_n(\hat{\theta}) \xrightarrow{p} G$ and $\hat{G}_n(\bar{\theta}) \xrightarrow{p} G$, so that by (v), $[\hat{G}_n(\hat{\theta})'\hat{W}\hat{G}_n(\bar{\theta})]^{-1}\hat{G}_n(\hat{\theta})'\hat{W} \xrightarrow{p} (G'WG)^{-1}G'W$. The conclusion then follows by the Slutzky theorem. Q.E.D.

The complicated asymptotic variance formula simplifies to $(G'\Omega^{-1}G)^{-1}$ when $W = \Omega^{-1}$. As shown in Hansen (1982) and further discussed in Section 5, this value for W is optimal in the sense that it minimizes the asymptotic variance matrix of the GMM estimator.

The hypotheses of Theorem 2.6 are only used to make sure that $\hat{\theta} \xrightarrow{p} \theta_0$, so that they can be replaced by any other conditions that imply consistency. For example, the conditions that θ_0 is identified, $g(z,\theta)$ is linear in θ, and $E[\|g(z,\theta)\|] < \infty$ for all θ can be used as replacements for Theorem 2.6, because Theorem 2.7 then gives $\hat{\theta} \xrightarrow{p} \theta_0$. More generally, a GMM estimator will be asymptotically normal if it is consistent and the other conditions (i)–(v) of Theorem 3.4 are satisfied.

It is straightforward to derive a corresponding result for classical minimum distance, under the conditions that $\hat{\theta}$ is consistent, $\sqrt{n}[\hat{\pi} - h(\theta_0)] \xrightarrow{d} N(0, \Omega)$ for some Ω, $h(\theta)$ is continuously differentiable in a neighborhood of θ_0, and $G'WG$ is nonsingular for $G = \nabla_\theta h(\theta_0)$. The statement of a theorem is left as an exercise for the interested reader. The resulting asymptotic variance for CMD will have the same form as given in the conclusion of Theorem 3.4.

By expanding the GMM first-order conditions, as in eq. (3.5), it is straightforward to show that GMM is asymptotically linear with influence function

$$\psi(z) = -(G'WG)^{-1}G'Wg(z, \theta_0). \tag{3.6}$$

In general CMD need not be asymptotically linear, but will be if the reduced form estimator $\hat{\pi}$ is asymptotically linear. Expanding the first-order conditions for $\hat{\theta}$ around the truth gives $\sqrt{n}(\hat{\theta} - \theta_0) = -(\hat{G}'\hat{W}\bar{G})^{-1}\hat{G}'\hat{W}\sqrt{n}(\hat{\pi} - \pi_0)$, where $\hat{G} = \nabla_\theta h(\hat{\theta})$, $\bar{G} = \nabla_\theta h(\bar{\theta})$, and $\bar{\theta}$ is the mean value. Then $\sqrt{n}(\hat{\pi} - \pi_0)$ converging in distribution and $(\hat{G}'\hat{W}\bar{G})^{-1}\hat{G}'\hat{W} \xrightarrow{p} (G'WG)^{-1}G'W$ implies that $\sqrt{n}(\hat{\theta} - \theta_0) = -(G'WG)^{-1}G' \times W\sqrt{n}(\hat{\pi} - \pi_0) + o_p(1)$. Therefore, if $\hat{\pi}$ is asymptotically linear with influence function $\psi^\pi(z)$, the CMD estimator will also be asymptotically linear with influence function

$$\psi(z) = -(G'WG)^{-1}G'W\psi^\pi(z). \tag{3.7}$$

The Hansen–Singleton example provides a useful illustration of how the conditions of Theorem 3.4 can be verified.

Example 1.3 continued

It was shown in Section 2 that sufficient conditions for consistency are that $E[x(\beta w y^\gamma - 1)] = 0$ have a unique solution at $\theta_0 \in \Theta = [\beta_\ell, \beta_u] \times [\gamma_\ell, \gamma_u]$, and that $E[\|x\|] < \infty$ and $E[\|x\| |w|(|y|^{\gamma_\ell} + |y|^{\gamma_u})] < \infty$. To obtain asymptotic normality, impose the additional conditions that $\theta_0 \in \text{interior}(\Theta)$, $\gamma_\ell < 0$, $E[\|x\|^2] < \infty$, $E[\|x\|^2 |w|^2 y^{2\gamma_0}] < \infty$, and $E[x(wy^{\gamma_0}, w \cdot \ln(y) y^{\gamma_0})]$ has rank 2. Then condition (i) of Theorem 3.4 is satisfied by assumption. Condition (ii) is also satisfied, with $\nabla_\theta g(z, \theta) = x(wy^\gamma, w \cdot \ln(y) y^\gamma)$. Condition (iii) is satisfied by the additional, second-moment restrictions, and by the GMM identification hypothesis.

To check condition (iv), note that $|\ln(y)|$ is bounded above by $C(|y|^{-\varepsilon} + |y|^\varepsilon)$ for any $\varepsilon > 0$ and constant C big enough. Let $\mathcal{N}$ be a neighborhood of θ_0 such that $\gamma_\ell + \varepsilon < \gamma < \gamma_u - \varepsilon$ for all $\theta \in \mathcal{N}$. Then $\sup_{\theta \in \mathcal{N}} \|\nabla_\theta g(z, \theta)\| \leq C\|x\| |w| [1 + \ln(y)] \times \sup_{\mathcal{N}} |y|^\gamma \leq C\|x\| |w| (1 + |y|^{-\varepsilon} + |y|^\varepsilon) \sup_{\mathcal{N}} |y|^\gamma \leq \|x\| |w| (|y|^{\gamma_\ell} + |y|^{\gamma_u})$, so that condition (iv) follows by the previously assumed moment condition. Finally, condition (v) holds by the previous rank condition and $W = (E[xx'])^{-1}$ nonsingular. Thus, under the assumptions imposed above, the nonlinear two-stage least squares estimator will be consistent and asymptotically normal, with asymptotic variance as given in the conclusion of Theorem 3.4.

3.4. One-step theorems

A result that is useful, particularly for efficient estimation, pertains to the properties of estimators that are obtained from a single iteration of a numerical maximization procedure, such as Newton–Raphson. If the starting point is an estimator that is asymptotically normal, then the estimator from applying one iteration will have the same asymptotic variance as the maximum of an objective function. This result is particularly helpful when simple initial estimators can be constructed, but an efficient estimator is more complicated, because it means that a single iteration will yield an efficient estimator.

To describe a one-step extremum estimator, let $\bar{\theta}$ be an initial estimator and $\bar{H}$ be an estimator of $H = \operatorname{plim}[\nabla_{\theta\theta}\hat{Q}_n(\theta_0)]$. Consider the estimator

$$\tilde{\theta} = \bar{\theta} - \bar{H}^{-1}\nabla_\theta \hat{Q}_n(\bar{\theta}). \tag{3.8}$$

If $\bar{H} = \nabla_{\theta\theta}\hat{Q}_n(\bar{\theta})$ then eq. (3.8) describes one Newton–Raphson iteration. More generally it might be described as a modified Newton–Raphson step with some other value of $\bar{H}$ used in place of the Hessian. The useful property of this estimator is that it will have the same asymptotic variance as the maximizer of $\hat{Q}_n(\theta)$, if $\sqrt{n}(\bar{\theta} - \theta_0)$ is bounded in probability. Consequently, if the extremum estimator is efficient in some class, so will be the one-step estimator, while the one-step estimator is computationally more convenient than the extremum estimator.[30]

An important example is the MLE. In this case the Hessian limit is the negative of the information matrix, so that $\bar{H} = -\bar{J}$ is an estimated Hessian. The corresponding iteration is

$$\tilde{\theta} = \bar{\theta} + \bar{J}^{-1}n^{-1}\sum_{i=1}^{n}\nabla_\theta \ln f(z_i|\bar{\theta}). \tag{3.9}$$

For the Hessian estimator of the information matrix $\bar{J} = -n^{-1}\sum_{i=1}^{n}\nabla_{\theta\theta}\ln f(z_i|\bar{\theta})$, eq. (3.9) is one Newton–Raphson iteration. One could also use one of the other information matrix estimators discussed in Section 4. This is a general form of the famous linearized maximum likelihood estimator. It will have the same asymptotic variance as MLE, and hence inherit the asymptotic efficiency of the MLE.

For minimum distance estimators it is convenient to use a version that does not involve second derivatives of the moments. For $\bar{G} = \nabla_\theta \hat{g}_n(\bar{\theta})$, the matrix $-2\bar{G}'\hat{W}\bar{G}$ is an estimator of the Hessian of the objective function $-\hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta)$ at the true parameter value, because the terms that involve the second derivatives of $\hat{g}_n(\theta)$ are asymptotically negligible.[31] Plugging $\bar{H} = -2\bar{G}'\hat{W}\bar{G}$ into eq. (3.8) gives a one-step

[30] An alternative one-step estimator can be obtained by maximizing over the step size, rather than setting it equal to one, as $\tilde{\theta} = \bar{\theta} + \tilde{\lambda}\hat{d}$ for $\hat{d} = -\bar{H}^{-1}\nabla_0\hat{Q}_n(\bar{\theta})$ and $\tilde{\lambda} = \operatorname{argmax}_\lambda \hat{Q}_n(\bar{\theta} + \lambda\hat{d})$. This estimator will also have the same asymptotic variance as the solution to eq. (1.1), as shown by Newey (1987).

[31] These terms are all multiplied by one or more elements of $\hat{g}_n(\theta_0)$, which all converge to zero.

minimum distance estimator,

$$\tilde{\theta} = \bar{\theta} - (\bar{G}'\hat{W}\bar{G})^{-1}\bar{G}'\hat{W}\hat{g}_n(\bar{\theta}). \tag{3.10}$$

Alternatively, one could replace $\bar{G}$ by any consistent estimator of $\operatorname{plim}[\nabla_\theta \hat{g}_n(\theta_0)]$. This estimator will have the same asymptotic variance as a minimum distance estimator with weighting matrix $\hat{W}$. In particular, if $\hat{W}$ is a consistent estimator of Ω^{-1}, an efficient choice of weighting matrix, then $\tilde{\theta}$ has the same asymptotic variance as the minimum distance estimator with an efficient weighting matrix.

An example is provided by GMM estimation. Let $\bar{G} = n^{-1}\sum_{i=1}^n \nabla_\theta g(z_i, \bar{\theta})$ and let $\bar{\Omega}$ be an estimator of $\Omega = E[g(z, \theta_0)g(z, \theta_0)']$, such as $\bar{\Omega} = n^{-1}\sum_{i=1}^n g(z_i, \bar{\theta})g(z_i, \bar{\theta})'$. Then the one-step estimator of eq. (3.10) is

$$\tilde{\theta} = \bar{\theta} - (\bar{G}'\bar{\Omega}^{-1}\bar{G})^{-1}\bar{G}'\bar{\Omega}^{-1}\sum_{i=1}^n g(z_i, \bar{\theta})/n. \tag{3.11}$$

This is a one-step GMM estimator with efficient choice of weighting matrix.

The results showing that the one-step estimators have the same asymptotic variances as the maximizing values are quite similar for both extremum and minimum distance estimators, so it is convenient to group them together in the following result:

Theorem 3.5

Suppose that $\sqrt{n}(\bar{\theta} - \theta_0)$ is bounded in probability. If $\tilde{\theta}$ satisfies eq. (3.8), the conditions of Theorem 3.1 are satisfied, and either $\bar{H} = \nabla_{\theta\theta}\hat{Q}_n(\bar{\theta})$ or $\bar{H} \xrightarrow{p} H$, then $\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1}\Sigma H^{-1})$. If $\tilde{\theta}$ satisfies eq. (3.10), the conditions of Theorem 3.2 are satisfied, and either $\bar{G} = \nabla_\theta \hat{g}_n(\bar{\theta})$ or $\bar{G} \xrightarrow{p} G$, then $\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} N[0, (G'WG)^{-1} \times G'W\Omega WG(G'WG)^{-1}]$.

Proof

Using eq. (3.8) and expanding $\nabla_\theta \hat{Q}_n(\bar{\theta})$ around θ_0 gives:

$$\sqrt{n}(\tilde{\theta} - \theta_0) = [I - \bar{H}^{-1}\nabla_{\theta\theta}\hat{Q}_n(\dot{\theta})]\sqrt{n}(\bar{\theta} - \theta_0) - \bar{H}^{-1}\sqrt{n}\nabla_\theta \hat{Q}_n(\theta_0),$$

where $\dot{\theta}$ is the mean value. By $\bar{H}^{-1} \xrightarrow{p} H^{-1}$ and the Slutzky theorem, the second term converges in distribution to $N(0, H^{-1}\Sigma H^{-1})$. By condition (iv) of Theorem 3.1, $\bar{H}^{-1}\nabla_{\theta\theta}\hat{Q}_n(\dot{\theta}) \xrightarrow{p} H^{-1}H = I$, so that the first term is a product of a term that converges in probability to zero with a term that is bounded in probability, so that the first term converges in probability to zero, giving the conclusion. The result for minimum distance follows by a similar argument applied to the expansion of eq. (3.10)

given by $\sqrt{n}(\tilde{\theta}-\theta_0)=[I-(\bar{G}'\hat{W}\bar{G})^{-1}\bar{G}'\hat{W}\nabla_\theta\hat{g}_n(\dot{\theta})]\sqrt{n}(\bar{\theta}-\theta_0)-(\bar{G}'\hat{W}\bar{G})^{-1}\bar{G}'\hat{W}\cdot\sqrt{n}\hat{g}_n(\theta_0)$. Q.E.D.

This result can be specialized to MLE or GMM by imposing the conditions of Theorem 3.3 or 3.4, but for brevity this specialization is not given here.

The proof of this result could be modified to give the slightly stronger conclusion that $\sqrt{n}(\tilde{\theta}-\hat{\theta})\xrightarrow{p}0$, a condition that is referred to as "asymptotic equivalence" of the estimators $\tilde{\theta}$ and $\hat{\theta}$. Rothenberg (1984) showed that for MLE, if a second iteration is undertaken, i.e. $\bar{\theta}$ in eq. (3.8) solves the same equation for some other initial estimator, then $n(\tilde{\theta}-\hat{\theta})\xrightarrow{p}0$. Thus, a second iteration makes the estimator asymptotically closer to the extremum estimator. This result has been extended to multiple iterations and other types of estimators in Robinson (1988a).

3.5. *Technicalities*

A complete proof of Theorem 3.1

Without loss of generality, assume that $\mathcal{N}$ is a convex, open set contained in Θ. Let $\hat{1}$ be the indicator function for the event that $\hat{\theta}\in\mathcal{N}$. Note that $\hat{\theta}\xrightarrow{p}\theta_0$ implies $\hat{1}\xrightarrow{p}1$. By condition (ii) and the first-order conditions for a maximum, $\hat{1}\cdot\nabla_\theta\hat{Q}_n(\hat{\theta})=0$. Also, by a mean-value expansion theorem, $0=\hat{1}\cdot\nabla_\theta\hat{Q}_n(\theta_0)_j+\hat{1}\cdot\nabla^2_\theta\hat{Q}_n(\bar{\theta}_j)'_j(\hat{\theta}-\theta_0)$, where $\bar{\theta}_j$ is a random variable equal to the mean value when $\hat{1}=1$ and equal to θ_0 otherwise. Then $\bar{\theta}_j\xrightarrow{p}\theta_0$. Let $\bar{H}$ denote the matrix with jth row $\nabla^2_\theta\hat{Q}_n(\bar{\theta}_j)'_j$. By condition (iv), $\bar{H}\xrightarrow{p}H$. Let $\bar{1}$ be the indicator for $\hat{\theta}\in\mathcal{N}$ and $\bar{H}$ nonsingular. Then by condition (v), $\bar{1}\xrightarrow{p}1$, and $0=\bar{1}\cdot\nabla_\theta\hat{Q}_n(\theta_0)+\bar{1}\cdot\bar{H}(\hat{\theta}-\theta_0)$, so that $\sqrt{n}(\hat{\theta}-\theta_0)=\bar{1}\bar{H}^{-1}\sqrt{n}\nabla_\theta\hat{Q}_n(\theta_0)+(1-\bar{1})\sqrt{n}(\hat{\theta}-\theta_0)$. Then since $\bar{1}\bar{H}^{-1}\xrightarrow{p}H^{-1}$ by condition (v), $\sqrt{n}\nabla_\theta\hat{Q}_n(\theta_0)\xrightarrow{d}N(0,\Sigma)$ by condition (iii), and $(1-\bar{1})\sqrt{n}(\hat{\theta}-\theta_0)\xrightarrow{p}0$ by $\bar{1}\xrightarrow{p}1$, the conclusion follows by the Slutzky theorem and the fact that if $Y_n\xrightarrow{d}Y_0$ and $Z_n-Y_n\xrightarrow{p}0$ then $Z_n\xrightarrow{d}Y_0$. Q.E.D.

The proof that the score has zero mean and of the information matrix equality. By the proof of Theorem 3.3 it suffices to show that $\int f(z|\theta)\mathrm{d}z$ is twice differentiable and that the order of differentiation and integration can be interchanged. The following well known lemma, e.g. as found in Bartle (1966, Corollary 5.9), is useful for showing that the order of differentiation and integration can be interchanged.

Lemma 3.6

If $a(z,\theta)$ is continuously differentiable on an open set $\mathcal{N}$ of θ_0, a.s. $\mathrm{d}z$, and $\int\sup_{\theta\in\mathcal{N}}\|\nabla_\theta a(z,\theta)\|\,\mathrm{d}z<\infty$, then $\int a(z,\theta)\mathrm{d}z$ is continuously differentiable and $\nabla_\theta\int a(z,\theta)\mathrm{d}z=\int[\nabla_\theta a(z,\theta)]\mathrm{d}z$ for $\theta\in\mathcal{N}$.

Proof

Continuity of $\int [\nabla_\theta a(z,\theta)]dz$ on $\mathcal{N}$ follows by continuity of $\nabla_\theta a(z,\theta)$ in θ and the dominated convergence theorem. Also, for all $\tilde{\theta}$ close enough to θ, the line joining $\tilde{\theta}$ and θ will lie in $\mathcal{N}$, so a mean-value expansion gives $a(z,\tilde{\theta}) = a(z,\theta) + \nabla_\theta a(z,\theta)'(\tilde{\theta}-\theta) + r(z,\tilde{\theta})$, where, for the mean value $\bar{\theta}(z)$, $r(z,\tilde{\theta}) = \{\nabla_\theta a[z,\bar{\theta}(z)] - \nabla_\theta a(z,\theta)\}'(\tilde{\theta}-\theta)$. As $\tilde{\theta}\to\theta$, $\|r(z,\tilde{\theta})\|/\|\tilde{\theta}-\theta\| \leq \|\nabla_\theta a[z,\bar{\theta}(z)] - \nabla_\theta a(z,\theta)\| \to 0$ by continuity of $\nabla_\theta a(z,\theta)$. Also, $|r(z,\tilde{\theta})|/\|\tilde{\theta}-\theta\| \leq 2\sup_{\theta\in\mathcal{N}}\|\nabla_\theta a(z,\theta)\|$, so by the dominated convergence theorem, $\int |r(z,\tilde{\theta})|dz/\|\tilde{\theta}-\theta\| \to 0$. Therefore, $|\int a(z,\tilde{\theta})dz - \int a(z,\theta)dz - \{\int [\nabla_\theta a(z,\theta)]dz\}' \times (\tilde{\theta}-\theta)| = |\int r(z,\tilde{\theta})dz| \leq \int |r(z,\tilde{\theta})|dz = o(\|\tilde{\theta}-\theta\|)$. Q.E.D.

The needed result that $\int f(z|\theta)dz$ is twice differentiable and that $f(z|\theta)$ can be differentiated under the integral then follows by Lemma 3.6 and conditions (ii) and (iii) of Theorem 3.3.

4. Consistent asymptotic variance estimation

A consistent estimator of the asymptotic variance is important for construction of asymptotic confidence intervals, as discussed in the introduction. The basic idea for constructing variance estimators is to substitute, or "plug-in", estimators of the various components in the formulae for the asymptotic variance. For both extremum and minimum distance estimators, derivatives of sample functions can be used to estimate the Hessian or Jacobian terms in the asymptotic variance, when the derivatives exist. Even when derivatives do not exist, numerical approximations can be used to estimate Hessian or Jacobian terms, as discussed in Section 7. The more difficult term is the one that results from asymptotic normality of $\sqrt{n}\nabla_\theta \hat{Q}_n(\theta_0)$ or $\sqrt{n}\hat{g}_n(\theta_0)$. The form of this term depends on the nature of the estimator and whether there is dependence in the data. In this chapter, estimation of this more difficult term will only be discussed under i.i.d. data, with Wooldridge's chapter in this volume giving results for dependent observations.

To better describe variance estimation it is helpful to consider separately extremum and minimum distance estimators. The asymptotic variance of an extremum estimator is $H^{-1}\Sigma H^{-1}$, where H is the probability limit of $\nabla_{\theta\theta}\hat{Q}_n(\theta_0)$ and Σ is the asymptotic variance of $\sqrt{n}\nabla_\theta \hat{Q}_n(\theta_0)$. Thus, an estimator of the asymptotic variance can be formed as $\hat{H}^{-1}\hat{\Sigma}\hat{H}^{-1}$, where $\hat{H}$ is an estimator of H and $\hat{\Sigma}$ is an estimator of Σ. An estimator of H can be constructed in a general way, by substituting $\hat{\theta}$ for θ_0 in the Hessian of the objective function, i.e. $\hat{H} = \nabla_{\theta\theta}\hat{Q}_n(\hat{\theta})$. It is more difficult to find a general estimator of Σ, because it depends on the nature of the extremum estimator and the properties of the data.

In some cases, including MLE and NLS, an estimator of Σ can be formed in a straightforward way from sample second moments. For example, for MLE the central limit theorem implies that $\Sigma = E[\nabla_\theta \ln f(z|\theta_0)\{\nabla_\theta \ln f(z|\theta_0)\}']$, so that an

estimator can be formed by substituting moments for expectations and estimators for true parameter, i.e. $\hat{\Sigma} = n^{-1}\sum_{i=1}^{n} \nabla_\theta \ln f(z_i|\hat{\theta})\{\nabla_\theta \ln f(z_i|\hat{\theta})\}'$. More generally, an analogous estimator can be constructed whenever the objective function is a sample average, $\hat{Q}_n(\theta) = n^{-1}\sum_{i=1}^{n} q(z_i, \theta)$, e.g. where $q(z, \theta) = -[y - h(x, \theta)]^2$ for NLS. In this case $\sqrt{n}\nabla_\theta \hat{Q}_n(\theta_0) = n^{-1/2}\sum_{i=1}^{n} \nabla_\theta q(z_i, \theta_0)$, so the central limit theorem will imply that $\Sigma = E[\nabla_\theta q(z, \theta_0)\{\nabla_\theta q(z, \theta_0)\}']$.[32] This second-moment matrix can be estimated as

$$\hat{\Sigma} = n^{-1}\sum_{i=1}^{n} \nabla_\theta q(z_i, \hat{\theta})\{\nabla_\theta q(z_i, \hat{\theta})\}', \qquad \hat{Q}_n(\theta) = n^{-1}\sum_{i=1}^{n} q(z_i, \theta). \tag{4.1}$$

In cases where the asymptotic variance simplifies it will be possible to simplify the variance estimator in a corresponding way. For example the MLE asymptotic variance is the inverse of the information matrix, which can be estimated by $\hat{J}^{-1}$, for an estimator $\hat{J}$ of the information matrix. Of course, this also means that there are several ways to construct a variance estimator. For the MLE, $\hat{J}$ can be estimated from the Hessian, the sample second moment of the score, or even the general formula $\hat{H}^{-1}\hat{\Sigma}\hat{H}^{-1}$. Asymptotic distribution theory is silent about the choice between these estimators, when the models are correctly specified (i.e. the assumptions that lead to simplification are true), because any consistent estimator will lead to asymptotically correct confidence intervals. Thus, the choice between them has to be based on other considerations, such as computational ease or more refined asymptotic accuracy and length of the confidence intervals. These considerations are inherently specific to the estimator, although many results seem to suggest it is better to avoid estimating higher-order moments in the formation of variance estimators. If the model is *not* correctly specified, then the simplifications may not be valid, so that one should use the general form $\hat{H}^{-1}\hat{\Sigma}\hat{H}^{-1}$, as pointed out by Huber (1967) and White (1982a). This case is particularly interesting when $\hat{\theta}$ is consistent even though the model is misspecified, as for some MLE estimators with exponential family likelihoods; see Gourieroux et al. (1984).

For minimum distance estimation it is straightforward to estimate the Jacobian term G in the asymptotic variance $(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$, as $\hat{G} = \nabla_\theta \hat{g}_n(\hat{\theta})$. Also, by assumption $\hat{W}$ will be a consistent estimator of W. A general method of forming Ω is more difficult because the form of Ω depends on the nature of the estimator.

For GMM an estimator of Ω can be formed from sample second moments. By the central limit theorem, the asymptotic variance of $\sqrt{n}\hat{g}_n(\theta_0) = n^{-1/2}\sum_{i=1}^{n} g(z_i, \theta_0)$ is $\Omega = E[g(z, \theta_0)g(z, \theta_0)']$. Thus, an estimator can be formed by substituting sample

[32] The derivative $\nabla_\theta q(z, \theta_0)$ can often be shown to have mean zero, as needed for the central limit theorem, by a direct argument. Alternatively, a zero mean will follow from the first-order condition for maximization of $Q_0(\theta) = E[q(z, \theta)]$ at θ_0.

moments for the expectation and an estimator of θ for the true value, as

$$\hat{\Omega} = n^{-1} \sum_{i=1}^{n} g(z_i, \hat{\theta}) g(z_i, \hat{\theta})'. \tag{4.2}$$

As discussed in Section 3, extremum estimators can be considered as special cases of minimum distance estimators for analyzing asymptotic normality. More specifically, an extremum estimator with $\hat{Q}_n(\theta) = n^{-1} \sum_{i=1}^{n} q(z_i, \theta)$ will be a GMM estimator with $g(z, \theta) = \nabla_\theta q(z, \theta)$. Consequently, the estimator in eq. (4.1) is actually a special case of the one in eq. (4.2).

For minimum distance estimators, where $\hat{g}_n(\theta) = \hat{\pi} - h(\theta)$, the asymptotic variance Ω of $\sqrt{n}\hat{g}_n(\theta_0)$ is just the asymptotic variance of $\hat{\pi}$. Thus, to form $\hat{\Omega}$ one simply uses a consistent estimator of the asymptotic variance of $\hat{\pi}$. If $\hat{\pi}$ is itself an extremum or GMM estimator, its asymptotic variance can be estimated in the way described above.

When the asymptotic variance matrix simplifies there will be a corresponding simplification for an estimator. In particular, if $W = \Omega^{-1}$ then the asymptotic variance is $(G'\Omega^{-1}G)^{-1}$, so that a corresponding estimator is $(\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}$. Alternatively, if $\hat{W}$ is a consistent estimator of Ω^{-1}, a variance estimator is $(\hat{G}'\hat{W}\hat{G})^{-1}$. In addition, it may also be possible to estimate Ω in alternative ways. For example, for linear instrumental variables where $g(z, \theta) = x(y - Y'\theta)$, the estimator in eq. (4.2) is $n^{-1}\sum_{i=1}^{n} x_i x_i'(y_i - Y_i'\hat{\theta})^2$, which is consistent even if $\varepsilon_i = y_i - Y_i'\theta_0$ is heteroskedastic. An alternative estimator that would be consistent under homoskedasticity (i.e. if $E[\varepsilon^2|x]$ is constant) is $\hat{\sigma}^2 \sum_{i=1}^{n} x_i x_i'/n$ for $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(y_i - Y_i'\hat{\theta})^2$.

For minimum distance estimators, the choice between different consistent variance estimators can be based on considerations such as those discussed for extremum estimators, when the model is correctly specified. When the model is not correctly specified and there are more elements in $\hat{g}_n(\theta)$ than θ, the formula $(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$ is no longer the correct asymptotic variance matrix, the reason being that other terms enter the asymptotic variance because $\hat{g}_n(\hat{\theta})$ need not converge to zero. It is possible to show that $\hat{\theta}$ is asymptotically normal when centered at its limit, by treating it as an extremum estimator, but the formula is very complicated [e.g. see Maasoumi and Phillips (1982)]. This formula is not used often in econometrics, because it is so complicated and because, in most models where $\hat{g}_n(\theta)$ has more elements than θ, the estimator will not be consistent under misspecification.

4.1. *The basic results*

It is easy to state a consistency result for asymptotic variance estimation if $\hat{\Sigma}$ or $\hat{\Omega}$ is assumed to be consistent. A result for extremum estimators is:

Theorem 4.1

If the hypotheses of Theorem 3.1 are satisfied, $\hat{H} = \nabla_{\theta\theta}\hat{Q}_n(\hat{\theta})$, and $\hat{\Sigma} \xrightarrow{p} \Sigma$, then $\hat{H}^{-1}\hat{\Sigma}\hat{H}^{-1} \xrightarrow{p} H^{-1}\Sigma H^{-1}$.

Proof

By asymptotic normality, $\hat{\theta} \xrightarrow{p} \theta_0$. By condition (iv) of Theorem 3.1, with probability approaching one, $\|\hat{H} - H\| \leqslant \|\hat{H} - H(\hat{\theta})\| + \|H(\hat{\theta}) - H\| \leqslant \sup_{\theta \in \mathcal{N}} \|\nabla_{\theta\theta}\hat{Q}_n(\theta) - H(\theta)\| + \|H(\hat{\theta}) - H\| \xrightarrow{p} 0$, so that $\hat{H} \xrightarrow{p} H$. The conclusion then follows by condition (v) of Theorem 3.1 and continuity of matrix inversion and multiplication. Q.E.D.

A corresponding result for minimum distance estimators is:

Theorem 4.2

If the hypotheses of Theorem 3.2 are satisfied, $\hat{G} = \nabla_\theta \hat{g}_n(\hat{\theta})$, and $\hat{\Omega} \xrightarrow{p} \Omega$, then $(\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{\Omega}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1} \xrightarrow{p} (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$.

Proof

It follows similarly to the proof of Theorem 4.1 that condition (iv) of Theorem 3.2 implies $\hat{G} \xrightarrow{p} G$, while $\hat{W} \xrightarrow{p} W$ and $\hat{\Omega} \xrightarrow{p} \Omega$ hold by hypothesis. The conclusion then follows from condition (v) of Theorem 3.2 and continuity of matrix inversion and multiplication. Q.E.D.

As discussed above, the asymptotic variance for MLE, NLS, and GMM can be estimated using sample second moments, with true parameters replaced by estimators. This type of estimator will be consistent by the law of large numbers, as long as the use of estimators in place of true parameters does not affect the limit. The following result is useful in this respect.

Lemma 4.3

If z_i is i.i.d., $a(z, \theta)$ is continuous at θ_0 with probability one, and there is a neighborhood $\mathcal{N}$ of θ_0 such that $E[\sup_{\theta \in \mathcal{N}} \|a(z, \theta)\|] < \infty$, then for any $\tilde{\theta} \xrightarrow{p} \theta_0$, $n^{-1}\sum_{i=1}^n a(z_i, \tilde{\theta}) \xrightarrow{p} E[a(z, \theta_0)]$.

Proof

By consistency of $\tilde{\theta}$ there is $\delta_n \to 0$ such that $\|\tilde{\theta} - \theta_0\| \leqslant \delta_n$ with probability approaching one. Let $\Delta_n(z) = \sup_{\|\theta - \theta_0\| \leqslant \delta_n} \|a(z, \theta) - a(z, \theta_0)\|$. By continuity of $a(z, \theta)$ at θ_0, $\Delta_n(z) \to 0$ with probability one, while by the dominance condition, for n large enough $\Delta_n(z) \leqslant 2\sup_{\theta \in \mathcal{N}} \|a(z, \theta)\|$. Then by the dominated convergence theorem, $E[\Delta_n(z)] \to 0$, so by the Markov inequality, $P(|n^{-1}\sum_{i=1}^n \Delta_n(z_i)| > \varepsilon) \leqslant E[\Delta_n(z)]/\varepsilon \to 0$ for all $\varepsilon > 0$, giving $n^{-1}\sum_{i=1}^n \Delta_n(z_i) \xrightarrow{p} 0$. By Khintchine's law of large numbers, $n^{-1}\sum_{i=1}^n a \times$

$(z_i, \theta_0) \xrightarrow{p} E[a(z, \theta_0)]$. Also, with probability approaching one, $\| n^{-1} \sum_{i=1}^n a(z_i, \hat{\theta}) - n^{-1} \sum_{i=1}^n a(z_i, \theta_0) \| \leqslant n^{-1} \sum_{i=1}^n \| a(z_i, \hat{\theta}) - a(z_i, \theta_0) \| \leqslant n^{-1} \sum_{i=1}^n \Delta_n(z_i) \xrightarrow{p} 0$, so the conclusion follows by the triangle inequality. Q.E.D.

The conditions of this result are even weaker than those of Lemma 2.4, because the conclusion is simply uniform convergence at the true parameter. In particular, the function is only required to be continuous at the true parameter. This weak type of condition is not very important for the cases considered so far, e.g. for GMM where the moment functions have been assumed to be differentiable, but it is very useful for the results of Section 7, where some discontinuity of the moments is allowed. For example, for the censored LAD estimator the asymptotic variance depends on indicator functions for positivity of $x'\theta$ and Lemma 4.3 can be used to show consistency of asymptotic variance estimators that depend on such indicator functions.

4.2. *Variance estimation for MLE*

The asymptotic variance of the maximum likelihood estimator is J^{-1}, the inverse of the Fisher information matrix. It can be consistently estimated from $\hat{J}^{-1}$, where $\hat{J}$ is a consistent estimator of the information matrix. There are several ways to estimate the information matrix. To describe these ways, let $s(z, \theta) = \nabla_\theta \ln f(z|\theta)$ denote the score. Then by the information matrix equality, $J = E[s(z, \theta_0)s(z, \theta_0)'] = -E[\nabla_\theta s(z, \theta_0)] = J(\theta_0)$, where $J(\theta) = -\int [\nabla_\theta s(z, \theta)] f(z|\theta) \mathrm{d}z$. That is, J is the expectation of the outer product of the score and the expectation of the negative of the derivative of the score, i.e. of the Hessian of the log-likelihood. This form suggests that J might be estimated by the method of moments, replacing expectations by sample averages and unknown parameter values by estimates. This yields two estimators,

$$\hat{J}_1 = n^{-1} \sum_{i=1}^n s(z_i, \hat{\theta}) s(z_i, \hat{\theta})'/n, \qquad \hat{J}_2 = -n^{-1} \sum_{i=1}^n \nabla_{\theta\theta} \ln f(z_i|\hat{\theta}). \tag{4.3}$$

The second estimator is just the negative of the Hessian, and so will be consistent under the conditions of Theorem 3.3. Lemma 4.3 can be used to formulate conditions for consistency of the first estimator.

A third estimator could be obtained by substituting $\hat{\theta}$ in the integrated function $J(\theta)$. This estimator is often not feasible in econometrics, because $f(z|\theta)$ is a conditional likelihood, e.g. conditioned on regressors, and so the integration in $J(\theta)$ involves the unknown marginal distribution. An alternative estimator that is feasible is the sample average of the conditional information matrix. To describe this estimator, suppose that $z = (y, x)$ and that $f(z|\theta) = f(y|x, \theta)$ is the conditional density of y given x. Let $J(x, \theta) = E[s(z, \theta)s(z, \theta)'|x, \theta] = \int s(z, \theta)s(z, \theta)' f(y|x, \theta) \mathrm{d}y$ be the con-

ditional information matrix, so that $J = E[J(x, \theta_0)]$ by the law of iterated expectations. The third estimator of the information matrix is then

$$\hat{J}_3 = \sum_{i=1}^{n} J(x_i, \hat{\theta})/n. \tag{4.4}$$

Lemma 4.3 can be used to develop conditions for consistency of this estimator. In particular, it will often be the case that $a(z, \theta) = J(x, \theta)$ is continuous in θ, because the integration in $J(x, \theta)$ tends to smooth out any discontinuities. Consistency will then follow from a dominance condition for $J(x, \theta)$.

The following result gives conditions for consistency of all three of these estimators:

Theorem 4.4

Suppose that the hypotheses of Theorem 3.3 are satisfied. Then $\hat{J}_2^{-1} \xrightarrow{p} J^{-1}$. Also, if there is a neighborhood $\mathcal{N}$ of θ_0 such that $E[\sup_{\theta \in \mathcal{N}} \| s(z, \theta) \|^2] < \infty$ then $\hat{J}_1^{-1} \xrightarrow{p} J^{-1}$. Also, if $J(x, \theta)$ is continuous at θ_0 with probability one and $E[\sup_{\theta \in \mathcal{N}} \| J(x, \theta) \|] < \infty$ then $\hat{J}_3^{-1} \xrightarrow{p} J^{-1}$.

Proof

It follows as in the proof of Theorem 4.1 that $\hat{J}_2^{-1} \xrightarrow{p} J^{-1}$. Also, by $s(z, \theta)$ continuously differentiable in a neighborhood of θ_0, $a(z, \theta) = s(z, \theta)s(z, \theta)'$ so consistency of $\hat{J}_1^{-1}$ follows from Lemma 4.3. Also, consistency of $\hat{J}_3^{-1}$ follows by Lemma 4.3 with $a(z, \theta) = J(x, \theta)$. Q.E.D.

The regularity conditions for consistency of each of these estimators are quite weak, and so typically they all will be consistent when the likelihood is twice differentiable. Since only consistency is required for asymptotically correct confidence intervals for θ, the asymptotic theory for $\hat{\theta}$ provides no guide as to which of these one should use. However, there are some known properties of these estimators that are useful in deciding which to use. First, $\hat{J}_1$ is easier to compute than $\hat{J}_2$, which is easier to compute than $\hat{J}_3$. Because it is easiest to compute, $\hat{J}_1$ has seen much use in maximum likelihood estimation and inference, as in Berndt et al. (1974). In at least some cases they seem to rank the opposite way in terms of how closely the asymptotic theory approximates the true confidence interval distribution; e.g. see Davidson and MacKinnon (1984). Since the estimators are ranked differently according to different criteria, none of them seems always preferred to the others.

One property shared by all inverse information matrix estimators for the MLE variance is that they may not be consistent if the distribution is misspecified, as pointed out by Huber (1967) and White (1982a). If $f(z|\theta_0)$ is not the true p.d.f. then the information matrix equality will generally not hold. An alternative estimator that will be consistent is the general extremum estimator formula $\hat{J}_2^{-1}\hat{J}_1\hat{J}_2^{-1}$. Sufficient regularity conditions for its consistency are that $\hat{\theta} \xrightarrow{p} \theta_0$, $\ln f(z|\theta)$ satisfy

parts (ii) and (iv) of Theorem 3.3, $E[\sup_{\theta \in \mathcal{N}} \| \nabla_\theta \ln f(z|\theta) \|^2]$ be finite for a neighborhood $\mathcal{N}$ of θ_0, and $E[\nabla_{\theta\theta} \ln f(z|\theta_0)]$ be nonsingular.

Example 1.1 continued

It would be straightforward to give the formulae $\hat{J}_1$ and $\hat{J}_2$ using the derivatives derived earlier. In this example, there are no conditioning variables x, so that $\hat{J}_3^{-1}$ would simply be the information formula evaluated at $\hat{\theta}$. Alternatively, since it is known that the information matrix is diagonal, one could replace $\hat{J}_1^{-1}$ and $\hat{J}_2^{-1}$ with same matrices, except that before the inversion the off-diagonal elements are set equal to zero. For example, the matrix corresponding to $\hat{J}_2^{-1}$ would produce a variance estimator for $\hat{\mu}$ of $n\hat{\sigma}^2 / \sum_{i=1}^n \ell_{\varepsilon\varepsilon}(\hat{\varepsilon}_i)$, for $\hat{\varepsilon}_i = \hat{\sigma}^{-1}(z_i - \hat{\mu})$. Consistency of all of these estimators will follow by Theorem 4.4

Sometimes some extra conditions are needed for consistency of $\hat{J}_1^{-1}$ or $\hat{J}_3^{-1}$, as illustrated by the probit example.

Example 1.2 continued

For probit, the three information matrix estimators discussed above are, for $\lambda(\varepsilon) = \phi(\varepsilon)/\Phi(\varepsilon)$,

$$\hat{J}_3 = n^{-1} \sum_{i=1}^n x_i x_i' \lambda(x_i'\hat{\theta}) \lambda(-x_i'\hat{\theta}),$$

$$\hat{J}_2 = \hat{J}_3 + n^{-1} \sum_{i=1}^n x_i x_i' [\mathrm{d}\{\Phi(-v)^{-1}\lambda(v)\}/\mathrm{d}v]|_{v = x_i'\hat{\theta}} [y_i - \Phi(x_i'\hat{\theta})],$$

$$\hat{J}_1 = n^{-1} \sum_{i=1}^n x_i x_i' \Phi(-x_i'\hat{\theta})^{-2} \lambda(x_i'\hat{\theta})^2 \{y_i - \Phi(x_i'\hat{\theta})\}^2.$$

Both $\hat{J}_3^{-1} \xrightarrow{\mathrm{p}} J^{-1}$ and $\hat{J}_2^{-1} \xrightarrow{\mathrm{p}} J^{-1}$ will follow from consistency of $\hat{\theta}$, $E[\| x \|^2]$ finite, and J nonsingular. However, consistency of $\hat{J}_1^{-1}$ seems to require that $E[\| x \|^4]$ is finite, because the score satisfies $\| \nabla_\theta \ln f(z|\theta) \|^2 \leqslant |\Phi(x'\theta)^{-1} \Phi(-x'\theta)^{-1} \phi(x'\theta)| 4 \| x \|^2 \leqslant 4[C_2(1 + \| x \| \, \| \theta \|)]^2 \| x \|^2 \leqslant C(1 + \| x \|^4)$.

The variance of nonlinear least squares has some special features that can be used to simplify its calculation. By the conditional mean assumption that $E[y|x] = h(x, \theta_0)$, the Hessian term in the asymptotic variance is

$$\begin{aligned} H &= 2\{E[h_\theta(x, \theta_0) h_\theta(x, \theta_0)'] - E[h_{\theta\theta}(x, \theta_0)\{y - h(x, \theta_0)\}]\} \\ &= 2E[h_\theta(x, \theta_0) h_\theta(x, \theta_0)'], \end{aligned}$$

where h_θ denotes the gradient, $h_{\theta\theta}$ the Hessian of $h(x, \theta)$, and the second equality

follows by the law of iterated expectations. Therefore, H can be estimated by $\hat{H} = 2n^{-1}\sum_{i=1}^{n} h_\theta(x_i, \hat{\theta})h_\theta(x_i, \hat{\theta})'$, which is convenient because it only depends on first derivatives, rather than first and second derivatives. Under homoskedasticity the matrix Σ also simplifies, to $4\sigma^2 E[h_\theta(x, \theta_0)h_\theta(x, \theta_0)']$ for $\sigma^2 = E[\{y - h(x, \theta_0)\}^2]$, which can be estimated by $2\hat{\sigma}^2\hat{H}$ for $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\{y - h(x_i, \hat{\theta})\}^2$. Combining this estimator of Σ with the one for H gives an asymptotic variance estimator of the form $\hat{V} = \hat{H}^{-1}\hat{\Sigma}\hat{H}^{-1} = 2\hat{\sigma}^2\hat{H}^{-1}$. Consistency of this estimator can be shown by applying the conditions of Lemma 4.3 to both $a(z, \theta) = \{y - h(x, \theta)\}^2$ and $a(z, \theta) = h_\theta(x, \theta)h_\theta(x, \theta)'$, which is left as an exercise.

If there is heteroskedasticity then the variance of y does not factor out of Σ, so that one must use the estimator $\hat{\Sigma} = 4n^{-1}\sum_{i=1}^{n} h_\theta(x_i, \hat{\theta})h_\theta(x_i, \hat{\theta})'\{y_i - h(x_i, \hat{\theta})\}^2$. Also, if the conditional expectation is misspecified, then second derivatives of the regression function do not disappear from the Hessian (except in the linear case), so that one must use the estimator $\hat{H} = 2n^{-1}\sum_{i=1}^{n}[h_\theta(x_i, \hat{\theta})h_\theta(x_i, \hat{\theta})' + h_{\theta\theta}(x_i, \hat{\theta})'\{y_i - h(x_i, \hat{\theta})\}]$. A variance estimator for NLS that is consistent in spite of heteroskedasticity or misspecification is $\hat{H}^{-1}\hat{\Sigma}\hat{H}^{-1}$, as discussed in White (1982b). One could formulate consistency conditions for this estimator by applying Lemma 4.3. The details are left as an exercise.

4.3. *Asymptotic variance estimation for GMM*

The asymptotic variance of a GMM estimator is $(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$, which can be estimated by substituting estimators for each of G, W and Ω. As previously discussed, estimators of G and W are readily available, and are given by $\hat{G} = n^{-1}\sum_{i=1}^{n}\nabla_\theta g(z_i, \hat{\theta})$ and $\hat{W}$, where $\hat{W}$ is the original weighting matrix. To estimate $\Omega = E[g(z, \theta_0)g(z, \theta_0)']$, one can replace the population moment by a sample average and the true parameter by an estimator, to form $\hat{\Omega} = n^{-1}\sum_{i=1}^{n} g(z_i, \hat{\theta})g(z_i, \hat{\theta})'$, as in eq. (4.2). The estimator of the asymptotic variance is then given by $\hat{V} = (\hat{G}'\hat{W}\hat{G})^{-1} \times \hat{G}'\hat{W}\hat{\Omega}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1}$.

Consistency of $\hat{\Omega}$ will follow from Lemma 4.3 with $a(z, \theta) = g(z, \theta)g(z, \theta)'$, so that consistency of $\hat{V}$ will hold under the conditions of Theorem 4.2, as applied to GMM. A result that summarizes these conditions is the following one:

Theorem 4.5

If the hypotheses of Theorem 3.4 are satisfied, $g(z, \theta)$ is continuous at θ_0 with probability one, and for a neighborhood $\mathcal{N}$ of θ_0, $E[\sup_{\theta\in\mathcal{N}} \|g(z, \theta)\|^2] < \infty$, then $\hat{V} = (\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{\Omega}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1} \xrightarrow{p} (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$.

Proof

By Lemma 4.3 applied to $a(z, \theta) = g(z, \theta)g(z, \theta)'$, $\hat{\Omega} \xrightarrow{p} \Omega$. Also, the proof of Theorem 3.4 shows that the hypotheses of Theorem 3.2 are satisfied, so the conclusion follows by Theorem 4.2. Q.E.D.

If $\hat{W}$ is a consistent estimator of Ω^{-1}, i.e. the probability limit W of $\hat{W}$ is equal to Ω^{-1}, then a simpler estimator of the asymptotic variance can be formed as $\hat{V} = (\hat{G}'\hat{W}\hat{G})^{-1}$. Alternatively, one could form $\hat{\Omega}$ as in eq. (4.2) and use $\hat{V} = (\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}$. Little seems to be known about the relative merits of these two procedures in small samples, i.e. which (if either) of the initial $\hat{W}$ or the final $\hat{\Omega}^{-1}$ gives more accurate or shorter confidence intervals.

The asymptotic variance estimator $\hat{V}$ is very general, in that it does not require that the second moment matrix $\Omega = E[g(z,\theta_0)g(z,\theta_0)']$ be restricted in any way. Consequently, consistency of $\hat{V}$ does not require substantive distributional restrictions other than $E[g(z,\theta_0)] = 0$.[33] For example, in the context of least squares estimation, where $g(z,\theta) = x(y - x'\theta)$, $\hat{W} = I$, and $\hat{G} = -\sum_{i=1}^n x_i x_i'/n$, this GMM variance estimator is $\hat{V} = \hat{G}^{-1}[n^{-1}\sum_{i=1}^n x_i x_i'(y_i - x_i'\hat{\theta})^2]\hat{G}^{-1}$, the Eicker (1967) and White (1980) heteroskedasticity consistent variance estimator. Furthermore, the GMM variance estimator includes many heteroskedasticity-robust IV variance estimators, as discussed in Hansen (1982).

When there is more information about the model than just the moment restrictions, it may improve the asymptotic confidence interval approximation to try to use this information in estimation of the asymptotic variance. An example is least squares, where the usual estimator under homoskedasticity is $n(\sum_{i=1}^n x_i x_i')^{-1}\sum(y_i - x_i'\hat{\theta})^2/(n-K)$, where K is the dimension of x. It is well known that under homoskedasticity this estimator gives more accurate confidence intervals than the heteroskedasticity consistent one, e.g. leading to exact confidence intervals from the t-distribution under normality.

Example 1.3 continued

The nonlinear two-stage least squares estimator for the Hansen–Singleton example is a GMM estimator with $g(z,\theta) = x\{\beta w y^{\gamma} - 1\}$ and $\hat{W} = \sum_{i=1}^n x_i x_i'/n$, so that an asymptotic variance estimator can be formed by applying the general GMM formula to this case. Here an estimator of the variance of the moment functions can be formed as described above, with $\hat{\Omega} = n^{-1}\sum_{i=1}^n x_i x_i\{\hat{\beta} w_i y_i^{\hat{\gamma}} - 1\}^2$. The Jacobian estimator is $\hat{G} = n^{-1}\sum_{i=1}^n x_i(w_i y_i^{\hat{\gamma}},\ \hat{\beta} w_i \ln(y_i) y_i^{\hat{\gamma}})$. The corresponding asymptotic variance estimator then comes from the general GMM formula $(\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{\Omega}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1}$.

Consistency of this estimator will follow under the conditions of Theorem 4.5. It was previously shown that all of these conditions are satisfied except the additional moment assumption stated in Theorem 4.5. For this assumption, it suffices that the upper and lower limits on γ, namely γ_ℓ and γ_u, satisfy $E[\|x\|^2|w|^2(|y|^{2\gamma_\ell} + |y|^{2\gamma_u})] < \infty$. This condition requires that slightly more moments exist than the previous conditions that were imposed.

[33] If this restriction is not satisfied, then a GMM estimator may still be asymptotically normal, but the asymptotic variance is much more complicated; see Maasoumi and Phillips (1982) for the instrumental variables case.

5. Asymptotic efficiency

Asymptotically normal estimators can be compared on the basis of their asymptotic variances, with one being asymptotically efficient relative to another if it has at least as small an asymptotic variance for all possible true parameter values. Asymptotic efficiency is desirable because an efficient estimator will be closer to the true parameter value in large samples; if $\hat{\theta}$ is asymptotically efficient relative to $\tilde{\theta}$ then for all constants K, $\text{Prob}(|\hat{\theta} - \theta_0| \leqslant K/\sqrt{n}) > \text{Prob}(|\tilde{\theta} - \theta_0| \leqslant K/\sqrt{n})$ for all n large enough. Efficiency is important in practice, because it results in smaller asymptotic confidence intervals, as discussed in the introduction.

This section discusses general results on asymptotic efficiency within a class of estimators, and application of these results to important estimation environments, both old and new. In focusing on efficiency within a class of estimators, we follow much of the econometrics and statistics literature.[34] Also, this efficiency framework allows one to derive results on efficiency within classes of "limited information" estimators (such as single equation estimators in a simultaneous system), which are of interest because they are relatively insensitive to misspecification and easier to compute. An alternative approach to efficiency analysis, that also allows for limited information estimators, is through semiparametric efficiency bounds, e.g. see Newey (1990). The approach taken here, focusing on classes of estimators, is simpler and more directly linked to the rest of this chapter.

Two of the most important and famous efficiency results are efficiency of maximum likelihood and the form of an optimal weighting matrix for minimum distance estimation. Other useful results are efficiency of heteroskedasticity-corrected generalized least squares in the class of weighted least squares estimators and two-stage least squares as an efficient instrumental variables estimator. All of these results share a common structure that is useful in understanding them and deriving new ones. To motivate this structure, and focus attention on the most important results, we first consider separately maximum likelihood and minimum distance estimation.

5.1. *Efficiency of maximum likelihood estimation*

Efficiency of maximum likelihood is a central proposition of statistics that dates from the work of R.A. Fisher (1921). Although maximum likelihood is not efficient in the class of all asymptotically normal estimators, because of "superefficient" estimators, it is efficient in quite general classes of estimators.[35] One such general class is the

[34] In particular, one of the precise results on efficiency of MLE is the Hajek–LeCam representation theory, which shows efficiency in a class of *regular* estimators. See, e.g. Newey (1990) for a discussion of regularity.

[35] The word "superefficient" refers to a certain type of estimator, attributed to Hodges, that is used to show that there does not exist an efficient estimator in the class of all asymptotically normal estimators. Suppose $\hat{\theta}$ is asymptotically normal, and for some number α and $0 < \beta < \frac{1}{2}$, suppose that $\hat{\theta}$ has positive asymptotic variance when the true parameter is α. Let $\tilde{\theta} = \hat{\theta}$ if $n^{\beta}|\hat{\theta} - \alpha| > 1$ and $\tilde{\theta} = \alpha$ if $n^{\beta}|\hat{\theta} - \alpha| < 1$. Then $\tilde{\theta}$ is superefficient relative to $\hat{\theta}$, having the same asymptotic variance when the true parameter is not α but having a smaller asymptotic variance, of zero, when the true parameter is α.

class of GMM estimators, which includes method of moments, least squares, instrumental variables, and other estimators. Because this class includes so many estimators of interest, efficiency in this class is a useful way of thinking about MLE efficiency.

Asymptotic efficiency of MLE among GMM estimators is shown by comparing asymptotic variances. The asymptotic variance of the MLE is $(E[ss'])^{-1}$, where $s = \nabla_\theta \ln f(z|\theta_0)$ is the score, with the z and θ arguments suppressed for notational convenience. The asymptotic variance of a GMM estimator can be written as $(E[m_\theta])^{-1}E[mm'](E[m_\theta'])^{-1}$ where $m_\theta = (E[\nabla_\theta g(z,\theta_0)])'W\nabla_\theta g(z,\theta_0)$ and $m = (E[\nabla_\theta g(z,\theta_0)])'Wg(z,\theta_0)$. At this point the relationship between the GMM and MLE variances is not clear. It turns out that a relationship can be derived from an interpretation of $E[m_\theta]$ as the covariance of m with the score. To obtain this interpretation, consider the GMM moment condition $\int g(z,\theta)f(z|\theta)\,dz = 0$. This condition is typically an *identity* over the parameter space that is necessary for consistency of a GMM estimator. If it did not hold at a parameter value, then the GMM estimator may not converge to the parameter at that point, and hence would not be consistent.[36] Differentiating this identity, assuming differentiation under the integral is allowed, gives

$$\begin{aligned} 0 &= \nabla_\theta \int g(z,\theta)f(z|\theta)\,dz|_{\theta=\theta_0} \\ &= \left\{ \int [\nabla_\theta g(z,\theta)]f(z|\theta)\,dz + \int g(z,\theta)[\nabla_\theta f(z|\theta)]'\,dz \right\}\Bigg|_{\theta=\theta_0} \\ &= E[\nabla_\theta g(z,\theta_0)] + E[g(z,\theta_0)\nabla_\theta \ln f(z|\theta_0)'], \end{aligned} \tag{5.1}$$

where the last equality follows by multiplying and dividing $\nabla_\theta f(z|\theta_0)$ by $f(z|\theta_0)$. This is the *generalized information matrix equality*, including the information matrix equality as a special case, where $g(z,\theta) = \nabla_\theta \ln f(z|\theta)$.[37] It implies that $E[m_\theta] + E[ms'] = 0$, i.e. that $E[m_\theta] = -E[ms']$. Then the difference of the GMM and MLE asymptotic variances can be written as

$$\begin{aligned} &(E[m_\theta])^{-1}E[mm'](E[m_\theta'])^{-1} - (E[ss'])^{-1} \\ &\quad = (E[ms'])^{-1}E[mm'](E[sm'])^{-1} - (E[ss'])^{-1} \\ &\quad = (E[ms'])^{-1}\{E[mm'] - E[ms'](E[ss'])^{-1}E[sm']\}(E[sm'])^{-1} \\ &\quad = (E[ms'])^{-1}E[UU'](E[sm'])^{-1}, \quad U = m - E[ms'](E[ss'])^{-1}s. \end{aligned} \tag{5.2}$$

[36] Recall that consistency means that the estimator converges in probability to the true parameter for *all* possible true parameter values.

[37] A similar equality, used to derive the Cramer–Rao bound for the variance of unbiased estimators, is obtained by differentiating the identity $\theta = \int \hat{\theta}\,dF_\theta$, where F_θ is the distribution of the data when θ is the true parameter value.

Since $E[UU']$ is positive semi-definite, the difference of the respective variance matrices is also positive semi-definite, and hence the MLE is asymptotically efficient in the class of GMM estimators.

To give a precise result it is necessary to specify regularity conditions for the generalized information matrix equality of eq. (5.1). Conditions can be formulated by imposing smoothness on the square root of the likelihood, $f(z|\theta)^{1/2}$, similar to the regularity conditions for MLE efficiency of LeCam (1956) and Hajek (1970). A precise result on efficiency of MLE in the class of GMM estimators can then be stated as:

Theorem 5.1

If the conditions of Theorem 3.4 are satisfied, $f(z|\theta)^{1/2}$ is continuously differentiable at θ_0, J is nonsingular, and for all θ in a neighborhood $\mathcal{N}$ of θ_0, $\int \sup_{\tilde{\theta}\in\mathcal{N}} \| g(z,\tilde{\theta})\|^2 \times f(z|\theta)\,dz$ and $\int \sup_{\tilde{\theta}\in\mathcal{N}} \|\nabla_\theta f(z|\tilde{\theta})^{1/2}\|^2\,dz$ are bounded and $\int g(z,\theta) f(z|\theta)\,dz = 0$, then $(G'WG)^{-1}G'W\Omega WG(G'WG) - J^{-1}$ is positive semi-definite.

The proof is postponed until Section 5.6. This result states that J^{-1} is a lower bound on the asymptotic variance of a GMM estimator. Asymptotic efficiency of MLE among GMM estimators then follows from Theorem 3.4, because the MLE will have J^{-1} for its asymptotic variance.[38]

5.2. *Optimal minimum distance estimation*

The asymptotic variance of a minimum distance estimator depends on the limit W of the weighting matrix $\hat{W}$. When $W = \Omega^{-1}$, the asymptotic variance of a minimum distance estimator is $(G'\Omega^{-1}G)^{-1}$. It turns out that this estimator is efficient in the class of minimum distance estimators. To show this result, let Z be any random vector such that $\Omega = E[ZZ']$, and let $m = G'WZ$ and $\bar{m} = G'\Omega^{-1}Z$. Then by $G'WG = E[m\bar{m}']$ and $G'\Omega^{-1}G = E[\bar{m}\bar{m}']$,

$$(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1} - (G'\Omega^{-1}G)^{-1}$$
$$= (G'WG)^{-1}E[UU'](G'WG)^{-1}, \quad U = m - E[m\bar{m}'](E[\bar{m}\bar{m}'])^{-1}\bar{m}. \tag{5.3}$$

Since $E[UU']$ is positive semi-definite, the difference of the asymptotic variances is positive semi-definite. This proves the following result:

[38] It is possible to show this result under the weaker condition that $f(z|\theta)^{1/2}$ is mean-square differentiable, which allows for $f(z|\theta)$ to not be continuously differentiable. This condition is further discussed in Section 5.5.

Theorem 5.2

If Ω is nonsingular, a minimum distance estimator with $W = \text{plim}(\hat{W}) = \Omega^{-1}$ is asymptotically efficient in the class of minimum distance estimators.

This type of result is familiar from efficiency theory for CMD and GMM estimation. For example, in minimum chi-square estimation, where $\hat{g}(\theta) = \hat{\pi} - \pi(\theta)$, the efficient weighting matrix W is the inverse of the asymptotic variance of $\hat{\pi}$, a result given by Chiang (1956) and Ferguson (1958). For GMM, where $\hat{g}(\theta) = \sum_{i=1}^{n} g(z_i, \theta)/n$, the efficient weighting matrix is the inverse of the variance of $g(z_i, \theta_0)$, a result derived by Hansen (1982). Each of these results is a special case of Theorem 5.2.

Construction of an efficient minimum distance estimator is quite simple, because the weighting matrix affects the asymptotic distribution only through its probability limit. All that is required is a consistent estimator $\hat{\Omega}$, for then $\hat{W} = \hat{\Omega}^{-1}$ will converge in probability to Ω^{-1}. Since an estimator of Ω is needed for asymptotic variance estimation, very little additional effort is required to form an efficient weighting matrix. An efficient minimum distance estimator can then be constructed by minimizing $\hat{g}(\theta)'\hat{\Omega}^{-1}\hat{g}(\theta)$. Alternatively, the one-step estimator $\bar{\theta} = \hat{\theta} - (\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1} \times \hat{G}\hat{\Omega}^{-1}\hat{g}(\hat{\theta})$ will also be efficient, because it is asymptotically equivalent to the fully iterated minimum distance estimator.

The condition that $W = \Omega^{-1}$ is sufficient but not necessary for efficiency. A necessary and sufficient condition can be obtained by further examination of eq. (5.3). A minimum distance estimator will be efficient if and only if the random vector U is zero. This vector is the residual from a population regression of m on $\bar{m}$, and so will be zero if and only if m is a linear combination of $\bar{m}$, i.e. there is a constant matrix C such that $G'WZ = CG'\Omega^{-1}Z$. Since Z has nonsingular variance matrix, this condition is the same as

$$G'W = CG'\Omega^{-1}. \tag{5.4}$$

This is the necessary and sufficient condition for efficiency of a minimum distance estimator.

5.3. *A general efficiency framework*

The maximum likelihood and minimum distance efficiency results have a similar structure, as can be seen by comparing eqs. (5.2) and (5.3). This structure can be exploited to construct an efficiency framework that includes these and other important results, and is useful for finding efficient estimators. To describe this framework one needs notation for the asymptotic variance associated with an estimator. To this end, let τ denote an "index" for the asymptotic variance of an estimator in some

class, where τ is an element of some abstract set. A completely general form for τ would be the sequence of functions of the data that *is* the sequence of estimators. However, since τ is only needed to index the asymptotic variance, a simpler specification will often suffice. For example, in the class of minimum distance estimators with given $\hat{g}_n(\theta)$, the asymptotic variance depends only on $W = \text{plim}(\hat{W})$, so that it suffices to specify that $\tau = W$.

The framework considered here is one where there is a random vector Z such that for each τ (corresponding to an estimator), there is $D(\tau)$ and $m(Z, \tau)$ with the asymptotic variance $V(\tau)$ satisfying

$$V(\tau) = D(\tau)^{-1} E[m(Z, \tau)m(Z, \tau)']D(\tau)^{-1'}. \tag{5.5}$$

Note that the random vector Z is held fixed as τ varies. The function $m(Z, \tau)$ can often be interpreted as a score or moment function, and the matrix $D(\tau)$ as a Jacobian matrix for the parameters. For example, the asymptotic variances of the class of GMM estimators satisfy this formula, with τ being $[g(z, \theta_0), G, W]$, $Z = z$ being a single observation, $m(Z, \tau) = G'Wg(z, \theta_0)$, and $D(\tau) = G'WG$. Another example is minimum distance estimators, where Z is any random vector with mean zero and variance Ω, $\tau = W$, $m(Z, \tau) = G'WZ$, and $D(\tau) = G'WG$.

In this framework, there is an interesting and useful characterization of an efficient estimator.

Theorem 5.3

If $\bar{\tau}$ satisfies $D(\tau) = E[m(Z, \tau)m(Z, \bar{\tau})']$ for all τ then any estimator with variance $V(\bar{\tau})$ is efficient. Furthermore, suppose that for any τ_1, τ_2, and constant square matrices C_1, C_2 such that $C_1 D(\tau_1) + C_2 D(\tau_2)$ is nonsingular, there is τ_3 with (i) (linearity of the moment function set) $m(Z, \tau_3) = C_1 m(Z, \tau_1) + C_2 m(Z, \tau_2)$; (ii) (linearity of D) $D(\tau_3) = C_1 D(\tau_1) + C_2 D(\tau_2)$. If there is an efficient estimator with $E[m(Z, \tau)m(Z, \tau)']$ nonsingular then there is an efficient estimator with index $\bar{\tau}$ such that $D(\tau) = E[m(Z, \tau)m(Z, \bar{\tau})']$ for all τ.

Proof

If τ and $\bar{\tau}$ satisfy $D(\tau) = E[m(Z, \tau)m(Z, \bar{\tau})']$ then the difference of the respective asymptotic variances satisfies, for $m = m(Z, \tau)$ and $\bar{m} = m(Z, \bar{\tau})$,

$$\begin{aligned} V(\tau) - V(\bar{\tau}) &= (E[m\bar{m}'])^{-1} E[mm'](E[\bar{m}m'])^{-1} - (E[\bar{m}\bar{m}'])^{-1} \\ &= (E[m\bar{m}'])^{-1} E[UU'](E[\bar{m}m'])^{-1}, \\ U &= m - E[m\bar{m}'](E[\bar{m}\bar{m}'])^{-1}\bar{m}, \end{aligned} \tag{5.6}$$

so the first conclusion follows by $E[UU']$ positive semi-definite. To show the second conclusion, let $\psi(Z, \tau) = D(\tau)^{-1} m(Z, \tau)$, so that $V(\tau) = E[\psi(Z, \tau)\psi(Z, \tau)']$. Consider

any constant matrix B, and for τ_1 and τ_2 let $C_1 = BD(\tau_1)^{-1}$ and $C_2 = (I - B)D(\tau_2)^{-1}$ note that $C_1 D(\tau_1) + C_2 D(\tau_2) = I$ is nonsingular, so by (i) and (ii) there is τ_3 such that $B\psi(Z, \tau_1) + (I - B)\psi(Z, \tau_2) = C_1 m(Z, \tau_1) + C_2 m(Z, \tau_2) = m(Z, \tau_3) = I^{-1} m(Z, \tau_3) = [C_1 D(\tau_1) + C_2 D(\tau_2)]^{-1} m(Z, \tau_3) = D(\tau_3)^{-1} m(Z, \tau_3) = \psi(Z, \tau_3)$. Thus, the set $\{\psi(Z, \tau)\}$ is affine, in the sense that $B\psi(Z, \tau_1) + (I - B)\psi(Z, \tau_2)$ is in this set for any τ_1, τ_2 and constant matrix B. Let $\psi(Z, \bar{\tau})$ correspond to an efficient estimator. Suppose that there is τ with $E[(\psi - \bar{\psi})\bar{\psi}'] \neq 0$ for $\psi = \psi(Z, \tau)$ and $\bar{\psi} = \psi(Z, \bar{\tau})$. Then $\psi - \bar{\psi} \neq 0$, so there exists a constant matrix F such that $e = F(\psi - \bar{\psi})$ has nonsingular variance and $E[e\bar{\psi}'] \neq 0$. Let $B = -E[\bar{\psi}e'](E[ee'])^{-1}F$ and $u = \bar{\psi} + B(\psi - \bar{\psi}) = (I - B)\bar{\psi} + B\psi$. By the affine property of $\{\psi(Z, \tau)\}$ there is $\tilde{\tau}$ such that $V(\tilde{\tau}) = E[uu'] = E[\bar{\psi}\bar{\psi}'] - E[\bar{\psi}e'](E[ee'])^{-1}E[e\bar{\psi}'] = V(\bar{\tau}) - E[\bar{\psi}e'](E[ee'])^{-1}E[e\bar{\psi}']$, which is smaller than $V(\bar{\tau})$ in the positive semi-definite sense. This conclusion contradicts the assumed efficiency of $\bar{\tau}$, so that the assumption that $E[(\psi - \bar{\psi})\bar{\psi}'] \neq 0$ contradicts efficiency. Thus, it follows that $E[(\psi - \bar{\psi})\bar{\psi}'] = 0$ for all τ, i.e. that for all τ,

$$D(\tau)^{-1}E[m(Z, \tau)m(Z, \bar{\tau})']D(\bar{\tau})^{-1\prime} = D(\bar{\tau})^{-1}E[m(Z, \bar{\tau})m(Z, \bar{\tau})']D(\bar{\tau})^{-1\prime}. \tag{5.7}$$

By the assumed nonsingularity of $E[m(Z, \bar{\tau})m(Z, \bar{\tau})']$, this equation can be solved for $D(\tau)$ to give $D(\tau) = E[m(Z, \tau)m(Z, \bar{\tau})'](E[m(Z, \bar{\tau})m(Z, \bar{\tau})'])^{-1}D(\bar{\tau})$. Since $C = D(\bar{\tau})'(E[m(Z, \bar{\tau})m(Z, \bar{\tau})'])^{-1}$ is a nonsingular matrix it follows by (i) and (ii) that there exists $\tilde{\tau}$ with $m(Z, \tilde{\tau}) = Cm(Z, \bar{\tau})$. Furthermore, by linearity of $D(\tau)$ it follows that $V(\tilde{\tau}) = V(\bar{\tau})$, so that the estimator corresponding to $\tilde{\tau}$ is efficient. The second conclusion then follows from $D(\tau) = E[m(Z, \tau)m(Z, \tilde{\tau})']$ for all τ. Q.E.D.

This result states that

$$D(\tau) = E[m(Z, \tau)m(Z, \bar{\tau})'], \quad \text{for all } \tau, \tag{5.8}$$

is sufficient for $\bar{\tau}$ to correspond to an efficient estimator and is necessary for some efficient estimator if the set of moment functions is linear and the Jacobian is a linear function of the scores. This equality is a generalization of the information matrix equality. Hansen (1985a) formulated and used this condition to derive efficient instrumental variables estimators, and gave more primitive hypotheses for conditions (i) and (ii) of Theorem 5.3. Also, the framework here is a modified version of that of Bates and White (1992) for general classes of estimators. The sufficiency part of Theorem 5.3 appears in both of these papers. The necessity part of Theorem 5.3 appears to be new, but is closely related to R.A. Fisher's (1925) necessary condition for an efficient statistic, as further discussed below.

One interpretation of eq. (5.8) is that the asymptotic covariance between an efficient estimator and any other estimator is the variance of the efficient estimator. This characterization of an efficient estimator was discussed in R.A. Fisher (1925),

and is useful in constructing Hausman (1978) specification tests. It is derived by assuming that the asymptotic covariance between two estimators in the class takes the form $D(\tau_1)^{-1}E[m(Z,\tau_1)m(Z,\tau_2)']D(\tau_2)^{-1\prime}$, as can usually be verified by "stacking" the two estimators and deriving their joint asymptotic variance (and hence asymptotic covariance). For example, consider two different GMM estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, with two different moment functions $g_1(z,\theta)$ and $g_2(z,\theta)$, and $r = q$ for simplicity. The vector $\hat{\gamma} = (\hat{\theta}_1', \hat{\theta}_2')'$ can be considered a joint GMM estimator with moment vector $g(z,\gamma) = [g_1(z,\theta_1)', g_2(z,\theta_2)']'$. The Jacobian matrix of the stacked moment vector will be block diagonal, and hence so will its inverse, so that the asymptotic covariance between $\hat{\theta}_1$ and $\hat{\theta}_2$ will be $\{E[\nabla_\theta g_1(z,\theta_0)]\}^{-1}E[g_1(z,\theta_0)g_2(z,\theta_0)'] \times \{E[\nabla_\theta g_2(z,\theta_0)]\}^{-1\prime}$. This is exactly of the form $D(\tau_1)^{-1}E[m(Z,\tau_1)m(Z,\tau_2)']D(\tau_2)^{-1\prime}$, where $Z = z$, $m(Z,\tau_1) = g_1(z,\theta_0)$, etc. When the covariance takes this form, the covariance between any estimator and one satisfying eq. (5.8) will be $D(\tau)^{-1} \times E[m(Z,\tau)m(Z,\bar{\tau})']D(\bar{\tau})^{-1\prime} = I \cdot D(\bar{\tau})^{-1\prime} = D(\bar{\tau})^{-1}E[m(Z,\bar{\tau})m(Z,\bar{\tau})']D(\bar{\tau})^{-1\prime} = V(\bar{\tau})$, the variance of the efficient estimator. R.A. Fisher (1925) showed that this covariance condition is sufficient for efficiency, and that it is also necessary if the class of statistics is linear, in a certain sense. The role of conditions (i) and (ii) is to guarantee that R.A. Fisher's (1925) linearity condition is satisfied.

Another interpretation of eq. (5.8) is that the variance of any estimator in the class can be written as the sum of the efficient variance and the variance of a "noise term". Let $U(Z) = D(\tau)^{-1}m(Z,\tau) - D(\bar{\tau})^{-1}m(Z,\bar{\tau})$, and note that $U(Z)$ is orthogonal to $D(\bar{\tau})^{-1}m(Z,\bar{\tau})$ by eq. (5.8). Thus, $V(\tau) = V(\bar{\tau}) + E[U(Z)U(Z)']$. This interpretation is a second-moment version of the Hajek and LeCam efficiency results.

5.4. *Solving for the smallest asymptotic variance*

The characterization of an efficient estimator given in Theorem 5.3 is very useful for finding efficient estimators. Equation (5.8) can often be used to solve for $\bar{\tau}$, by following two steps: (1) specify the class of estimators so that conditions (i) and (ii) of Theorem 5.3 are satisfied, i.e. so the set of moment functions is linear and the Jacobian D is linear in the moment functions; (2) look for $\bar{\tau}$ such that $D(\tau) = E[m(Z,\tau)m(Z,\bar{\tau})']$. The importance of step (1) is that the linearity conditions guarantee that a solution to eq. (5.8) exists when there is an efficient estimator [with the variance of $m(Z,\tau)$ nonsingular], so that the effort of solving eq. (5.8) will not be in vain. Although for some classes of estimators the linearity conditions are not met, it often seems to be possible to enlarge the class of estimators so that the linearity conditions are met without affecting the efficient estimator. An example is weighted least squares estimation, as further discussed below.

Using eq. (5.8) to solve for an efficient estimator can be illustrated with several examples, both old and new. Consider first minimum distance estimators. The asymptotic variance has the form given in eq. (5.5) for the score $G'WZ$ and the Jacobian term $G'WG$. The equation for the efficient W is then $0 = G'WG - G'W\Omega\bar{W}G =$

$G'W(I - \Omega\bar{W})G$, which holds if $\Omega\bar{W} = I$, i.e. $\bar{W} = \Omega^{-1}$. Thus, in this example one can solve directly for the optimal weight matrix.

Another example is provided by the problem of deriving the efficient instruments for a nonlinear instrumental variables estimator. Let $\rho(z, \theta)$ denote an $s \times 1$ residual vector, and suppose that there is a vector of variables x such that a conditional moment restriction,

$$E[\rho(z, \theta_0)|x] = 0, \tag{5.9}$$

is satisfied. Here $\rho(z, \theta)$ can be thought of as a vector of residuals and x as a vector of instrumental variables. A simple example is a nonlinear regression model $y = f(x, \theta_0) + \varepsilon$, $E[\varepsilon|x] = 0$, where the residual $\rho(z, \theta) = y - f(x, \theta)$ will satisfy the conditional moment restriction in eq. (5.9) by ε having conditional mean zero. Another familiar example is a single equation of a simultaneous equations system, where $\rho(z, \theta) = y - Y'\theta$ and Y are the right-hand-side endogenous variables.

An important class of estimators are instrumental variable, or GMM estimators, based on eq. (5.9). This conditional moment restriction implies the unconditional moment restriction that $E[A(x)\rho(z, \theta_0)] = 0$ for any $q \times s$ matrix of functions $A(x)$. Thus, a GMM estimator can be based on the moment functions $g(z, \theta) = A(x)\rho(z, \theta)$. Noting that $\nabla_\theta g(z, \theta) = A(x)\nabla_\theta \rho(z, \theta)$, it follows by Theorem 3.4 that the asymptotic variance of such a GMM estimator will be

$$V(A) = \{E[A(x)\nabla_\theta \rho(z, \theta_0)]\}^{-1} E[A(x)\rho(z, \theta_0)\rho(z, \theta_0)'A(x)']\{E[A(x)\nabla_\theta \rho(z, \theta_0)]\}^{-1'}, \tag{5.10}$$

where no weighting matrix is present because $g(z, \theta) = A(x)\rho(z, \theta)$ has the same number of components as θ. This asymptotic variance satisfies eq. (5.5), where $\tau = A(\cdot)$ indexes the asymptotic variance. By choosing $\rho(z, \theta)$ and $A(x)$ in certain ways, this class of asymptotic variances can be set up to include all weighted least squares estimators, all single equation instrumental variables estimators, or all system instrumental variables estimators. In particular, cases with more instrumental variables than parameters can be included by specifying $A(x)$ to be a linear combination of all the instrumental variables, with linear combination coefficients given by the probability limit of corresponding sample values. For example, suppose the residual is a scalar $\rho(z, \theta) = y - Y'\theta$, and consider the 2SLS estimator with instrumental variables x. Its asymptotic variance has the form given in eq. (5.10) for $A(x) = E[Yx'](E[xx'])^{-1}x$. In this example, the probability limit of the linear combination coefficients is $E[Yx'](E[xx'])^{-1}$. For system instrumental variables estimators these coefficients could also depend on the residual variance, e.g. allowing for 3SLS.

The asymptotic variance in eq. (5.10) satisfies eq. (5.5) for $Z = z$, $D(\tau) = E[A(x) \times \nabla_\theta \rho(z, \theta_0)]$, and $m(Z, \tau) = A(x)\rho(Z, \theta_0)$. Furthermore, both $m(Z, \tau)$ and $D(\tau)$ are linear in $A(x)$, so that conditions (i) and (ii) should be satisfied if the set of functions $\{A(x)\}$

is linear. To be specific, consider the class of all $A(x)$ such that $E[A(x)\nabla_\theta \rho(z,\theta_0)]$ and $E[\|A(x)\|^2 \|\rho(z,\theta_0)\|^2]$ exist. Then conditions (i) and (ii) are satisfied with $\tau_3 = A_3(\cdot) = C_1 A_1(\cdot) + C_2 A_2(\cdot)$.[39] Thus, by Theorem 5.3, if an efficient choice of instruments exist there will be one that solves eq. (5.8). To find such a solution, let $G(x) = E[\nabla_\theta \rho(z,\theta_0)|x]$ and $\Omega(x) = E[\rho(z,\theta_0)\rho(z,\theta_0)'|x]$, so that by iterated expectations eq. (5.8) is $0 = E[A(x)\{G(x) - \Omega(x)\bar{A}(x)'\}]$. This equation will be satisfied if $G(x) - \Omega(x)\bar{A}(x)' = 0$, i.e. if

$$\bar{A}(x) = G(x)'\Omega(x)^{-1}. \tag{5.11}$$

Consequently, this function minimizes the asymptotic variance. Also, the asymptotic variance is invariant to nonsingular linear transformations, so that $\bar{A}(x) = CG(x)'\Omega(x)^{-1}$ will also minimize the asymptotic variance for any nonsingular constant matrix C.

This efficient instrument formula includes many important efficiency results as special cases. For example, for nonlinear weighted least squares it shows that the optimal weight is the inverse of the conditional variance of the residual: For $\hat{Q}_n(\theta) = -n^{-1}\sum_{i=1}^n w(x_i)[y_i - h(x_i,\theta)]^2$, the conclusion of Theorem 3.1 will give an asymptotic variance in eq. (5.10) with $A(x) = w(x)h_\theta(x,\theta_0)$, and the efficient estimator has $\bar{A}(x) = \{E[\varepsilon^2|x]\}^{-1}h_\theta(x,\theta_0)$, corresponding to weighting by the inverse of the conditional variance. This example also illustrates how efficiency in a class that does not satisfy assumptions (i) and (ii) of Theorem 5.3 (i.e. the linearity conditions), can be shown by enlarging the class: the set of scores (or moments) for weighted least squares estimators is not linear in the sense of assumption (i), but by also including variances for "instrumental variable" estimators, based on the moment conditions $g(z,\theta) = A(x)[y - h(x,\theta)]$, one obtains a class that includes weighted least squares, satisfies linearity, and has an efficient member given by a weighted least squares estimator. Of course, in a simple example like this one it is not necessary to check linearity, but in using eq. (5.8) to derive new efficiency results, it is a good idea to set up the class of estimators so that the linearity hypothesis is satisfied, and hence some solution to eq. (5.8) exists (when there is an efficient estimator).

Another example of optimal instrument variables is the well known result on efficiency of 2SLS in the class of instrumental variables estimators with possibly nonlinear instruments: If $\rho(z,\theta) = y - Y'\theta$, $E[Y|x] = \Pi x$, and $\sigma^2 = E[\rho(z,\theta_0)^2|x]$ is constant, then $G(x) = -\Pi x$ and $\Omega(x) = \sigma^2$, and the 2SLS instruments are $E[Yx'](E[xx'])^{-1}x = \Pi x = -\sigma^2\bar{A}(x)$, a nonsingular linear combination of $\bar{A}(x)$. As noted above, for efficiency it suffices that the instruments are a nonsingular linear combination of $\bar{A}(x)$, implying efficiency of 2SLS.

This general form $\bar{A}(x)$ for the optimal instruments has been previously derived in Chamberlain (1987), but here it serves to illustrate how eq. (5.8) can be used to

[39] Existence of the asymptotic variance matrix corresponding to τ_3 follows by the triangle and Cauchy–Schwartz inequalities.

derive the form of an optimal estimator. In this example, an optimal choice of estimator follows immediately from the form of eq. (5.8), and there is no need to guess what form the optimal instruments might take.

5.5. *Feasible efficient estimation*

In general, an efficient estimator can depend on nuisance parameters or functions. For example, in minimum distance estimation the efficient weighting matrix is a nuisance parameter that is unknown. Often there is a nuisance function, i.e. an infinite-dimensional nuisance parameter, such as the optimal instruments discussed in Section 5.4. The true value of these nuisance parameters is generally unknown, so that it is not feasible to use the true value to construct an efficient estimator. One feasible approach to efficient estimation is to use estimates in place of true nuisance parameters, i.e. to "plug-in" consistent nuisance parameter estimates, in the construction of the estimator. For example, an approach to feasible, optimal weighted least squares estimator is to maximize $-n^{-1}\sum_{i=1}^{n}\hat{w}(x_i)[y_i - h(x_i,\theta)]^2$, where $\hat{w}(x)$ is an estimator of $1/E[\varepsilon^2|x]$.

This approach will give an efficient estimator, if the estimation of the nuisance parameters does not affect the asymptotic variance of $\hat{\theta}$. It has already been shown, in Section 5.2, that this approach works for minimum distance estimation, where it suffices for efficiency that the weight matrix converges in probability to Ω^{-1}. More generally, a result developed in Section 6, on two-step estimators, suggests that estimation of the nuisance parameters should not affect efficiency. One can think of the "plug-in" approach to efficient estimation as a two-step estimator, where the first step is estimating the nuisance parameter or function, and the second is construction of $\hat{\theta}$. According to a principle developed in the next section, the first-step estimation has no effect on the second-step estimator if consistency of the first-step estimator does not affect consistency of the second. This principle generally applies to efficient estimators, where nuisance parameter estimates that converge to wrong values do not affect consistency of the estimator of parameters of interest. For example, consistency of the weighted least squares estimator is not affected by the form of the weights (as long as they satisfy certain regularity conditions). Thus, results on two-step estimation suggest that the "plug-in" approach should usually yield an efficient estimator.

The plug-in approach is often easy to implement when there are a finite number of nuisance parameters or when one is willing to assume that the nuisance function can be parametrized by a finite number of parameters. Finding a consistent estimator of the true nuisance parameters to be used in the estimator is often straightforward. A well known example is the efficient linear combination matrix $\Pi = E[Yx'](E[xx'])^{-1}$ for an instrumental variables estimator, which is consistently estimated by the 2SLS coefficients $\hat{\Pi} = \sum_{i=1}^{n} Y_i x_i'(\sum_{i=1}^{n} x_i x_i')^{-1}$. Another example is the optimal weight for nonlinear least squares. If the conditional variance is parametrized as $\sigma^2(x,\gamma)$, then

the true γ can be consistently estimated from the nonlinear least squares regression of $\tilde{\varepsilon}_i^2$ on $\sigma^2(x_i, \gamma)$, where $\tilde{\varepsilon}_i = y_i - h(x_i, \tilde{\theta})$, $(i = 1, \ldots, n)$, are the residuals from a preliminary consistent estimator $\tilde{\theta}$.

Of course, regularity conditions are useful for showing that estimation of the nuisance parameters does not affect the asymptotic variance of the estimator. To give a precise statement it is helpful to be more specific about the nature of the estimator. A quite general type of "plug-in" estimator is a GMM estimator that depends on preliminary estimates of some parameters. Let $g(z, \theta, \gamma)$ denote a $q \times 1$ vector of functions of the parameters of interest and nuisance parameters γ, and let $\hat{\gamma}$ be a first-step estimator. Consider an estimator $\hat{\theta}$ that, with probability approaching one, solves

$$n^{-1} \sum_{i=1}^{n} g(z_i, \theta, \hat{\gamma}) = 0. \tag{5.12}$$

This class is quite general, because eq. (5.12) can often be interpreted as the first-order conditions for an estimator. For example, it includes weighted least squares estimators with an estimated weight $w(x, \hat{\gamma})$, for which eq. (5.12) is the first-order condition with $g(z, \theta, \gamma) = w(x, \gamma) h_\theta(x, \theta)[y - h(x, \theta)]$. One type of estimator not included is CMD, but the main result of interest here is efficient choice of weighting matrix, as already discussed in Section 5.2.

Suppose also that $\hat{\gamma}$ is a GMM estimator, satisfying $n^{-1} \sum_{i=1}^{n} m(z_i, \gamma) = 0$. If this equation is "stacked" with eq. (5.12), the pair $(\hat{\theta}, \hat{\gamma})$ becomes a joint GMM estimator, so that regularity conditions for asymptotic efficiency can be obtained from the assumptions for Theorem 3.4. This result, and its application to more general types of two-step estimators, is described in Section 6. In particular, Theorem 6.1 can be applied to show that $\hat{\theta}$ from eq. (5.12) is efficient. If the hypotheses of that result are satisfied and $G_\gamma = E[\nabla_\gamma g(z, \theta_0, \gamma_0)] = 0$ then $\hat{\theta}$ will be asymptotically normal with asymptotic variance the same as if $\hat{\gamma} = \gamma_0$. As further discussed in Section 6, the condition $G_\gamma = 0$ is related to the requirement that consistency of $\hat{\gamma}$ not affect consistency of $\hat{\theta}$. As noted above, this condition is a useful one for determining whether the estimation of the nuisance parameters affects the asymptotic variance of the feasible estimator $\hat{\theta}$.

To show how to analyze particular feasible estimators, it is useful to give an example.

Linear regression with linear heteroskedasticity: Consider a linear model where $E[y|x] = x'\theta_0$ and $\sigma^2(x) = \mathrm{Var}(y|x) = w'\alpha_0$ for some $w = w(x)$ that is a function of x. As noted above, the efficient estimator among those that solve $n^{-1} \sum_{i=1}^{n} A(x_i) \times [y_i - x_i'\theta] = 0$ has $A(x) = \bar{A}(x) = (w'\alpha_0)^{-1} x$. A feasible efficient estimator can be constructed by using a squared residual regression to form an estimator $\hat{\alpha}$ for α_0, and plugging this estimator into the first-order conditions. More precisely, let $\hat{\beta}$ be the least squares estimator from a regression of y on x and $\hat{\alpha}$ the least squares

estimator from a regression of $(y - x'\hat{\beta})^2$ on w. Suppose that $w'\alpha_0$ is bounded below and let $\tau(v)$ be a positive function that is continuously differentiable with bounded derivative and $\tau(v) = v$ for v greater than the lower bound on $w'\alpha_0$.[40] Consider $\hat{\theta}$ obtained from solving $\sum_{i=1}^{n} \tau(w'\hat{\alpha})^{-1} x_i(y_i - x_i'\theta) = 0$. This estimator is a two-step GMM estimator like that given above with

$$\gamma = (\alpha', \beta')', \; m(z, \gamma) = [(y - x'\beta)x', \{(y - x'\beta)^2 - w'\alpha\}w']',$$
$$g(z, \theta, \gamma) = \tau(w'\alpha)^{-1} x(y - x'\theta).$$

It is straightforward to verify that the vector of moment functions $[m(z, \gamma)', g(z, \theta, \gamma)']'$ satisfies the conditions of Theorem 6.1 if w is bounded, x and y have finite fourth moments, and $E[xx']$ and $E[ww']$ are nonsingular. Furthermore, $E[\nabla_\gamma g(z, \theta_0, \gamma_0)] = -E[\tau(w'\alpha_0)^{-2}(y - x'\theta_0)xw'] = 0$, so that this feasible estimator will be efficient.

In many cases the efficiency of a "plug-in" estimator may be adversely affected if the parametrization of the nuisance functions is incorrect. For example, if in a linear model, heteroskedasticity is specified as exponential, but the true conditional variance takes another form, then the weighted least squares estimator based on an exponential variance function will not be efficient. Consistency will generally not be affected, and there will be only a little loss in efficiency if the parametrization is approximately correct, but there could be big efficiency losses if the parametrized functional form is far from the true one. This potential problem with efficiency suggests that one might want to use *nonparametric* nuisance function estimators, that do not impose any restrictions on functional form. For the same reasons discussed above, one would expect that estimation of the nuisance function does not affect the limiting distribution, so that the resulting feasible estimators would be efficient. Examples of this type of approach are Stone (1975), Bickel (1982), and Carroll (1982). These estimators are quite complicated, so an account is not given here, except to say that similar estimators are discussed in Section 8.

5.6. *Technicalities*

It is possible to show the generalized information matrix equality in eq. (5.1) under a condition that allows for $f(z|\theta)^{1/2}$ to not be continuously differentiable and $g(z, \theta)$ to not be continuous. For the root-density, this condition is "mean-square" differentiability at θ_0 with respect to integration over z, meaning that there is $\delta(z)$ with $\int \|\delta(z)\|^2 \, dz < \infty$ such that $\int [f(z|\theta)^{1/2} - f(z|\theta_0)^{1/2} - \delta(z)'(\theta - \theta_0)]^2 \, dz = o(\|\theta - \theta_0\|^2)$

[40] The $\tau(v)$ function is a "trimming" device similar to those used in the semiparametric estimation literature. This specification requires knowing a lower bound on the conditional variance. It is also possible to allow $\tau(v)$ to approach the identity for all $v > 0$ as the sample size grows, but this would complicate the analysis.

as $\theta \to \theta_0$. As shown in Bickel et al. (1992), it will suffice for this condition that $f(z|\theta)$ is continuously differentiable in θ (for almost all z) and that $J(\theta) = \int \nabla_\theta \ln f(z|\theta) \times \{\nabla_\theta \ln f(z|\theta)\}' f(z|\theta)\, dz$ is nonsingular and continuous in θ. Here $\delta(z)$ is the derivative of $f(z|\theta)^{1/2}$, so by $\nabla_0 f(z|\theta)^{1/2} = \frac{1}{2} f(z|\theta)^{1/2} \nabla_\theta \ln f(z|\theta)$, the expression for the information matrix in terms of $\delta(z)$ is $J = 4\int \delta(z)\delta(z)'\, dz$. A precise result on efficiency of MLE in the class of GMM estimators can then be stated as:

Lemma 5.4

If (i) $f(z|\theta)^{1/2}$ is mean-square differentiable at θ_0 with derivative $\delta(z)$; (ii) $E[g(z,\theta)]$ is differentiable at θ_0 with derivative G; (iii) $g(z,\theta)$ is continuous at θ_0 with probability one; (iv) there is a neighborhood $\mathcal{N}$ of θ_0 and a function $d(z)$ such that $\| g(z,\theta) \| \leq d(z)$ and $\int d(z)^2 f(z|\theta)\, dz$ is bounded for $\theta \in \mathcal{N}$; then $\int g(z,\theta) f(z|\theta)\, dz$ is differentiable at θ_0 with derivative $G + 2\int g(z,\theta_0)\delta(z) f(z|\theta_0)^{1/2}\, dz$.

Proof

The proof is similar to that of Lemma 7.2 of Ibragimov and Has'minskii (1981). Let $r(\theta) = f(z|\theta)^{1/2}$, $g(\theta) = g(z,\theta)$, $\delta = \delta(z)$, and $\Delta(\theta) = r(\theta) - r(\theta_0) - \delta'(\theta - \theta_0)$, suppressing the z argument for notational convenience. Also, let $m(\tilde{\theta}, \theta) = \int g(\tilde{\theta}) r(\theta)^2\, dz$ and $M = \int g(\theta_0)\delta r(\theta_0)\, dz$. By (ii), $m(\theta,\theta_0) - m(\theta_0,\theta_0) - G(\theta - \theta_0) = o(\| \theta - \theta_0 \|)$. Also, by the triangle inequality, $\| m(\theta,\theta) - m(\theta_0,\theta_0) - (G + 2M)(\theta - \theta_0) \| \leq \| m(\theta,\theta_0) - m(\theta_0,\theta_0) - G(\theta - \theta_0) \| + \| m(\theta,\theta) - m(\theta,\theta_0) - 2M(\theta - \theta_0) \|$, so that to show the conclusion it suffices to show $\| m(\theta,\theta) - m(\theta,\theta_0) - 2M(\theta - \theta_0) \| = o(\| \theta - \theta_0 \|)$. To show this, note by the triangle inequality,

$$\begin{aligned}
\| m(\theta,\theta) - m(\theta,\theta_0) - 2M(\theta - \theta_0) \| &= \left\| \int g(\theta)[r(\theta)^2 - r(\theta_0)^2]\, dz - 2M(\theta - \theta_0) \right\| \\
&\leq \left\| \int g(\theta)[r(\theta) + r(\theta_0)]\Delta(\theta)\, dz \right\| + \left\| \int [g(\theta) - g(\theta_0)] r(\theta_0)\delta'\, dz \right\| \| \theta - \theta_0 \| \\
&\quad + \left\| \int [g(\theta)r(\theta) - g(\theta_0)r(\theta_0)]\delta'\, dz \right\| \| \theta - \theta_0 \| = R_1 + R_2 \| \theta - \theta_0 \| + R_3 \| \theta - \theta_0 \|.
\end{aligned}$$

Therefore, it suffices to show that $R_1 = o(\| \theta - \theta_0 \|)$, $R_2 \to 0$, and $R_3 \to 0$ as $\theta \to \theta_0$. By (iv) and the triangle and Cauchy–Schwartz inequalities,

$$\begin{aligned}
R_1 &\leq \left\{ \left[\int g(\theta)^2 r(\theta)^2\, dz \right]^{1/2} + \left[\int g(\theta)^2 r(\theta_0)^2\, dz \right]^{1/2} \right\} \left[\int \Delta(\theta)^2\, dz \right]^{1/2} \\
&\leq \left\{ \left[\int d(z)^2 r(\theta)^2\, dz \right]^{1/2} + \left[\int d(z)^2 r(\theta_0)^2\, dz \right]^{1/2} \right\} o(\| \theta - \theta_0 \|) = o(\| \theta - \theta_0 \|).
\end{aligned}$$

Also, by (iii) and (iv) and the dominated convergence theorem, $E[\| g(\theta) - g(\theta_0) \|^2] \to 0$,

so by the Cauchy–Schwartz inequality, $R_2 \leq (E[\| g(\theta) - g(\theta_0) \|^2])^{1/2} (\int \| \delta \|^2 \, dz)^{1/2} \to 0$. Also, by the triangle inequality, $R_3 \leq R_2 + \int \| g(\theta) \| |r(\theta) - r(\theta_0)| \| \delta \| \, dz$, while for $K > 0$,

$$\int \| g(\theta) \| |r(\theta) - r(\theta_0)| \| \delta \| \, dz \leq \int d(z) |r(\theta) - r(\theta_0)| \| \delta \| \, dz$$

$$\leq \int_{d(z) \geq K} d(z) |r(\theta) - r(\theta_0)| \| \delta \| \, dz + K \int |r(\theta) - r(\theta_0)| \| \delta \| \, dz$$

$$\leq \left\{ \int d(z)^2 |r(\theta) - r(\theta_0)|^2 \, dz \right\}^{1/2} \left\{ \int_{d(z) \geq K} \| \delta \|^2 \, dz \right\}$$

$$+ K \left\{ \int |r(\theta) - r(\theta_0)|^2 \, dz \right\}^{1/2} \left\{ \int \| \delta \|^2 \, dz \right\}^{1/2}.$$

By (iv), $\int d(z)^2 |r(\theta) - r(\theta_0)|^2 \, dz \leq 2 \int d(z)^2 r(\theta)^2 \, dz + 2 \int d(z)^2 r(\theta_0)^2 \, dz$ is bounded. Also, by the dominated convergence theorem, $\int_{d(z) \geq K} \| \delta \|^2 \, dz \to 0$ as $K \to \infty$, and by (i), $\int |r(\theta) - r(\theta_0)|^2 \, dz \to 0$, so that the last term converges to zero for any K. Consider $\varepsilon > 0$ and choose K so $\int_{d(z) \geq K} \| \delta \|^2 \, dz < \frac{1}{2} \varepsilon$. Then by the last term is less than $\frac{1}{2} \varepsilon$ for θ close enough to θ_0, implying that $\int \| g(\theta) \| |r(\theta) - r(\theta_0)| \| \delta \| \, dz < \varepsilon$ for θ close enough to θ_0. The conclusion then follows by the triangle inequality. Q.E.D.

Proof of Theorem 5.1

By condition (iv) of Theorem 3.4 and Lemma 3.5, $g(z, \theta)$ is continuous on a neighborhood of θ_0 and $E[g(z, \theta)]$ is differentiable at θ_0 with derivative $G = E[\nabla_\theta g(z, \theta_0)]$. Also, $f(z|\theta)^{1/2}$ is mean-square differentiable by the dominance condition in Theorem 5.1, as can be shown by the usual mean-value expansion argument. Also, by the conditions of Theorem 5.1, the derivative is equal to $\frac{1}{2} 1[f(z|\theta_0) > 0] f(z|\theta_0)^{-1/2} \times \nabla_\theta f(z|\theta_0)$ on a set of full measure, so that the derivative in the conclusion of Lemma 5.4 is $G + E[g(z, \theta_0) \nabla_\theta \ln f(z|\theta_0)]$. Also, $\| g(z, \theta) \| \leq d(z) = \sup_{\theta \in \mathcal{N}} \| g(z, \theta) \|$ has $\int d(z)^2 f(z|\theta) \, dz$ bounded, so that the conclusion of Lemma 5.4 holds. Then for $u = g(z, \theta_0) + G J^{-1} \nabla_\theta \ln f(z|\theta_0)$,

$$(G'WG)^{-1} G'W \Omega WG (G'WG)^{-1} - J^{-1}$$

$$= (G'WG)^{-1} G'W \left(\int uu' \, dz \right) WG (G'WG)^{-1},$$

so the conclusion follows by $\int uu' \, dz$ positive semi-definite. Q.E.D.

6. Two-step estimators

A two-step estimator is one that depends on some preliminary, "first-step" estimator of a parameter vector. They provide a useful illustration of how the previous results

can be applied, even to complicated estimators. In particular, it is shown in this section that two-step estimators can be fit into the GMM framework. Two-step estimators are also of interest in their own right. As discussed in Section 5, feasible efficient estimators often are two-step estimators, with the first step being the estimation of nuisance parameters that affect efficiency. Also, they provide a simpler alternative to complicated joint estimators. Examples of two-step estimators in econometrics are the Heckman (1976) sample selection estimator and the Barro (1977) estimator for linear models that depend on expectations and/or corresponding residuals. Their properties have been analyzed by Newey (1984) and Pagan (1984, 1986), among others.

An important question for two-step estimators is whether the estimation of the first step affects the asymptotic variance of the second, and if so, what effect does the first step have. Ignoring the first step can lead to inconsistent standard error estimates, and hence confidence intervals that are not even asymptotically valid. This section develops a simple condition for whether the first step affects the second, which is that an effect is present if and only if consistency of the first-step estimator affects consistency of the second-step estimator. This condition is useful because one can often see by inspection whether first-step inconsistency leads to the second-step inconsistency. This section also describes conditions for ignoring the first step to lead to either an underestimate or an overestimate of the standard errors.

When the variance of the second step is affected by the estimation in the first step, asymptotically valid standard errors for the second step require a correction for the first-step estimation. This section derives consistent standard error estimators by applying the general GMM formula. The results are illustrated by a sample selection model.

The efficiency results of Section 5 can also be applied, to characterize efficient members of some class of two-step estimators. For brevity these results are given in Newey (1993) rather than here.

6.1. *Two-step estimators as joint GMM estimators*

The class of GMM estimators is sufficiently general to include two-step estimators where moment functions from the first step and the second step can be "stacked" to form a vector of moment conditions. Theorem 3.4 can then be applied to specify regularity conditions for asymptotic normality, and the conclusion of Theorem 3.4 will provide the asymptotic variance, which can then be analyzed to derive the results described above. Previous results can also be used to show consistency, which is an assumption for the asymptotic normality results, but to focus attention on the most interesting features of two-step estimators, consistency will just be assumed in this section.

A general type of estimator $\hat{\theta}$ that has as special cases most examples of interest is one that, with probability approaching one, solves an equation

$$n^{-1}\sum_{i=1}^{n} g(z_i, \theta, \hat{\gamma}) = 0, \tag{6.1}$$

where $g(z, \theta, \gamma)$ is a vector of functions with the same dimension as θ and $\hat{\gamma}$ is a first-step estimator. This equation is exactly the same as eq. (5.12), but here the purpose is analyzing the asymptotic distribution of $\hat{\theta}$ in general rather than specifying regularity conditions for $\hat{\gamma}$ to have no effect. The estimator can be treated as part of a joint GMM estimator if $\hat{\gamma}$ also satisfies a moment condition of the form, with probability approaching one,

$$n^{-1}\sum_{i=1}^{n} m(z_i, \gamma) = 0, \tag{6.2}$$

where $m(z, \gamma)$ is a vector with the same dimension as γ. If $g(z, \theta, \gamma)$ and $m(z, \gamma)$ are "stacked" to form $\tilde{g}(z, \theta, \gamma) = [m(z, \theta)', g(z, \theta, \gamma)']'$, then eqs. (6.1) and (6.2) are simply the two components of the joint moment equation $n^{-1}\sum_{i=1}^{n} \tilde{g}(z_i, \hat{\theta}, \hat{\gamma}) = 0$. Thus, the two-step estimator from eq. (6.1) can be viewed as a GMM estimator.

An interesting example of a two-step estimator that fits into this framework is Heckman's (1976) sample selection estimator.

Sample selection example: In this example the first step $\hat{\gamma}$ is a probit estimator with regressors x. The second step is least squares regression in the subsample where the probit-dependent variable is one, i.e. in the selected sample, with regressors given by w and $\lambda(x'\hat{\gamma})$ for $\lambda(v) = \phi(v)/\Phi(v)$. Let d be the probit-dependent variable, that is equal to either zero or one. This estimator is useful when y is only observed if $d = 1$, e.g. where y is wages and d is labor force participation. The idea is that joint normality of the regression $y = w'\beta_0 + u$ and the probit equation leads to $E[y|w, d = 1, x] = w'\beta_0 + \alpha_0\lambda(x'\gamma_0)$, where α_0 is nonzero if the probit- and regression-dependent variables are not independent. Thus, $\lambda(x'\alpha_0)$ can be thought of as an additional regressor that corrects for the endogenous subsample.

This two-step estimator will satisfy eqs. (6.1) and (6.2) for

$$g(z, \theta, \gamma) = d\begin{bmatrix} w \\ \lambda(x'\gamma) \end{bmatrix}[y - w'\beta - \alpha\lambda(x'\gamma)],$$

$$m(z, \gamma) = \lambda(x'\gamma)\Phi^{-1}(-x'\gamma)x[d - \Phi(x'\gamma)], \tag{6.3}$$

where $\theta = (\beta', \alpha)'$. Then eq. (6.1) becomes the first-order condition for least squares on the selected sample and eq. (6.2) the first-order condition for probit.

Regularity conditions for asymptotic normality can be formulated by applying the asymptotic normality result for GMM, i.e. Theorem 3.4, to the stacked vector of moment conditions. Also, the conclusion of Theorem 3.4 and partitioned inversion can then be used to calculate the asymptotic variance of $\hat{\theta}$, as in the following result. Let

$$G_\theta = E[\nabla_\theta g(z, \theta_0, \gamma_0)], \qquad G_\gamma = E[\nabla_\gamma g(z, \theta_0, \gamma_0)], \qquad g(z) = g(z, \theta_0, \gamma_0),$$
$$M = E[\nabla_\gamma m(z, \gamma_0)], \qquad \psi(z) = -M^{-1} m(z, \gamma_0). \tag{6.4}$$

Theorem 6.1

If eqs. (6.1) and (6.2) are satisfied with probability approaching one, $\hat{\theta} \xrightarrow{p} \theta_0$, $\hat{\gamma} \xrightarrow{p} \gamma_0$, and $\tilde{g}(z, \theta, \gamma)$ satisfies conditions (i)–(v) of Theorem 3.4, then $\hat{\theta}$ and $\hat{\gamma}$ are asymptotically normal and $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$ where $V = G_\theta^{-1} E[\{g(z) + G_\gamma \psi(z)\}\{g(z) + G_\gamma \psi(z)\}'] G_\theta^{-1\prime}$.

Proof

By eqs. (6.1) and (6.2), with probability approaching one $(\hat{\theta}, \hat{\gamma})$ is a GMM estimator with moment function $\tilde{g}(z, \theta, \gamma) = [m(z, \gamma)', g(z, \theta, \gamma)']'$ and $\hat{W}$ equal to an identity matrix. By $(\tilde{G}' I \tilde{G})^{-1} \tilde{G}' = \tilde{G}^{-1}$, the asymptotic variance of the estimator is $(\tilde{G}' I \tilde{G})^{-1} \tilde{G}' I E[\tilde{g}(z, \theta_0, \gamma_0) \tilde{g}(z, \theta_0, \gamma_0)'] I \tilde{G} (\tilde{G}' I \tilde{G})^{-1} = \tilde{G}^{-1} E[\tilde{g}(z, \theta_0, \gamma_0) \tilde{g}(z, \theta_0, \gamma_0)'] \tilde{G}^{-1\prime}$. Also, the expected Jacobian matrix and its inverse are given by

$$\tilde{G} = E[\partial \tilde{g}(z, \theta_0, \gamma_0) / \partial(\theta', \gamma')'] = \begin{bmatrix} G_\theta & G_\gamma \\ 0 & M \end{bmatrix}, \quad \tilde{G}^{-1} = \begin{bmatrix} G_\theta^{-1} & -G_\theta^{-1} G_\gamma M^{-1} \\ 0 & M^{-1} \end{bmatrix}. \tag{6.5}$$

Noting that the first row of $\tilde{G}^{-1}$ is $G_\theta^{-1}[I, -G_\gamma M^{-1}]$ and that $[I, -G_\gamma M^{-1}] \times \tilde{g}(z, \theta_0, \gamma_0) = g(z) + G_\gamma \psi(z)$, the asymptotic variance of $\hat{\theta}$, which is the upper left block of the joint variance matrix, follows by partitioned matrix multiplication. Q.E.D.

An alternative approach to deriving the asymptotic distribution of two-step estimators is to work directly from eq. (6.1), expanding in θ to solve for $\sqrt{n}(\hat{\theta} - \theta_0)$ and then expanding the result around the true γ_0. To describe this approach, first note that $\hat{\gamma}$ is an asymptotically linear estimator with influence function $\psi(z) = -M^{-1} m(z_i, \gamma_0)$, where $\sqrt{n}(\hat{\gamma} - \gamma_0) = \sum_{i=1}^n \psi(z_i)/\sqrt{n} + o_p(1)$. Then expanding the left-hand side of eq. (6.1) around θ_0 and solving gives:

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\left[n^{-1} \sum_{i=1}^n \nabla_\theta g(z_i, \bar{\theta}, \hat{\gamma}) \right]^{-1} \sum_{i=1}^n g(z_i, \theta_0, \hat{\gamma}) / \sqrt{n}$$
$$= -\left[n^{-1} \sum_{i=1}^n \nabla_\theta g(z_i, \bar{\theta}, \hat{\gamma}) \right]^{-1}$$

$$\times \left\{ \sum_{i=1}^{n} g(z_i)/\sqrt{n} + \left[n^{-1} \sum_{i=1}^{n} \nabla_\gamma g(z_i, \theta_0, \bar{\gamma}) \right] \sqrt{n}(\hat{\gamma} - \gamma_0) \right\}$$

$$= -G_\theta^{-1} \sum_{i=1}^{n} \{ g(z_i) + G_\gamma \psi(z_i) \}/\sqrt{n} + o_p(1), \tag{6.6}$$

where $\bar{\theta}$ and $\bar{\gamma}$ are mean values and the third equality follows by convergence of $\hat{\gamma}$ and the mean values and the conclusion of Lemma 2.4. The conclusion then follows by applying the central limit theorem to the term following the last equality.

One advantage of this approach is that it only uses the influence function representation $\sqrt{n}(\hat{\gamma} - \gamma_0) = \sum_{i=1}^{n} \psi(z_i)/\sqrt{n} + o_p(1)$ for $\hat{\gamma}$, and not the GMM formula in eq. (6.2). This generalization is useful when $\hat{\gamma}$ is not a GMM estimator. The GMM approach has been adopted here because it leads to straightforward primitive conditions, while an influence representation for $\hat{\gamma}$ is not a very primitive condition. Also the GMM approach can be generalized to allow $\hat{\gamma}$ to be a two-step, or even multistep, estimator by stacking moment conditions for estimators that affect $\hat{\gamma}$ with the moment conditions for θ and γ.

6.2. *The effect of first-step estimation on second-step standard errors*

One important feature of two-step estimators is that ignoring the first step in calculating standard errors can lead to inconsistent standard errors for the second step. The asymptotic variance for the estimator solving eq. (6.1) with $\hat{\gamma} = \gamma_0$, i.e. the asymptotic variance ignoring the presence of $\hat{\gamma}$ in the first stage, is $G_\theta^{-1} E[g(z)g(z)']G_\theta^{-1\prime}$. In general, this matrix differs from the asymptotic variance given in the conclusion of Theorem 6.1, because it does not account for the presence of the first-step estimators.

Ignoring the first step will be valid if $G_\gamma = 0$. Also, if $G_\gamma \neq 0$, then ignoring the first step will generally be invalid, leading to an incorrect asymptotic variance formula, because nonzero G_γ means that, except for unusual cases, $E[g(z)g(z)']$ will not equal $E[\{g(z) + G_\gamma \psi(z)\}\{g(z) + G_\gamma \psi(z)\}']$. Thus, the condition for estimation of the first step to have no effect on the second-step asymptotic variance is $G_\gamma = 0$.

A nonzero G_γ can be interpreted as meaning that inconsistency in the first-step estimator leads to inconsistency in the second-step estimator. This interpretation is useful, because it gives a comparatively simple criterion for determining if first-stage estimation has to be accounted for. To derive this interpretation, consider the solution $\theta(\gamma)$ to $E[g(z, \theta(\gamma), \gamma)] = 0$. Because $\hat{\theta}$ satisfies the sample version of this condition, $\theta(\gamma)$ should be the probability limit of the second-step estimator when $\hat{\gamma}$ converges to γ (under appropriate regularity conditions, such as those of Section 2). Assuming differentiation inside the expectation is allowed, the implicit function theorem gives

$$\nabla_\gamma \theta(\gamma_0) = -G_\theta^{-1} G_\gamma. \tag{6.7}$$

By nonsingularity of G_θ, the necessary and sufficient condition for $G_\gamma = 0$ is that $\nabla_\gamma \theta(\gamma_0) = 0$. Since $\theta(\gamma_0) = \theta_0$, the condition that $\nabla_\gamma \theta(\gamma_0) = 0$ is a local, first-order condition that inconsistency in $\hat{\gamma}$ does not affect consistency of $\hat{\theta}$. The following result adds regularity conditions for this first-order condition to be interpreted as a consistency condition.

Theorem 6.2

Suppose that the conditions of Theorem 6.1 are satisfied and $g(z, \theta, \gamma)$ satisfies the conditions of Lemma 2.4 for the parameter vector (θ', γ'). If $\hat{\theta} \xrightarrow{p} \theta_0$ even when $\hat{\gamma} \to \gamma \neq \gamma_0$, then $G_\gamma = 0$. Also suppose that $E[\nabla_\gamma g(z, \theta_0, \gamma)]$ has constant rank on a neighborhood of γ_0. If for any neighborhood of γ_0 there is γ in that neighborhood such that $\hat{\theta}$ does not converge in probability to θ_0 when $\hat{\gamma} \xrightarrow{p} \gamma$, then $G_\gamma \neq 0$.

Proof

By Lemma 2.4, $\hat{\theta} \xrightarrow{p} \theta_0$ and $\hat{\gamma} \xrightarrow{p} \gamma$ imply that $\sum_{i=1}^n g(z_i, \hat{\theta}, \hat{\gamma})/n \xrightarrow{p} E[g(z, \theta_0, \gamma)]$. The sample moment conditions (6.1) thus imply $E[g(z, \theta_0, \gamma)] = 0$. Differentiating this identity with respect to γ at $\gamma = \gamma_0$ gives $G_\gamma = 0$.[41] To show the second conclusion, let $\theta(\gamma)$ denote the limit of $\hat{\theta}$ when $\hat{\gamma} \xrightarrow{p} \gamma$. By the previous argument, $E[g(z, \theta(\gamma), \gamma)] = 0$. Also, by the implicit function theorem $\theta(\gamma)$ is continuous at γ_0, with $\theta(\gamma_0) = \theta_0$. By the conditions of Theorem 6.1, $G_\theta(\theta, \gamma) = E[\nabla_\theta g(z, \theta, \gamma)]$ is continuous in a neighborhood of θ_0 and γ_0, and so will be nonsingular on a small enough neighborhood by G_θ nonsingular. Consider a small enough convex neighborhood where this nonsingularity condition holds and $E[\nabla_\gamma g(z, \theta_0, \gamma)]$ has constant rank. A mean-value expansion gives $E[g(z, \theta_0, \gamma)] = E[g(z, \theta(\gamma), \gamma)] + G_\theta(\bar{\theta}, \gamma)[\theta_0 - \theta(\gamma)] \neq 0$. Another expansion then gives $E[g(z, \theta_0, \gamma)] = E[\nabla_\gamma g(z, \theta_0, \bar{\gamma})](\gamma - \gamma_0) \neq 0$, implying $E[\nabla_\gamma g(z, \theta_0, \bar{\gamma})] \neq 0$, and hence $G_\gamma \neq 0$ (by the derivative having constant rank).

Q.E.D.

This results states that, under certain regularity conditions, the first-step estimator affects second-step standard errors, i.e. $G_\gamma \neq 0$, if and only if inconsistency in the first step leads to inconsistency in the second step. The sample selection estimator provides an example of how this criterion can be applied.

Sample selection continued: The second-step estimator is a regression where some of the regressors depend on γ. In general, including the wrong regressors leads to inconsistency, so that, by Theorem 6.2, the second-step standard errors will be affected by the first step. One special case where the estimator will still be consistent is if $\alpha_0 = 0$, because including a regressor that does not belong does not affect consistency. Thus, by Theorem 6.2, no adjustment is needed (i.e. $G_\gamma = 0$) if $\alpha_0 = 0$. This result is useful for constructing tests of whether these regressors belong, because

[41] Differentiation inside the expectation is allowed by Lemma 3.6.

it means that under the null hypothesis the test that ignores the first stage will have asymptotically correct size. These results can be confirmed by calculating

$$G_\gamma = -\alpha_0 E\left[\mathrm{d}\begin{bmatrix} w \\ \lambda(x'\gamma_0) \end{bmatrix} \lambda_v(x'\gamma_0)x' \right],$$

where $\lambda_v(v) = \mathrm{d}\lambda(v)/\mathrm{d}v$. By inspection this matrix is generally nonzero, but is zero if $\alpha_0 = 0$.

This criterion can also be applied to subsets of the second-step coefficients. Let S denote a selection matrix such that SA is a matrix of rows of A, so that $S\hat{\theta}$ is a subvector of the second-step coefficients. Then the asymptotic variance of $S\hat{\theta}$ is $SG_\theta^{-1}E[\{g(z) + G_\gamma\psi(z)\}\{g(z) + G_\gamma\psi(z)\}']G_\theta^{-1}S'$, while the asymptotic variance that ignores the first step is $SG_\theta^{-1}E[g(z)g(z)']G_\theta^{-1}S'$. The general condition for equality of these two matrices is

$$0 = -SG_\theta^{-1}G_\gamma = S\nabla_\gamma\theta(\gamma_0) = \nabla_\gamma[S\theta(\gamma_0)], \tag{6.8}$$

where the second equality follows by eq. (6.7). This is a first-order version of the statement that asymptotic variance of $S\hat{\theta}$ is affected by the first-step estimator if and only if consistency of the first step affects consistency of the second. This condition could be made precise by modifying Theorem 6.2, but for simplicity this modification is not given here.

Sample selection continued: As is well known, if the correct and incorrect regressors are independent of the other regressors then including the wrong regressor only affects consistency of the coefficient of the constant. Thus, the second-step standard errors of the coefficients of nonconstant variables in w will not be affected by the first-step estimation if w and x are independent.

One can also derive conditions for the correct asymptotic variance to be larger or smaller than the one that ignores the first step. A condition for the correct asymptotic variance to be larger, given in Newey (1984), is that the first- and second-step moment conditions are uncorrelated, i.e.

$$E[g(z,\theta_0,\gamma_0)m(z,\gamma_0)'] = 0. \tag{6.9}$$

In this case $E[g(z)\psi(z)'] = 0$, so the correct variance is $G_\theta^{-1}E[g(z)g(z)']G_\theta^{-1\prime} + G_\theta^{-1}G_\gamma E[\psi(z)\psi(z)']G_\gamma' G_\theta^{-1\prime}$, which is larger, in the positive semi-definite sense, than the one $G_\theta^{-1}E[g(z)g(z)']G_\theta^{-1\prime}$ that ignores first-step estimation.

Sample selection continued: In this example, $E[y - w'\beta_0 - \alpha_0\lambda(x'\gamma_0)|w, d = 1, x] = 0$, which implies (6.9). Thus, the standard error formula that ignores the first-step estimation will understate the asymptotic standard error.

A condition for the correct asymptotic variance to be smaller than the one that ignores the first step, given by Pierce (1982), is that

$$m(z) = m(z, \gamma_0) = \nabla_\gamma \ln f(z|\theta_0, \gamma_0). \tag{6.10}$$

In this case, the identities $\int m(z, \gamma) f(z|\theta_0, \gamma)\, dz = 0$ and $\int g(z, \theta_0, \gamma) f(z|\theta_0, \gamma)\, dz = 0$ can be differentiated to obtain the generalized information matrix equalities $M = -E[s(z)s(z)']$ and $G_\gamma = -E[g(z)s(z)']$. It then follows that $G_\gamma = -E[g(z)m(z)'] = -E[g(z)\psi(z)'][E[\psi(z)\psi(z)']]^{-1}$, so that the correct asymptotic variance is $G_\theta^{-1}E[g(z)g(z)']G_\theta^{-1} - G_\theta^{-1}E[g(z)\psi(z)'][E[\psi(z)\psi(z)']]^{-1}E[\psi(z)g(z)']G_\theta^{-1\prime}$. This variance is smaller, in the positive semi-definite sense, than the one that ignores the first step.

Equation (6.10) is a useful condition, because it implies that conservative asymptotic confidence intervals can be constructed by ignoring the first stage. Unfortunately, the cases where it is satisfied are somewhat rare. A necessary condition for eq. (6.10) is that the information matrix for θ and γ be block diagonal, because eq. (6.10) implies that the asymptotic variance of $\hat{\gamma}$ is $[E[m(z)m(z)']]^{-1}$, which is only obtainable when the information matrix is block diagonal. Consequently, if $g(z, \theta, \gamma)$ were the score for θ, then $G_\gamma = 0$ by the information matrix equality, and hence estimation of $\hat{\gamma}$ would have no effect on the second-stage variance. Thus, eq. (6.10) only leads to a lowering of the variance when $g(z, \theta, \gamma)$ is not the score, i.e. $\hat{\theta}$ is not an efficient estimator.

One case where eq. (6.10) holds is if there is a factorization of the likelihood $f(z|\theta, \gamma) = f_1(z|\theta)f_2(z|\gamma)$ and $\hat{\gamma}$ is the MLE of γ. In particular, if $f_1(z|\theta)$ is a conditional likelihood and $f_2(z|\gamma) = f_2(x|\gamma)$ a marginal likelihood of variables x, i.e. x are ancillary to θ, then eq. (6.8) is satisfied when $\hat{\gamma}$ is an efficient estimator of γ_0.

6.3. *Consistent asymptotic variance estimation for two-step estimators*

The interpretation of a two-step estimator as a joint GMM estimator can be used to construct a consistent estimator of the asymptotic variance when $G_\gamma \neq 0$, by applying the general GMM formula. The Jacobian terms can be estimated by sample Jacobians, i.e. as

$$\hat{G}_\theta = n^{-1}\sum_{i=1}^{n} \nabla_\theta g(z_i, \hat{\theta}, \hat{\gamma}), \quad \hat{G}_\gamma = n^{-1}\sum_{i=1}^{n} \nabla_\gamma g(z_i, \hat{\theta}, \hat{\gamma}), \quad \hat{M} = n^{-1}\sum_{i=1}^{n} \nabla_\gamma m(z_i, \hat{\gamma}).$$

The second-moment matrix can be estimated by a sample second-moment matrix

$\hat{g}_i = g(z_i, \hat{\theta}, \hat{\gamma})$ and $\hat{m}_i = m(z_i, \hat{\gamma})$, of the form $\hat{\Omega} = n^{-1}\sum_{i=1}^{n}(\hat{g}_i', \hat{m}_i')'(\hat{g}_i', \hat{m}_i')$. An estimator of the joint asymptotic variance of $\hat{\theta}$ and $\hat{\gamma}$ is then given by

$$\hat{V} = \begin{bmatrix} \hat{G}_\theta & \hat{G}_\gamma \\ 0 & \hat{M} \end{bmatrix}^{-1} \hat{\Omega} \begin{bmatrix} \hat{G}_\theta & \hat{G}_\gamma \\ 0 & \hat{M} \end{bmatrix}^{-1\prime}$$
$$= \begin{bmatrix} \hat{G}_\theta^{-1} & -\hat{G}_\theta^{-1}\hat{G}_\gamma\hat{M}^{-1} \\ 0 & \hat{M}^{-1} \end{bmatrix} \hat{\Omega} \begin{bmatrix} \hat{G}_\theta^{-1} & -\hat{G}_\theta^{-1}\hat{G}_\gamma\hat{M}^{-1} \\ 0 & \hat{M}^{-1} \end{bmatrix}'.$$

An estimator of the asymptotic variance of the second step $\hat{\theta}$ can be extracted from the upper left block of this matrix. A convenient expression, corresponding to that in Theorem 6.1, can be obtained by letting $\hat{\psi}_i = -\hat{M}^{-1}\hat{m}_i$, so that the upper left block of $\hat{V}$ is

$$\hat{V}_\theta = \hat{G}_\theta^{-1}\left[n^{-1}\sum_{i=1}^{n}\{\hat{g}_i + \hat{G}_\gamma\hat{\psi}_i\}\{\hat{g}_i + \hat{G}_\gamma\hat{\psi}_i\}'\right]\hat{G}_\theta^{-1\prime}. \tag{6.11}$$

If the moment functions are uncorrelated as in eq. (6.9), so that the first-step estimation increases the second-step variance, then for $\hat{V}_\gamma = n^{-1}\sum_{i=1}^{n}\hat{\psi}_i\hat{\psi}_i'$, an asymptotic variance estimator for $\hat{\theta}$ is

$$\hat{V}_\theta = \hat{G}_\theta^{-1}\left(n^{-1}\sum_{i=1}^{n}\hat{g}_i\hat{g}_i'\right)\hat{G}_\theta^{-1\prime} + \hat{G}_\theta^{-1}\hat{G}_\gamma\hat{V}_\gamma\hat{G}_\gamma'\hat{G}_\theta^{-1\prime}. \tag{6.12}$$

This estimator is quite convenient, because most of its pieces can be recovered from standard output of computer programs. The first of the two terms being summed is a variance estimate that ignores the first step, as often provided by computer output (possibly in a different form than here). An estimated variance $\hat{V}_\gamma$ is also often provided by standard output from the first step. In many cases $\hat{G}_\theta^{-1}$ can also be recovered from the first step. Thus, often the only part of this variance estimator requiring application-specific calculation is $\hat{G}_\gamma$. This simplification is only possible under eq. (6.9). If the first- and second-step moment conditions are correlated then one will need the individual observations $\hat{\psi}_i$, in order to properly account for the covariance between the first- and second-step moments.

A consistency result for these asymptotic variance estimators can be obtained by applying the results of Section 4 to these joint moment conditions. It will suffice to assume that the joint moment vector $\tilde{g}(z, \theta, \gamma) = [m(z, \gamma)', g(z, \theta, \gamma)']'$ satisfies the conditions of Theorem 4.5. Because it is such a direct application of previous results a formal statement is not given here.

In some cases it may be possible to simplify $\hat{V}_\theta$ by using restrictions on the form of Jacobians and variance matrices that are implied by a model. The use of such restrictions in the general formula can be illustrated by deriving a consistent asymptotic variance estimator for the example.

Sample selection example continued: Let $W_i = d_i[w_i', \lambda(x_i'\gamma_0)]'$ and $\hat{W}_i = d_i[w_i', \lambda(x_i'\hat{\gamma})]'$. Note that by the residual having conditional mean zero given w, $d = 1$, and x, it is the case that $G_\theta = -E[d_i W_i W_i']$ and $G_\gamma = -\alpha_0 E[d_i \lambda_v(x_i'\gamma_0) W_i x_i']$, where terms involving second derivatives have dropped out by the residual having conditional mean zero. Estimates of these matrices are given by $\hat{G}_\theta = -\sum_{i=1}^n \hat{W}_i \hat{W}_i'/n$ and $\hat{G}_\gamma = -\hat{\alpha}\sum_{i=1}^n \lambda_v(x_i'\hat{\gamma})\hat{W}_i x_i'/n$. Applying eq. (6.12) to this case, for $\hat{\varepsilon}_i = y_i - \hat{W}_i'(\hat{\beta}', \hat{\alpha})'$, then gives

$$\hat{V}_\theta = \hat{G}_\theta^{-1}\left(n^{-1}\sum_{i=1}^n \hat{W}_i \hat{W}_i' \hat{\varepsilon}_i^2\right)\hat{G}_\theta^{-1\prime} + \hat{G}_\theta^{-1}\hat{G}_\gamma \hat{V}_\gamma \hat{G}_\gamma' \hat{G}_\theta^{-1\prime}$$

$$= n\left(\sum_{i=1}^n \hat{W}_i \hat{W}_i'\right)^{-1}\sum_{i=1}^n \hat{W}_i \hat{W}_i' \hat{\varepsilon}_i^2 \left(\sum_{i=1}^n \hat{W}_i \hat{W}_i'\right)^{-1} + \hat{\Pi}\hat{V}_\gamma \hat{\Pi}', \tag{6.13}$$

where $\hat{V}_\gamma$ is a probit estimator of the asymptotic variance of $\sqrt{n}(\hat{\gamma} - \gamma_0)$, e.g. as provided by a canned computer program, and $\hat{\Pi} = \hat{G}_\theta^{-1}\hat{G}_\gamma$ is the matrix of coefficients from a multivariate regression of $\hat{\alpha}\lambda_v(x_i'\hat{\gamma})x_i$ on $\hat{W}_i$. This estimator is the sum of the White (1980) variance matrix for least squares and a correction term for the first-stage estimation.[42] It will be a consistent estimator of the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$.[43]

7. Asymptotic normality with nonsmooth objective functions

The previous asymptotic normality results for MLE and GMM require that the log-likelihood be twice differentiable and that the moment functions be once differentiable. There are many examples of estimators where these functions are not that smooth. These include Koenker and Bassett (1978), Powell's (1984, 1986) censored least absolute deviations and symmetrically trimmed estimators, Newey and Powell's (1987) asymmetric least squares estimator, and the simulated moment estimators of Pakes (1986) and McFadden (1989). Therefore, it is important to have asymptotic normality results that allow for nonsmooth objective functions.

Asymptotic normality results for nonsmooth functions were developed by Daniels (1961), Huber (1967), Pollard (1985), and Pakes and Pollard (1989). The basic insight of these papers is that smoothness of the objective function can be replaced by smoothness of the limit if certain remainder terms are small. This insight is useful because the limiting objective functions are often expectations that are smoother than their sample counterparts.

[42] Contrary to a statement given in Amemiya (1985), the correction term is needed here.

[43] The normalization by the total sample size means that one can obtain asymptotic confidence intervals as described in Section 1, with the n given there equal to the total sample size. This procedure is equivalent to ignoring the n divisor in Section 1 and dropping the n from the probit asymptotic variance estimator (as is usually done in canned programs) and from the lead term in eq. (6.13).

To illustrate how this approach works it is useful to give a heuristic description. The basic idea is the approximation

$$\begin{aligned}\hat{Q}_n(\theta) - \hat{Q}_n(\theta_0) &\cong \hat{D}_n'(\theta - \theta_0) + Q_0(\theta) - Q_0(\theta_0) \\ &\cong \hat{D}_n'(\theta - \theta_0) + (\theta - \theta_0)H(\theta - \theta_0)/2,\end{aligned} \tag{7.1}$$

where $\hat{D}_n$ is a derivative, or approximate derivative, of $\hat{Q}_n(\theta)$ at θ_0, $H = \nabla_{\theta\theta} Q_0(\theta_0)$, and the second approximate equality uses the first-order condition $\nabla_\theta Q_0(\theta_0) = 0$ in a second-order expansion of $Q_0(\theta)$. This is an approximation of $\hat{Q}_n(\theta)$ by a quadratic function. Assuming that the approximation error is of the right order, the maximum of the approximation should be close to the true maximum, and the maximum of the approximation is $\tilde{\theta} = \theta_0 - H^{-1}\hat{D}_n$. This random variable will be asymptotically normal if $\hat{D}_n$ is, so that asymptotic normality of $\hat{\theta}$ will follow from asymptotic normality of its approximate value $\tilde{\theta}$.

7.1. *The basic results*

In order to make the previous argument precise the approximation error in eq. (7.1) has to be small enough. Indeed, the reason that eq. (7.1) is used, rather than some other expansion, is because it leads to approximation errors of just the right size. Suppose for discussion purposes that $\hat{D}_n = \nabla_\theta \hat{Q}_n(\theta_0)$, where the derivative exists with probability one. Then $\hat{Q}_n(\theta) - \hat{Q}_n(\theta_0) - \hat{D}_n'(\theta - \theta_0)$ goes to zero faster than $\| \theta - \theta_0 \|$ does, by the definition of a derivative. Similarly, $Q_0(\theta) - Q_0(\theta_0)$ goes to zero faster than $\| \theta - \theta_0 \|$ [since $\nabla_\theta Q_0(\theta_0) = 0$]. Also, assuming that $\sqrt{n}[\hat{Q}_n(\theta) - Q_0(\theta)]$ is bounded in probability for each θ, as would typically be the case when $\hat{Q}_n(\theta)$ is made up of sample averages, and noting that $\sqrt{n}\hat{D}_n$ bounded in probability follows by asymptotic normality, it follows that the remainder term,

$$\hat{R}_n(\theta) = \sqrt{n}[\hat{Q}_n(\theta) - \hat{Q}_n(\theta_0) - \hat{D}_n(\theta - \theta_0) - \{Q_0(\theta) - Q_0(\theta_0)\}]/\| \theta - \theta_0 \|, \tag{7.2}$$

is bounded in probability for each θ. Then, the combination of these two properties suggests that $\hat{R}_n(\theta)$ goes to zero as the sample size grows and θ goes to θ_0, a stochastic equicontinuity property. If so, then the remainder term in eq. (7.1) will be of order $o_p(\| \theta - \theta_0 \|/\sqrt{n} + \| \theta - \theta_0 \|^2)$. The next result shows that a slightly weaker condition is sufficient for the approximation in eq. (7.1) to lead to asymptotic normality of $\hat{\theta}$.

Theorem 7.1

Suppose that $\hat{Q}_n(\hat{\theta}) \geqslant \sup_{\theta \in \Theta} \hat{Q}_n(\theta) - o_p(n^{-1})$, $\hat{\theta} \xrightarrow{p} \theta_0$, and (i) $Q_0(\theta)$ is maximized on Θ at θ_0; (ii) θ_0 is an interior point of Θ; (iii) $Q_0(\theta)$ is twice differentiable at θ_0

with nonsingular second derivative H; (iv) $\sqrt{n}\hat{D} \xrightarrow{d} N(0, \Omega)$; (v) for any $\delta_n \longrightarrow 0$, $\sup_{\|\theta-\theta_n\| \leq \delta_n} |\hat{R}_n(\theta)/[1+\sqrt{n}\|\theta-\theta_0\|]| \xrightarrow{p} 0$. Then $\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0, H^{-1}\Omega H^{-1})$.

The proof of this result is given in Section 7.4. This result is essentially a version of Theorem 2 of Pollard (1985) that applies to any objective function rather than just a sample average, with an analogous method of proof. The key remainder condition is assumption (v), which is referred to by Pollard as *stochastic differentiability*. It is slightly weaker than $\hat{R}_n(\theta)$ converging to zero, because of the presence of the denominator term $(1+\sqrt{n}\|\theta-\theta_0\|)^{-1}$, which is similar to a term Huber (1967) used. In several cases the presence of this denominator term is quite useful, because it leads to a weaker condition on the remainder without affecting the conclusion. Although assumption (v) is quite complicated, primitive conditions for it are available, as further discussed below.

The other conditions are more straightforward. Consistency can be shown using Theorem 2.1, or the generalization that allows for $\hat{\theta}$ to be an approximate maximum, as suggested in the text following Theorem 2.1. Assumptions (ii) and (iii) are quite primitive, although verifying assumption (iii) may require substantial detailed work. Assumption (iv) will follow from a central limit theorem in the usual case where $\hat{D}_n$ is equal to a sample average.

There are several examples of GMM estimators in econometrics where the moments are not continuous in the parameters, including the simulated moment estimators of Pakes (1986) and McFadden (1989). For these estimators it is useful to have more specific conditions than those given in Theorem 7.1. One way such conditions can be formulated is in an asymptotic normality result for minimum distance estimators where $\hat{g}_n(\theta)$ is allowed to be discontinuous. The following is such a result.

Theorem 7.2

Suppose that $\hat{g}_n(\hat{\theta})'\hat{W}\hat{g}_n(\hat{\theta}) \leq \inf_{\theta\in\Theta}\hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta) + o_p(n^{-1})$, $\hat{\theta} \xrightarrow{p} \theta_0$, and $\hat{W} \xrightarrow{p} W$, W is positive semi-definite, where there is $g_0(\theta)$ such that (i) $g_0(\theta_0) = 0$; (ii) $g_0(\theta)$ is differentiable at θ_0 with derivative G such that $G'WG$ is nonsingular; (iii) θ_0 is an interior point of Θ; (iv) $\sqrt{n}\hat{g}_n(\theta_0) \xrightarrow{d} N(0, \Sigma)$; (v) for any $\delta_n \longrightarrow 0$, $\sup_{\|\theta-\theta_0\| \leq \delta_n}\sqrt{n}\,\|\hat{g}_n(\theta) - \hat{g}_n(\theta_0) - g_0(\theta)\|/[1+\sqrt{n}\|\theta-\theta_0\|] \xrightarrow{p} 0$. Then $\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N[0, (G'WG)^{-1}G' \times W\Sigma WG\,(G'WG)^{-1}]$.

The proof is given in Section 7.4. For the case where $\hat{g}_n(\theta)$ has the same number of elements as θ, this result is similar to Huber's (1967), and in the general case is like Pakes and Pollard's (1989), although the method of proof is different than either of these papers'. The conditions of this result are similar to those for Theorem 7.1. The function $g_0(\theta)$ should be thought of as the limit of $\hat{g}_n(\theta)$, as in Section 3. Most of the conditions are straightforward to interpret, except for assumption (v). This assumption is a "stochastic equicontinuity" assumption analogous to the condition (v) of Theorem 7.1. Stochastic equicontinuity is the appropriate term here because when $g_0(\theta)$ is the pointwise limit of $\hat{g}_n(\theta)$, i.e. $\hat{g}_n(\theta) \xrightarrow{p} g_0(\theta)$ for all θ, then for all

$\theta \neq \theta_0$, $\sqrt{n}\|\hat{g}_n(\theta) - \hat{g}_n(\theta_0) - g_0(\theta)\|/[1 + \sqrt{n}\|\theta - \theta_0\|] \xrightarrow{p} 0$. Thus, condition (v) can be thought of as an additional requirement that this convergence be uniform over any shrinking neighborhood of θ_0. As discussed in Section 2, stochastic equicontinuity is an essential condition for uniform convergence.

Theorem 7.2 is a special case of Theorem 7.1, in the sense that the proof proceeds by showing that the conditions of Theorem 7.1 are satisfied. Thus, in the nonsmooth case, asymptotic normality for minimum distance is a special case of asymptotic normality for an extremum estimator, in contrast to the results of Section 3. This relationship is the natural one when the conditions are sufficiently weak, because a minimum distance estimator is a special case of a general extremum estimator.

For some extremum estimators where $\nabla_\theta \hat{Q}_n(\theta)$ exists with probability one it is possible to use Theorem 7.2 to show asymptotic normality, by setting $\hat{g}_n(\theta)$ equal to $\nabla_\theta \hat{Q}_n(\theta)$. An example is censored least absolute deviations, where $\nabla_\theta \hat{Q}_n(\theta) = n^{-1}\sum_{i=1}^n x_i 1(x_i'\theta > 0)[1 - 2\cdot 1(y < x'\theta)]$. However, when this is done there is an additional condition that has to be checked, namely that $\|\nabla_\theta \hat{Q}_n(\hat{\theta})\|^2 \leqslant \inf_{\theta \in \Theta}\|\nabla_\theta \hat{Q}_n(\theta)\|^2 + o_p(n^{-1})$, for which it suffices to show that $\sqrt{n}\nabla_\theta \hat{Q}_n(\hat{\theta}) \xrightarrow{p} 0$. This is an "asymptotic first-order condition" for nonsmooth objective functions that generally has to be verified by direct calculations. Theorem 7.1 does not take this assumption to be one of its hypotheses, so that the task of checking the asymptotic first-order condition can be bypassed by working directly with the extremum estimator as in Theorem 7.1. In terms of the literature, this means that Huber's (1967) asymptotic first-order condition can be bypassed by working directly with the extremum formulation of the estimator, as in Pollard (1985). The cost of doing this is that the remainder in condition (v) of Theorem 7.1 tends to be more complicated than the remainder in condition (v) of Theorem 7.2, making that regularity condition more difficult to check.

The most complicated regularity condition in Theorems 7.1 and 7.2 is assumption (v). This condition is difficult to check in the form given, but there are more primitive conditions available. In particular, for $\hat{Q}_n(\theta) = n^{-1}\sum_{i=1}^n q(z_i, \theta)$, where the objective function is a sample average, Pollard (1985) has given primitive conditions for stochastic differentiability. Also, for GMM where $\hat{g}_n(\theta) = \sum_{i=1}^n g(z_i, \theta)/n$ and $g_0(\theta) = E[g(z, \theta)]$, primitive conditions for stochastic equicontinuity are given in Andrews' (1994) chapter of this handbook. Andrews (1994) actually gives conditions for a stronger result, that $\sup_{\|\theta - \theta_0\| \leqslant \delta_n}\sqrt{n}\|\hat{g}_n(\theta) - \hat{g}_n(\theta_0) - g_0(\theta)\| \xrightarrow{p} 0$, i.e. for (v) of Theorem 7.2 without the denominator term. The conditions described in Pollard (1985) and Andrews (1994) allow for very weak conditions on $g(z, \theta)$, e.g. it can even be discontinuous in θ. Because there is a wide variety of such conditions, we do not attempt to describe them here, but instead refer the reader to Pollard (1985) and Andrews (1994).

There is a primitive condition for stochastic equicontinuity that is not covered in these other papers, that allows for $g(z, \theta)$ to be Lipschitz at θ_0 and differentiable with probability one, rather than continuously differentiable. This condition is simple but has a number of applications, as we discuss next.

7.2. *Stochastic equicontinuity for Lipschitz moment functions*

The following result gives a primitive condition for the stochastic equicontinuity hypothesis of Theorem 7.2 for GMM, where $\hat{g}_n(\theta) = n^{-1}\sum_{i=1}^n g(z_i, \theta)$ and $g_0(\theta) = E[g(z, \theta)]$.

Theorem 7.3

Suppose that $E[g(z, \theta_0)] = 0$ and there are $\Delta(z)$ and $\varepsilon > 0$ such that with probability one, $r(z, \theta) = \| g(z, \theta) - g(z, \theta_0) - \Delta(z)(\theta - \theta_0) \| / \| \theta - \theta_0 \| \to 0$ as $\theta \to \theta_0$, $E[\sup_{\| \theta - \theta_0 \| < \varepsilon} \times r(z, \theta)] < \infty$, and $n^{-1}\sum_{i=1}^n \Delta(z_i) \xrightarrow{p} E[\Delta(z)]$. Then assumptions (ii) and (v) of Theorem 7.2 are satisfied for $G = E[\Delta(z)]$.

Proof

For any $\varepsilon > 0$, let $r(z, \varepsilon) = \sup_{\| \theta - \theta_0 \| \leq \varepsilon} \| r(z, \theta) \|$. With probability one $r(z, \varepsilon) \to 0$ as $\varepsilon \to 0$, so by the dominated convergence theorem, $E[r(z, \varepsilon)] \to 0$ as $\varepsilon \to 0$. Then for $\theta \to \theta_0$ and $\varepsilon = \| \theta - \theta_0 \|$, $\| g_0(\theta) - g_0(\theta_0) - G(\theta - \theta_0) \| = \| E[g(z, \theta) - g(z, \theta_0) - \Delta(z) \times (\theta - \theta_0)] \| \leq E[r(z, \varepsilon)] \| \theta - \theta_0 \| \to 0$, giving assumption (iii). For assumption (v), note that for all θ with $\| \theta - \theta_0 \| \leq \delta_n$, by the definition of $r(z, \varepsilon)$ and the Markov inequality, $\sqrt{n} \| \hat{g}_n(\theta) - \hat{g}_n(\theta_0) - g_0(\theta) \| / [1 + \sqrt{n} \| \theta - \theta_0 \|] \leq \sqrt{n} [\| \sum_{i=1}^n \{\Delta(z_i) - E[\Delta(z)]\} \times (\theta - \theta_0)/n \| + \{\sum_{i=1}^n r(z_i, \delta_n)/n + E[r(z, \delta_n)]\} \| \theta - \theta_0 \|] / (1 + \sqrt{n} \| \theta - \theta_0 \|) \leq \| \sum_{i=1}^n \{\Delta(z_i) - E[\Delta(z)]\}/n \| + O_p(E[r(z, \delta_n)]) \xrightarrow{p} 0$. Q.E.D.

The condition on $r(z, \theta)$ in this result was formulated by Hansen et al. (1992). The requirement that $r(z, \theta) \to 0$ as $\theta \to \theta_0$ means that, with probability one, $g(z, \theta)$ is differentiable with derivative $\Delta(z)$ at θ_0. The dominance condition further restricts this remainder to be well behaved uniformly near the true parameter. This uniformity property requires that $g(z, \theta)$ be Lipschitz at θ_0 with an integrable Lipschitz constant.[44]

A useful aspect of this result is that the hypotheses only require that $\sum_{i=1}^n \Delta(z_i) \xrightarrow{p} E[\Delta(z)]$, and place no other restriction on the dependence of the observations. This result will be quite useful in the time series context, as it is used in Hansen et al. (1992). Another useful feature is that the conclusion includes differentiability of $g_0(\theta)$ at θ_0, a "bonus" resulting from the dominance condition on the remainder.

The conditions of Theorem 7.3 are strictly weaker than the requirement of Section 3 that $g(z, \theta)$ be continuously differentiable in a neighborhood of θ_0 with derivative that is dominated by an integrable function, as can be shown in a straightforward way. An example of a function that satisfies Theorem 7.3, but not the stronger continuous differentiability condition, is the moment conditions corresponding to Huber's (1964) robust location estimator.

[44] For $d(z) = \sup_{\| \theta - \theta_0 \| < \varepsilon} r(z, \theta)$, the triangle and Cauchy–Schwarz inequalities imply $\| g(z, \theta) - g(z, \theta_0) \| \leq [\| \Delta(z) \| + d(z)] \| \theta - \theta_0 \|$.

Huber's robust location estimator: The first-order conditions for this estimator are $n^{-1}\sum_{i=1}^{n}\rho(y_i-\hat{\theta})=0$ for $\rho(\varepsilon)=-1(\varepsilon\leqslant -1)+1(-1<\varepsilon<1)\varepsilon+1(\varepsilon\geqslant 1)$. This estimator will be consistent for θ_0 where y is symmetrically distributed around θ_0. The motivation for this estimator is that its first-order condition is a bounded, continuous function of the data, giving it a certain robustness property; see Huber (1964). This estimator is a GMM estimator with $g(z,\theta)=\rho(y-\theta)$. The function $\rho(\varepsilon)$ is differentiable everywhere except at -1 or 1, with derivative $\rho_\varepsilon(\varepsilon)=1(-1<\varepsilon<1)$. Let $\Delta(z)=-\rho_\varepsilon(y-\theta_0)$. Then for $\varepsilon=y-\theta_0$ and $\delta=\theta_0-\theta$,

$$\begin{aligned}
r(z,\theta) &= |g(z,\theta)-g(z,\theta_0)-\Delta(z)(\theta-\theta_0)|/|\theta-\theta_0| \\
&= |\rho(\varepsilon+\delta)-\rho(\varepsilon)-\rho_\varepsilon(\varepsilon)\delta|/|\delta| \\
&= |[-1(\varepsilon+\delta\leqslant -1)+1(\varepsilon\leqslant -1)]+[1(\varepsilon+\delta\geqslant 1)-1(\varepsilon\geqslant 1)] \\
&\quad +[1(-1<\varepsilon+\delta<1)-1(-1<\varepsilon<1)](\varepsilon+\delta)|/|\delta|.
\end{aligned}$$

For $0<\delta\leqslant 1$,

$$\begin{aligned}
r(z,\theta) &= |1(-1-\delta<\varepsilon\leqslant -1)+1(1-\delta\leqslant\varepsilon<1)+[1(-1-\delta<\varepsilon\leqslant -1) \\
&\quad -1(1-\delta\leqslant\varepsilon<1)](\varepsilon+\delta)|/|\delta| \\
&\leqslant 1(-\delta<\varepsilon+1\leqslant 0)(\delta+|\varepsilon+1|)/|\delta|+1(-\delta\leqslant\varepsilon-1<0)(|\varepsilon-1|+\delta)/|\delta| \\
&\leqslant 2[1(-\delta<\varepsilon+1\leqslant 0)+1(-\delta\leqslant\varepsilon-1<0)]\leqslant 2.
\end{aligned}$$

Applying an analogous argument for negative $-1\leqslant\delta<0$ gives $r(z,\theta)\leqslant 2[1(|\varepsilon-1|\leqslant|\delta|)+1(|\varepsilon+1|\leqslant|\delta|)]\leqslant 4$. Therefore, if $\text{Prob}(\varepsilon=1)=0$ and $\text{Prob}(\varepsilon=-1)=0$ then $r(z,\theta)\to 0$ with probability one as $\theta\to\theta_0$ (i.e. as $\delta\to 0$). Also, $r(z,\theta)\leqslant 4$. Thus, the conditions of Theorem 7.3 are satisfied.

Other examples of estimators that satisfy these conditions are the asymmetric least squares estimator of Newey and Powell (1987) and the symmetrically trimmed estimators for censored Tobit models of Powell (1986) and Honoré (1992). All of these examples are interesting, and illustrate the usefulness of Theorem 7.3.

7.3. *Asymptotic variance estimation*

Just as in the smooth case the asymptotic variance of extremum and minimum distance estimators contain derivative and variance terms. In the smooth case the derivative terms were easy to estimate, using derivatives of the objective functions. In the nonsmooth case these estimates are no longer available, so alternatives must be found. One alternative is numerical derivatives.

For the general extremum estimator of Theorem 7.1, the matrix H can be

estimated by a second-order numerical derivative of the objective function. Let e_i denote the ith unit vector, ε_n a small positive constant that depends on the sample size, and $\hat{H}$ the matrix with i, jth element

$$\hat{H}_{ij} = [\hat{Q}(\hat{\theta} + e_i\varepsilon_n + e_j\varepsilon_n) - \hat{Q}(\hat{\theta} - e_i\varepsilon_n + \dot{e}_j\varepsilon_n) - \hat{Q}(\hat{\theta} + e_i\varepsilon_n - e_j\varepsilon_n) + \hat{Q}(\hat{\theta} - e_i\varepsilon_n - e_j\varepsilon_n)]/4\varepsilon_n^2.$$

Under certain conditions on ε_n, the hypotheses of Theorem 7.1 will suffice for consistency of $\hat{H}$ for the H in the asymptotic variance of Theorem 7.1. For a minimum distance estimator a numerical derivative estimator $\hat{G}$ of G has jth column

$$\hat{G}_j = [\hat{g}(\hat{\theta} + e_j\varepsilon_n) - \hat{g}(\hat{\theta} - e_j\varepsilon_n)]/2\varepsilon_n.$$

This estimator will be consistent under the conditions of Theorem 7.2. The following result shows consistency:

Theorem 7.4

Suppose that $\varepsilon_n \to 0$ and $\varepsilon_n\sqrt{n} \to \infty$. If the conditions of Theorem 7.1 are satisfied then $\hat{H} \xrightarrow{p} H$. Also, if the conditions of Theorem 7.2 are satisfied then $\hat{G} \xrightarrow{p} G$.

This result is proved in Section 7.4. Similar results have been given by McFadden (1989), Newey (1990), and Pakes and Pollard (1989).

A practical problem for both of these estimators is the degree of difference (i.e. the magnitude of ε_n) used to form the numerical derivatives. Our specification of the same ε_n for each component is only good if $\hat{\theta}$ has been scaled so that its components have similar magnitude. Alternatively, different ε_n could be used for different components, according to their scale. Choosing the size of ε_n is a difficult problem, although analogies with the choice of bandwidth for nonparametric regression, as discussed in the chapter by Härdle and Linton (1994), might be useful. One possibility is to graph some component as a function of ε_n and then choose ε_n small, but not in a region where the function is very choppy. Also, it might be possible to estimate variance and bias terms, and choose ε_n to balance them, although this is beyond the scope of this chapter.

In specific cases it may be possible to construct estimators that do not involve numerical differentiation. For example, in the smooth case we know that a numerical derivative can be replaced by analytical derivatives. A similar replacement is often possible under the conditions of Theorem 7.3. In many cases where Theorem 7.3 applies, $g(z, \theta)$ will often be differentiable with probability one with a derivative $\nabla_\theta g(z, \theta)$ that is continuous in θ with probability one and dominated by an integrable function. Consistency of $\hat{G} = n^{-1}\sum_{i=1}^n \nabla_\theta g(z_i, \hat{\theta})$ will then follow from Lemma 4.3. For example, it is straightforward to show that this reasoning applies to the Huber location estimator, with $\nabla_\theta g(z, \theta) = -1(-1 < y - \theta < 1)$ and $\hat{G} = \sum_{i=1}^n 1(-1 < y_i - \hat{\theta} < 1)/n$.

Estimation of the other terms in the asymptotic variance of $\hat{\theta}$ can usually be carried out in the way described in Section 4. For example, for GMM the moment function $g(z,\theta)$ will typically be continuous in θ with probability one and be dominated by a square integrable function, so that Lemma 4.3 will imply the consistency of $\hat{\Omega} = \sum_{i=1}^{n} g(z_i,\hat{\theta})g(z_i,\hat{\theta})'/n$. Also, extremum estimators where $\hat{Q}_n(\theta) = n^{-1}\sum_{i=1}^{n} q(z,\theta)$, $q(z,\theta)$ will usually be differentiable almost everywhere, and Lemma 4.3 will yield consistency of the variance estimator given in eq. (4.1).

7.4. *Technicalities*

Because they are long and somewhat complicated, the proofs of Theorems 7.1, 7.2, and 7.4 are given here rather than previously.

Proof of Theorem 7.1

Let $\hat{Q}(\theta) = \hat{Q}_n(\theta)$ and $Q(\theta) = Q_0(\theta)$. First it will be proven that $\sqrt{n}\,\|\hat{\theta} - \theta_0\| = O_p(1)$, i.e. that $\hat{\theta}$ is "$\sqrt{n}$-consistent". By $Q(\theta)$ having a local maximum at θ_0, its first derivative is zero at θ_0, and hence $Q(\theta) = Q(\theta_0) + (\theta - \theta_0)'H(\theta - \theta_0)/2 + o(\|\theta - \theta_0\|^2)$. Also, H is negative definite by θ_0 a maximum and nonsingularity of H, so that there is $C > 0$ and a small enough neighborhood of θ_0 with $(\theta - \theta_0)'H(\theta - \theta_0)/2 + o(\|\theta - \theta_0\|^2) \leq -C\|\theta - \theta_0\|^2$. Therefore, by $\hat{\theta} \xrightarrow{p} \theta_0$, with probability approaching one (w.p.a.1), $Q(\hat{\theta}) \leq Q(\theta_0) - C\|\hat{\theta} - \theta_0\|^2$. Choose U_n so that $\hat{\theta} \in U_n$ w.p.a.1, so that by (v) $\sqrt{n}|\hat{R}(\hat{\theta})| \leq (1 + \sqrt{n}\|\hat{\theta} - \theta_0\|)o_p(1)$.

$$
\begin{aligned}
0 &\leq \hat{Q}(\hat{\theta}) - \hat{Q}(\theta_0) + o_p(n^{-1}) = Q(\hat{\theta}) - Q(\theta_0) + \hat{D}'(\hat{\theta} - \theta_0) + \|\hat{\theta} - \theta_0\|\hat{R}(\hat{\theta}) + o_p(n^{-1}) \\
&\leq -C\|\hat{\theta} - \theta_0\|^2 + \|\hat{D}\|\,\|\hat{\theta} - \theta_0\| + \|\hat{\theta} - \theta_0\|(1 + \sqrt{n}\|\hat{\theta} - \theta_0\|)o_p(n^{-1/2}) + o_p(n^{-1}) \\
&\leq -[C + o_p(1)]\|\hat{\theta} - \theta_0\|^2 + O_p(n^{-1/2})\|\hat{\theta} - \theta_0\| + o_p(n^{-1}).
\end{aligned}
$$

Since $C + o_p(1)$ is bounded away from zero w.p.a.1, it follows that $\|\hat{\theta} - \theta_0\|^2 \leq O_p(n^{-1/2})\|\hat{\theta} - \theta_0\| + o_p(n^{-1})$, and hence, completing the square, that $[\|\hat{\theta} - \theta_0\| + O_p(n^{-1/2})]^2 \leq O_p(n^{-1})$. Taking the square root of both sides, it follows that $|\|\hat{\theta} - \theta_0\| + O_p(n^{-1/2})| \leq O_p(n^{-1/2})$, so by the triangle inequality, $\|\hat{\theta} - \theta_0\| \leq |\|\hat{\theta} - \theta_0\| + O_p(n^{-1/2})| + |-O_p(n^{-1/2})| \leq O_p(n^{-1/2})$.

Next, let $\tilde{\theta} = \theta_0 - H^{-1}\hat{D}$, and note that by construction it is $\sqrt{n}$-consistent. Then by $\sqrt{n}$-consistency of $\hat{\theta}$, twice differentiability of $Q(\theta)$, and (v) it follows that

$$
\begin{aligned}
2[\hat{Q}(\hat{\theta}) - \hat{Q}(\theta_0) + Q(\theta_0)] &= (\hat{\theta} - \theta_0)'H(\hat{\theta} - \theta_0) + 2\hat{D}'(\hat{\theta} - \theta_0) + o_p(n^{-1}) \\
&= (\hat{\theta} - \theta_0)'H(\hat{\theta} - \theta_0) - 2(\tilde{\theta} - \theta_0)'H(\hat{\theta} - \theta_0) + o_p(n^{-1}).
\end{aligned}
$$

Similarly, $2[\hat{Q}(\tilde{\theta}) - \hat{Q}(\theta_0) + Q(\theta_0)] = (\tilde{\theta} - \theta_0)'H(\tilde{\theta} - \theta_0) + 2\hat{D}'(\tilde{\theta} - \theta_0) + o_p(n^{-1}) =$

$-(\tilde{\theta}-\theta_0)'H(\tilde{\theta}-\theta_0)+o_p(n^{-1})$. Then since $\tilde{\theta}$ is contained within Θ w.p.a.1, $2[\hat{Q}(\hat{\theta})-\hat{Q}(\theta_0)+Q(\theta_0)]-2[\hat{Q}(\tilde{\theta})-\hat{Q}(\theta_0)+Q(\theta_0)]\geqslant o_p(n^{-1})$, so by the last equation and the corresponding equation for $\tilde{\theta}$,

$$o_p(n^{-1})\leqslant(\hat{\theta}-\theta_0)'H(\hat{\theta}-\theta_0)-2(\tilde{\theta}-\theta_0)'H(\hat{\theta}-\theta_0)+(\tilde{\theta}-\theta_0)'H(\tilde{\theta}-\theta_0)$$
$$=(\hat{\theta}-\tilde{\theta})'H(\hat{\theta}-\tilde{\theta})\leqslant -C\|\hat{\theta}-\tilde{\theta}\|^2.$$

Therefore, $\|\sqrt{n}(\hat{\theta}-\theta_0)-(-H^{-1}\sqrt{n}\hat{D})\|=\sqrt{n}\|\hat{\theta}-\tilde{\theta}\|\xrightarrow{p}0$, so the conclusion follows by $-H^{-1}\sqrt{n}\hat{D}\xrightarrow{d}N(0,H^{-1}\Omega H^{-1})$ and the Slutzky theorem. Q.E.D.

Proof of Theorem 7.2

Let $\hat{g}(\theta)=\hat{g}_n(\theta)$ and $g(\theta)=g_0(\theta)$. The proof proceeds by verifying the hypotheses of Theorem 7.1, for $Q(\theta)=-g(\theta)'Wg(\theta)/2$, $\hat{Q}(\theta)=-\hat{g}(\theta)'\hat{W}\hat{g}(\theta)/2+\hat{\Delta}(\theta)$, and $\hat{\Delta}(\theta)$ equal to a certain function specified below. By (i) and (ii), $Q(\theta)=-[G(\theta-\theta_0)+o(\|\theta-\theta_0\|)]'W[G(\theta-\theta_0)+o(\|\theta-\theta_0\|)]/2=Q(\theta_0)+(\theta-\theta_0)'H(\theta-\theta_0)/2+o(\|\theta-\theta_0\|^2)$, for $H=-G'WG$ and $Q(\theta_0)=0$, so that $Q(\theta)$ is twice differentiable at θ_0. Also, by W positive semi-definite and $G'WG$ nonsingular, H is negative definite, implying that there is a neighborhood of θ_0 on which $Q(\theta)$ has a unique maximum (of zero) at $\theta=\theta_0$. Thus, hypotheses (i)–(ii) of Theorem 7.1 are satisfied. By the Slutzky theorem, $\hat{D}=-G'\hat{W}\sqrt{n}\hat{g}(\theta_0)\xrightarrow{d}N(0,\Omega)$ for $\Omega=G'W\Sigma WG$, so that hypothesis (v) of Theorem 7.1 is satisfied. It therefore remains to check the initial supposition and hypothesis (v) of Theorem 7.1.

Let $\hat{\varepsilon}(\theta)=[\hat{g}(\theta)-\hat{g}(\theta_0)-g(\theta)]/[1+\sqrt{n}\|\theta-\theta_0\|]$. Then

$$\hat{g}(\theta)'\hat{W}\hat{g}(\theta)=(1+2\sqrt{n}\|\theta-\theta_0\|+\|\theta-\theta_0\|^2)\hat{\varepsilon}(\theta)'\hat{W}\hat{\varepsilon}(\theta)+g(\theta)'\hat{W}g(\theta)$$
$$+\hat{g}(\theta_0)'\hat{W}\hat{g}(\theta_0)+2g(\theta)'\hat{W}\hat{g}(\theta_0)$$
$$+2[g(\theta)+\hat{g}(\theta_0)]'\hat{W}\hat{\varepsilon}(\theta)[1+\sqrt{n}\|\theta-\theta_0\|].$$

Let $\hat{Q}(\theta)=-\hat{g}(\theta)'\hat{W}\hat{g}(\theta)/2+\hat{\varepsilon}(\theta)'\hat{W}\hat{\varepsilon}(\theta)/2+\hat{g}(\theta_0)'\hat{W}\hat{\varepsilon}(\theta)$. For any $\delta_n\to 0$, by (v), $\sup_{\|\theta-\theta_0\|\leqslant\delta_n}|\hat{Q}(\theta)-\{-\hat{g}(\theta)'\hat{W}\hat{g}(\theta)/2\}|\leqslant O_p(1)\sup_{\|\theta-\theta_0\|\leqslant\delta_n}\{\|\hat{\varepsilon}(\theta)\|\,\|\hat{\varepsilon}(\theta)\|+O_p(n^{-1/2})\}=o_p(n^{-1})$, so that by (i), $\hat{Q}(\hat{\theta})\geqslant\sup_{\|\theta-\theta_0\|\leqslant\delta_n}\hat{Q}(\theta)-o_p(n^{-1})$. Thus, the initial supposition of Theorem 7.1 is satisfied. To check hypothesis (v), note that by $\hat{\varepsilon}(\theta_0)=0$, for $\hat{R}(\theta)$ as defined above,

$$\sqrt{n}|\hat{R}(\theta)/[1+\sqrt{n}\|\theta-\theta_0\|]|\leqslant\sum_{j=1}^{5}\hat{r}_j(\theta),$$

$$\hat{r}_1(\theta)=\sqrt{n}(2\sqrt{n}\|\theta-\theta_0\|+\|\theta-\theta_0\|^2)|\hat{\varepsilon}(\theta)'\hat{W}\hat{\varepsilon}(\theta)|/[\|\theta-\theta_0\|(1+\sqrt{n}\|\theta-\theta_0\|)],$$

$$\hat{r}_2(\theta)=\sqrt{n}|[g(\theta)+G(\theta-\theta_0)]'\hat{W}\hat{g}(\theta_0)|/[\|\theta-\theta_0\|(1+\sqrt{n}\|\theta-\theta_0\|)],$$

$$\hat{r}_3(\theta) = n|[g(\theta) + \hat{g}(\theta_0)]'\hat{W}\hat{\varepsilon}(\theta)|/(1 + \sqrt{n}\|\theta - \theta_0\|),$$

$$\hat{r}_4(\theta) = \sqrt{n}|g(\theta)'\hat{W}\hat{\varepsilon}(\theta)|/\|\theta - \theta_0\|,$$

$$\hat{r}_5(\theta) = \sqrt{n}|g(\theta)'[\hat{W} - W]g(\theta)|/[\|\theta - \theta_0\|(1 + \sqrt{n}\|\theta - \theta_0\|)].$$

Then for $\delta_n \to 0$ and $U = \{\theta: \|\theta - \theta_n\| \leq \delta_n\}$, $\sup_U \hat{r}_1(\theta) \leq Cn \cdot \sup_U \|\hat{\varepsilon}(\theta)\|^2 \|\hat{W}\| = o_p(1)$, $\sup_U \hat{r}_2(\theta) \leq \sqrt{n} \sup_U \{o(\|\theta - \theta_0\|)\|\hat{W}\| \, \|\hat{g}(\theta_0)\| = \sup_U o(\|\theta - \theta_0\|)O_p(1) = o_p(1)$, $\sup_U \hat{r}_3(\theta) \leq \{\sup_U \sqrt{n}\|g(\theta)\|/\sqrt{n}\|\theta - \theta_0\|) + \sqrt{n}\|\hat{g}(\theta_0)\|\}\|\hat{W}\| \sup_U \sqrt{n}\|\hat{\varepsilon}(\theta)\| = \{\sup_U O(\|\theta - \theta_0\|) + O_p(1)\}o_p(1) = o_p(1)$, $\sup_U \hat{r}_4(\theta) \leq \sup_U (\|g(\theta)\|/\|\theta - \theta_0\|)\|\hat{W}\| \sup_U \sqrt{n}\|\hat{\varepsilon}(\theta)\| = o_p(1)$, and $\sup_U \hat{r}_5(\theta) \leq \sup_U (\|g(\theta)\|^2/\|\theta - \theta_0\|^2)\|\hat{W} - W\| = o_p(1)$. Q.E.D.

Proof of Theorem 7.4

Let a be a constant vector. By the conclusion of Theorem 7.1, $\|\hat{\theta} + a\varepsilon_n - \theta_0\| = O_p(\varepsilon_n)$. Then by hypothesis (v),

$$\begin{aligned}
&|\hat{Q}(\hat{\theta} + \varepsilon_n a) - \hat{Q}(\theta_0) - Q(\hat{\theta} + \varepsilon_n a) + Q(\theta_0)| \\
&\quad \leq \|\hat{\theta} + a\varepsilon_n - \theta_0\|[|\hat{R}(\hat{\theta} + \varepsilon_n a)| + \|\hat{D}\| \, \|\hat{\theta} + a\varepsilon_n - \theta_0\|] \\
&\quad \leq O_p(\varepsilon_n)\{(1 + \sqrt{n}\|\hat{\theta} + \varepsilon_n a - \theta_0\|)o_p(1/\sqrt{n}) + O_p(\varepsilon_n/\sqrt{n})\} = o_p(\varepsilon_n^2).
\end{aligned}$$

Also, by twice differentiability of $Q(\theta)$ at θ_0,

$$\begin{aligned}
&|\varepsilon_n^{-2}[Q(\hat{\theta} + \varepsilon_n a) - Q(\theta_0)] - a'Ha/2| \\
&\quad = |\varepsilon_n^{-2}[(\hat{\theta} + \varepsilon_n a - \theta_0)'H(\hat{\theta} + \varepsilon_n a - \theta_0)/2 + o(\|\hat{\theta} + \varepsilon_n a - \theta_0\|^2)] - a'Ha/2| \\
&\quad \leq |\varepsilon_n^{-1}(\hat{\theta} - \theta_0)'Ha| + |\varepsilon_n^{-2}(\hat{\theta} - \theta_0)'H(\hat{\theta} - \theta_0)| + o_p(1) = o_p(1).
\end{aligned}$$

It then follows by the triangle inequality that

$$\begin{aligned}
\hat{H}_{ij} &\xrightarrow{p} [2(e_i + e_j)'H(e_i + e_j) - (e_i - e_j)'H(e_i - e_j) - (e_j - e_i)'H(e_j - e_i)]/8 \\
&= 2[e_i'He_i + e_j'He_j - e_i'He_i - e_j'He_j]/8 + e_i'He_j \\
&= e_i'He_j = H_{ij},
\end{aligned}$$

giving the first conclusion. For the second conclusion, it follows from hypothesis (v) of Theorem 7.2, similarly to the proof for $\hat{H}$, that $\|\hat{g}(\hat{\theta} + \varepsilon_n a) - \hat{g}(\theta_0) - g(\hat{\theta} + \varepsilon_n a)\| \leq (1 + \sqrt{n}\|\hat{\theta} + \varepsilon_n a - \theta_0\|)o_p(n^{-1/2}) = o_p(\varepsilon_n^{-1})$, and by differentiability of $g(\theta)$ at θ_0 that $\|g(\hat{\theta} + \varepsilon_n a)/\varepsilon_n - Ga\| \leq \|G(\hat{\theta} - \theta_0)/\varepsilon_n\| + o(\varepsilon_n^{-1}\|\hat{\theta} + \varepsilon_n a - \theta_0\|) = o_p(1)$. The second conclusion then follows by the triangle inequality. Q.E.D.

8. Semiparametric two-step estimators

Two-step estimators where the first step is a function rather than a finite-dimensional parameter, referred to here as semiparametric two-step estimators, are of interest in a number of econometric applications.[45] As noted in Section 5, they are useful for constructing feasible efficient estimators when there is a nuisance function present. Also, they provide estimators for certain econometric parameters of interest without restricting functional form, such as consumer surplus in an example discussed below. An interesting property of these estimators is that they can be $\sqrt{n}$-consistent, even though the convergence rate for the first-step functions is slower than $\sqrt{n}$. This section discusses how and when this property holds, and gives regularity conditions for asymptotic normality of the second-step estimator. The regularity conditions here are somewhat more technical than those of previous sections, as required by the infinite-dimensional first step.

The type of estimator to be considered here will be one that solves

$$n^{-1}\sum_{i=1}^{n} g(z_i, \theta, \hat{\gamma}) = 0, \tag{8.1}$$

where $\hat{\gamma}$ can include infinite-dimensional functions and $g(z, \theta, \gamma)$ is some function of a data observation z, the parameters of interest θ, and a function γ. This estimator is exactly like that considered in Section 6, except for the conceptual difference that γ is allowed to denote a function rather than a finite-dimensional vector. Here, $g(z, \theta, \gamma)$ is a vector valued function of a function. Such things are usually referred to as functionals.

Examples are useful for illustrating how semiparametric two-step estimators can be fit into this framework.

V-estimators: Consider a simultaneous equations model where the residual $\rho(z, \theta)$ is independent of the instrumental variables x. Let $a(x, \rho)$ be a vector of functions of the instrumental variables and the residual ρ. Independence implies that $E[a\{x, \rho(z, \theta_0)\}] = E[\int a\{x, \rho(\tilde{z}, \theta_0)\} dF_0(\tilde{z})]$ where $F_0(z)$ is the distribution of a single observation. For example, if $a(x, \rho)$ is multiplicatively separable, then this restriction is that the expectation of the product is the product of the expectations. This restriction can be exploited by replacing expectations with sample averages and $dF(\tilde{z})$ with an estimator, and then solving the corresponding equation, as in

$$\sum_{i=1}^{n}\sum_{j=1}^{n} m(z_i, z_j, \theta)/n^2 = 0, \tag{8.2}$$

where $m(z_1, z_2, \theta) = a[x_1, \rho(z_1, \theta)] - a[x_1, \rho(z_2, \theta)]$. This estimator has the form given

[45] This terminology may not be completely consistent with Powell's chapter of this handbook.

in eq. (8.1), where γ is the CDF of a single observation, $g(z,\theta,\gamma)=\int m(z,\tilde{z},\theta)\mathrm{d}\gamma(\tilde{z})$, and $\hat{\gamma}$ is the empirical distribution with $\hat{\gamma}(\hat{z})=\sum_{i=1}^{n}1(z_i\leqslant\tilde{z})/n$. It is referred to as a V-estimator because double averages like that in eq. (8.2) are often referred to as V-statistics [Serfling (1980)]. V-statistics are related to U-statistics, which have been considered in recent econometric literature [e.g. Powell et al. (1989) and Robinson (1988b)] and are further discussed below.

The general class of V-estimators were considered in Newey (1989). If $a(x,\rho)$ is multiplicatively separable in x and ρ then these estimators just set a vector of sample covariances equal to zero. It turns out though, that the optimal $a(x,\rho)$ may not be multiplicatively separable, e.g. it can include Jacobian terms, making the generalization in eq. (8.2) of some interest. Also, Honoré and Powell (1992) have recently suggested estimators that are similar to those in equation (8.2), and given conditions that allow for lack of smoothness of $m(z_1,z_2,\theta)$ in θ.

Nonparametric approximate consumer surplus estimation: Suppose that the demand function as a function of price is given by $h_0(x)=E[q|x]$, where q is quantity demanded and x is price. The approximate consumer surplus for a price change from a to b is $\int_a^b h_0(x)\mathrm{d}x$. A nonparametric estimator can be constructed by replacing the true conditional expectation by a nonparametric estimator. One such is a kernel estimator of the form $\hat{h}(x)=\sum_{i=1}^{n}q_iK_\sigma(x-x_i)/\sum_{i=1}^{n}K_\sigma(x-x_i)$, where $K_\sigma(v)=\sigma^{-r}K(v/\sigma)$, r is the dimension of x, $K(u)$ is a function such that $\int K(u)\mathrm{d}u=1$, and σ is a bandwidth term that is chosen by the econometrician. This estimator is a weighted average of q_i, with the weight for the ith observation given by $K_\sigma(x-x_i)/\sum_{j=1}^{n}K_\sigma(x-x_j)$. The bandwidth σ controls the amount of local weighting and hence the variance and bias of this estimator. As σ goes down, more weight will tend to be given to observations with x_i close to x, lowering bias, but raising variance by giving more weight to fewer observations. Alternatively, $\hat{h}(x)$ can be interpreted as a ratio estimator, with a denominator $\hat{f}(x)=n^{-1}\sum_{i=1}^{n}K_\sigma(x-x_i)$ that is an estimator of the density of x. These kernel estimators are further discussed in Härdle and Linton (1994).

A kernel estimator of $h_0(x)$ can be used to construct a consumer surplus estimator of the form $\hat{\theta}=\int_a^b\hat{h}(x)\mathrm{d}x$. This estimator takes the form given in eq. (8.1), for $\gamma=(\gamma_1,\gamma_2)$ where $\gamma_1(x)$ is a density for x, $\gamma_2(x)$ is the product of a density for x and a conditional expectation of y given x, $g(z,\theta,\gamma)=\int_a^b[\gamma_2(x)/\gamma_1(x)]\mathrm{d}x-\theta$, $\hat{\gamma}_1(x)=n^{-1}\sum_{i=1}^{n}K_\sigma(x-x_i)$ and $\hat{\gamma}_2(x)=n^{-1}\sum_{i=1}^{n}q_iK_\sigma(x-x_i)$. This particular specification, where γ consists separately of the numerator and denominator of $\hat{h}(x)$, is convenient in the analysis to follow.

In both of these examples there is some flexibility in the formulation of the estimator as a solution to eq. (8.1). For V-estimators, one could integrate over the first argument in $a[x_1,\rho(z_2,\theta)]$ rather than the second. In the consumer surplus example, one could set $\gamma=h$ rather than equal to the separate numerator and denominator terms. This flexibility is useful, because it allows the estimator to be set up in a way

that is most convenient for verifying the regularity conditions for asymptotic normality.

This section will focus on conditions for asymptotic normality, taking consistency as given, similarly to Section 6. Consistency can often be shown by applying Theorem 2.1 directly, e.g. with uniform convergence resulting from application of Lemma 2.4. Also, when $g(z, \theta, \gamma)$ is linear in θ, as in the consumer surplus example, then consistency is not needed for the asymptotic normality arguments.

8.1. *Asymptotic normality and consistent variance estimation*

To motivate the precise results to be given, it is helpful to consider an expansion for $\hat{\theta}$. Expanding eq. (8.1) and solving for $\sqrt{n}(\hat{\theta} - \theta_0)$ gives

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\left[n^{-1} \sum_{i=1}^{n} \nabla_\theta g(z_i, \bar{\theta}, \hat{\gamma}) \right]^{-1} \sum_{i=1}^{n} g(z_i, \hat{\gamma})/\sqrt{n}, \qquad g(z, \gamma) = g(z, \theta_0, \gamma), \tag{8.3}$$

where $\bar{\theta}$ is the mean value. The usual (uniform) convergence arguments, when combined with consistency of $\bar{\theta}$ and $\hat{\gamma}$, suggest that $n^{-1} \sum_{i=1}^{n} \nabla_\theta g(z_i, \bar{\theta}, \hat{\gamma}) \xrightarrow{p} E[\nabla_\theta g(z, \theta_0, \gamma_0)] = G_\theta$. Thus, the behavior of the Jacobian term in eq. (8.3) is not conceptually difficult, only technically difficult because of the presence of nonparametric estimates. The score term $\sum_{i=1}^{n} g(z_i, \hat{\gamma})/\sqrt{n}$ is much more interesting and difficult. Showing asymptotic normality requires accounting for the presence of the infinite-dimensional term $\hat{\gamma}$. Section 6 shows how to do this for the finite-dimensional case, by expanding around the true value and using an influence function representation for $\hat{\gamma}$. The infinite-dimensional case requires a significant generalization. One such is given in the next result, from Newey (1992a). Let $\| \gamma \|$ denote a norm, such as $\sup_{x \in [a,b]} \| \gamma(x) \|$.

Theorem 8.1

Suppose that $E[g(z, \gamma_0)] = 0$, $E[\| g(z, \gamma_0) \|^2] < \infty$, and there is $\delta(z)$ with $E[\delta(z)] = 0$, $E[\| \delta(z) \|^2] < \infty$, and (i) (linearization) there is a function $G(z, \gamma - \gamma_0)$ that is linear in $\gamma - \gamma_0$ such that for all γ with $\| \gamma - \gamma_0 \|$ small enough, $\| g(z, \gamma) - g(z, \gamma_0) - G(z, \gamma - \gamma_0) \| \leq b(z) \| \gamma - \gamma_0 \|^2$, and $E[b(z)] \sqrt{n} \| \hat{\gamma} - \gamma_0 \|^2 \xrightarrow{p} 0$; (ii) (stochastic equicontinuity) $\sum_{i=1}^{n} [G(z_i, \hat{\gamma} - \gamma_0) - \int G(z, \hat{\gamma} - \gamma_0) dF_0]/\sqrt{n} \xrightarrow{p} 0$; (iii) (mean-square differentiability) there is $\delta(z)$ and a measure $\hat{F}$ such that $E[\delta(z)] = 0$, $E[\| \delta(z) \|^2] < \infty$ and for all $\| \gamma - \gamma_0 \|$ small enough, $\int G(z, \hat{\gamma} - \gamma_0) dF_0 = \int \delta(z) d\hat{F}$; (iv) for the empirical distribution $\tilde{F}$ $[\tilde{F}(z) = n^{-1} \sum_{i=1}^{n} 1(z_i \leq z)]$, $\sqrt{n}[\int \delta(z) d\hat{F} - \int \delta(z) d\tilde{F}] \xrightarrow{p} 0$. Then $\sum_{i=1}^{n} g(z_i, \hat{\gamma})/\sqrt{n} \xrightarrow{p} N(0, \Omega)$, where $\Omega = \mathrm{Var}[g(z_i, \gamma_0) + \delta(z_i)]$.

Proof

It follows by the triangle inequality that $\sum_{i=1}^{n}[g(z_i,\hat{\gamma})-g(z_i,\gamma_0)-\delta(z_i)]/\sqrt{n}\xrightarrow{p}0$, and by the central limit theorem that $\sum_{i=1}^{n}[g(z_i,\gamma_0)+\delta(z_i)]/\sqrt{n}\xrightarrow{d}N(0,\Omega)$.

Q.E.D.

This result is just a decomposition of the remainder term $\sum_{i=1}^{n}g(z_i,\hat{\gamma})/\sqrt{n}-\sum_{i=1}^{n}[g(z_i,\gamma_0)+\delta(z_i)]/\sqrt{n}$. As will be illustrated for the examples, it provides a useful outline of how asymptotic normality of a semiparametric two-step estimator can be shown. In addition, the assumptions of this result are useful for understanding how $\sum_{i=1}^{n}g(z_i,\hat{\gamma})/\sqrt{n}$ can have a limiting distribution, even though $\hat{\gamma}$ is not $\sqrt{n}$-consistent.

Assumption (i) requires that the remainder term from a linearization be small. The remainder term in this condition is analogous to $g(z,\gamma)-g(z,\gamma_0)-[\nabla_\gamma g(z,\gamma_0)](\gamma-\gamma_0)$ from parametric, two-step estimators. Here the functional $G(z,\gamma-\gamma_0)$ takes the place of $[\nabla_\gamma g(z,\gamma_0)](\gamma-\gamma_0)$. The condition on this remainder requires either that it be zero, where $b(z)=0$, or that the convergence rate of $\hat{\gamma}$ be faster than $n^{-1/4}$, in terms of the norm $\|\gamma\|$. Often such a convergence rate will require that the underlying nonparametric function satisfy certain smoothness restrictions, as further discussed in Section 8.3.

Assumption (ii) is analogous to the requirement for parametric two-step estimators that $\{n^{-1}\sum_{i=1}^{n}\nabla_\gamma g(z_i,\gamma_0)-E[\nabla_\gamma g(z,\gamma_0)]\}(\hat{\gamma}-\gamma_0)$ converge to zero. It is referred to as a stochastic equicontinuity condition for similar reasons as condition (v) of Theorem 7.2. Andrews (1990) has recently given quite general sufficient conditions for condition (ii). Alternatively, it may be possible to show by direct calculation that condition (ii) holds, under weaker conditions than those given in Andrews (1990). For example, in the V-estimator example, condition (ii) is a well known projection result for V-statistics (or U-statistics), as further discussed in Section 8.2. For kernel estimators, condition (ii) will follow from combining a V-statistic projection and a condition that the bias goes to zero, as further discussed in Section 8.3.

Both conditions (i) and (ii) involve "second-order" terms. Thus, both of these conditions are "regularity conditions", meaning that they should be satisfied if $g(z,\gamma)$ is sufficiently smooth and $\hat{\gamma}$ sufficiently well behaved. The terms in (iii) and (iv) are "first-order" terms. These conditions are the ones that allow $\sum_{i=1}^{n}g(z_i,\hat{\gamma})/\sqrt{n}$ to be asymptotically normal, even though $\hat{\gamma}$ may converge at a slower rate. The key condition is (iii), which imposes a representation of $\int G(z,\hat{\gamma}-\gamma_0)\mathrm{d}F_0$ as an integral with respect to an estimated measure. The interpretation of this representation is that $\int G(z,\hat{\gamma}-\gamma_0)\mathrm{d}F_0$ can be viewed as an average over some estimated distribution. As discussed in Newey (1992a), this condition is essentially equivalent to finiteness of the semiparametric variance bound for estimation of $\int G(z,\gamma-\gamma_0)\mathrm{d}F_0$. It is referred to as "mean-square differentiability" because the representation as an integral $\int\delta(z)\mathrm{d}F(z,\gamma)$ means that if $\mathrm{d}F(z,\gamma)^{1/2}$ has a mean-square derivative then

$\int \delta(z)\mathrm{d}F(z,\gamma)$ will be differentiable in γ, as shown in Ibragimov and Has'minskii (1981). This is an essential condition for a finite semiparametric variance bound, as discussed in Van der Vaart (1991), which in turn is a necessary condition for $\sqrt{n}$-consistency of $\int G(z,\hat{\gamma}-\gamma_0)\mathrm{d}F_0$. If $\int G(z,\hat{\gamma}-\gamma_0)\mathrm{d}F_0$ cannot be viewed as an average over an estimated distribution, then it will not be $\sqrt{n}$-consistent. Thus, condition (iii) is the key one to obtaining $\sqrt{n}$-consistency.

Condition (iv) requires that the difference between the estimator $\hat{F}$ and the empirical distribution be small, in the sense of difference of integrals. This condition embodies a requirement that $\hat{F}$ be nonparametric, because otherwise it could not be close to the empirical measure. For kernel estimators it will turn out that part (iv) is a pure bias condition, requiring that a bias term goes to zero faster than $1/\sqrt{n}$. For other estimators this condition may not impose such a severe bias requirement, as for the series estimators discussed in Newey (1992a).

An implication of conditions (iii) and (iv) is that $\sqrt{n}\int \delta(z)\mathrm{d}(\hat{F}-F_0) = \int \delta(z)\mathrm{d}\sqrt{n}\cdot(\hat{F}-F_0)$ converges in distribution to a normal random vector, a key result. An alternative way to obtain this result is to show that $\sqrt{n}(\hat{F}-F_0)$ is a stochastic process that converges in distribution in a metric for which $\int \delta(z)\mathrm{d}(\cdot)$ is continuous, and then apply the continuous mapping theorem.[46] This approach is followed in Ait-Sahalia (1993).

One piece of knowledge that is useful in verifying the conditions of Theorem 8.1 is the form of $\delta(z)$. As discussed in Newey (1992a), a straightforward derivative calculation is often useful for finding $\delta(z)$. Let η denote the parameters of some general distribution where η_0 is equal to the truth, and let $\gamma(\eta)$ denote the true value of γ when η are the true parameters. The calculation is to find $\delta(z)$ such that $\nabla_\eta \int g[z,\gamma(\eta)]\mathrm{d}F_0 = E[\delta(z)S_\eta']$, where the derivative is taken at the true distribution. The reason that this reproduces the $\delta(z)$ of Theorem 8.1 is that condition (i) will imply that $\nabla_\eta \int g[z,\gamma(\eta)]\mathrm{d}F_0 = \nabla_\eta \int G[z,\gamma(\eta)-\gamma_0]\mathrm{d}F_0$ [under the regularity condition that $\|\gamma(\eta)-\gamma\|$ is a differentiable function of η], so (iii) implies that $\nabla_\eta \int g[z,\gamma(\eta)]\mathrm{d}F_0 = \nabla_\eta \int \delta(z)\mathrm{d}F(\eta) = E[\delta(z)S_\eta']$. This calculation is like the Gateaux derivative calculation discussed in Huber (1981), except that it allows for the distributions to be continuous in some variables. With $\delta(z)$ in hand, one can then proceed to check the conditions of Theorem 8.1. This calculation is even useful when some result other than Theorem 8.1 is used to show asymptotic normality, because it leads to the form of the remainder term $\sum_{i=1}^n [g(z_i,\hat{\gamma})-g(z_i,\gamma_0)-\delta(z_i)]/\sqrt{n}$ that should be small to get asymptotic normality.

Theorem 8.1 can be combined with conditions for convergence of the Jacobian to obtain conditions for asymptotic normality of $\hat{\theta}$, as in the following result.

[46] The continuous mapping theorem states that if $Y(n) \xrightarrow{\mathrm{d}} Z$ and $h(y)$ is continuous on the support of Z then $h[Y(n)] \xrightarrow{\mathrm{d}} h(Z)$.

Theorem 8.2

If $\hat{\theta} \xrightarrow{p} \theta_0$, the conditions of Theorem 8.1 are satisfied, and (i) there are a norm $\|\gamma\|$, $\varepsilon > 0$, and a neighborhood $\mathcal{N}$ of θ_0 such that for $\|\gamma - \gamma_0\|$ small enough, $\sup_{\theta \in \mathcal{N}} \|\nabla_\theta g(z, \theta, \gamma) - \nabla_\theta g(z_i, \theta, \gamma_0)\| \leq b(z)\|\gamma - \gamma_0\|^\varepsilon$ and $E[b(z)]\|\hat{\gamma} - \gamma_0\|^\varepsilon \xrightarrow{p} 0$; (ii) $\nabla_\theta g(z_i, \theta, \gamma_0)$ satisfies the conditions of Lemma 4.3; (iii) G_θ is nonsingular; then $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, G_\theta^{-1} \Omega G_\theta^{-1\prime})$.

Proof

It suffices to show that $n^{-1}\sum_{i=1}^n \nabla_\theta g(z_i, \bar{\theta}, \hat{\gamma}) \xrightarrow{p} G_\theta$, because then the conclusion will follow from the conclusion of Theorem 8.1, eq. (8.3), and arguments like those of Section 3. Condition (i) implies that $[\sum_{i=1}^n b(z_i)/n]\|\hat{\gamma} - \gamma_0\|^\varepsilon \xrightarrow{p} 0$ by the Markov inequality, so $n^{-1}\sum_{i=1}^n \|\nabla_\theta g(z_i, \bar{\theta}, \hat{\gamma}) - \nabla_\theta g(z_i, \bar{\theta}, \gamma_0)\| \leq [n^{-1}\sum_{i=1}^n b(z_i)]\|\hat{\gamma} - \gamma_0\|^\varepsilon \xrightarrow{p} 0$. Also, by the conclusion of Lemma 4.3, $n^{-1}\sum_{i=1}^n \nabla_\theta g(z_i, \bar{\theta}, \gamma_0) \xrightarrow{p} G_\theta$. The conclusion then follows by the triangle inequality. Q.E.D.

This result provides one set of sufficient conditions for convergence of the Jacobian term. They are specified so as to be similar to those of Theorem 8.1, involving a norm for γ. In particular cases it may be useful to employ some other method for showing Jacobian convergence, as will be illustrated in Section 8.2. A similar comment applies to the consistency condition. Consistency can be shown by imposing conditions like (i) and (ii) to give uniform convergence of an objective function, but this result will not cover all cases. In some cases it may be better to work directly with Theorem 2.1 to show consistency.

The asymptotic variance of a semiparametric two-step estimator is $G_\theta^{-1}\Omega G_\theta^{-1\prime}$. As usual, a consistent estimator can be formed by plugging in estimators of the different pieces. An estimator of the Jacobian term can be formed in a straightforward way, as

$$\hat{G}_\theta = n^{-1}\sum_{i=1}^n \nabla_\theta g(z_i, \hat{\theta}, \hat{\gamma}). \tag{8.4}$$

Consistency of $\hat{G}_\theta$ for G_θ will follow under the same conditions as used for asymptotic normality of $\hat{\theta}$, because of the need to show consistency of the Jacobian matrix in the Taylor expansion. The more difficult term to estimate is the "score" variance Ω. One way to estimate this term is to form an estimator $\hat{\delta}(z)$ of the function $\delta(z)$ that appears in the asymptotic variance, and then construct

$$\hat{\Omega} = n^{-1}\sum_{i=1}^n \{g(z_i, \hat{\theta}, \hat{\gamma}) + \hat{\delta}(z_i)\}\{g(z_i, \hat{\theta}, \hat{\gamma}) + \hat{\delta}(z_i)\}'. \tag{8.5}$$

An estimator of the asymptotic variance can then be formed as $\hat{G}_\theta^{-1}\hat{\Omega}\hat{G}_\theta^{-1\prime}$.

It is difficult at this level of generality to give primitive conditions for consistency of a variance estimator, because these will depend on the nature of $\hat{\delta}(z)$. One useful intermediate result is the following one.

Lemma 8.3

If the conditions of Theorem 8.1 are satisfied, $\sum_{i=1}^n \| g(z_i, \hat{\theta}, \hat{\gamma}) - g(z_i, \theta_0, \gamma_0) \|^2/n \xrightarrow{p} 0$, and $\sum_{i=1}^n \| \hat{\delta}(z_i) - \delta(z_i) \|^2/n \xrightarrow{p} 0$, then $\hat{\Omega} \xrightarrow{p} \Omega$.

Proof

Let $\hat{u}_i = g(z_i, \hat{\theta}, \hat{\gamma}) + \hat{\delta}(z_i)$ and $u_i = g(z_i, \theta_0, \gamma_0) + \delta(z_i)$, so that $\Omega = E[u_i u_i']$ and $\hat{\Omega} = \sum_{i=1}^n \hat{u}_i \hat{u}_i'/n$. By the assumptions and the triangle inequality, $\sum_{i=1}^n \| \hat{u}_i - u_i \|^2/n \xrightarrow{p} 0$. Also, by the LLN, $\sum_{i=1}^n u_i u_i'/n \xrightarrow{p} E[u_i u_i']$. Also, $\| \sum_{i=1}^n \hat{u}_i \hat{u}_i'/n - \sum_{i=1}^n u_i u_i'/n \| \leq \sum_{i=1}^n \| \hat{u}_i \hat{u}_i' - u_i u_i' \|/n \leq \sum_{i=1}^n \| \hat{u}_i - u_i \|^2/n + 2\sum_{i=1}^n \| u_i \| \, \| \hat{u}_i - u_i \|/n \leq \sum_{i=1}^n \| \hat{u}_i - u_i \|^2/n + 2(\sum_{i=1}^n \| u_i \|^2/n)^{1/2} (\sum_{i=1}^n \| \hat{u}_i - u_i \|^2/n)^{1/2} \xrightarrow{p} 0$, because convergence of the diagonal elements of $\sum_{i=1}^n u_i u_i'/n$ implies that $\sum_{i=1}^n \| u_i \|^2/n$ is bounded in probability. Q.E.D.

Powell et al. (1989) use an analogous intermediate result to show consistency of their variance estimator. More primitive conditions are not given because it is difficult to specify them in a way that would cover all examples of interest.

These results provide a useful way of organizing and understanding asymptotic normality of semiparametric two-step estimators. In the analysis to follow, their usefulness will be illustrated by considering V-estimators and estimators where the first step is a kernel estimator. These results are also useful in showing asymptotic normality when the first step is a series regression estimator, i.e. an estimator obtained from least squares regression of some dependent variable on approximating functions. The series estimator case is considered in Newey (1992a).

8.2. *V-estimators*

A V-estimator, as in eq. (8.2), is useful as an illustration of the results. As previously noted, this estimator has $g(z, \gamma) = \int m(z, \tilde{z}, \theta_0) \mathrm{d}\gamma(\tilde{z})$, and $\hat{\gamma}$ is the empirical distribution with $\hat{\gamma}(\tilde{z}) = \sum_{i=1}^n 1(z_i \leq \tilde{z})/n$. For this estimator, condition (i) of Theorem 8.1 is automatically satisfied, with $b(z) = 0$ because $g(z, \gamma)$ is linear in γ. Condition (ii) needs to be verified. To see what this condition means, let $m(z_1, z_2) = m(z_2, z_2, \theta_0)$, $m_1(z) = \int m(z, \tilde{z}) \mathrm{d}F_0(\tilde{z})$, $m_2(z) = \int m(\tilde{z}, z) \mathrm{d}F_0(\tilde{z})$, and $\mu = \int \int m(z, \tilde{z}) \mathrm{d}F_0(z) \mathrm{d}F_0(\tilde{z})$. Then

$$\sum_{i=1}^n [G(z_i, \hat{\gamma} - \gamma_0) - \int G(z, \hat{\gamma} - \gamma_0) \mathrm{d}F_0]/\sqrt{n}$$

$$= \sqrt{n} \left\{ n^{-1} \sum_{i=1}^n \left[n^{-1} \sum_{j=1}^n m(z_i, z_j) - m_1(z_i) \right] - \left[n^{-1} \sum_{i=1}^n m_2(z_i) - \mu \right] \right\}$$

$$= \sqrt{n}\left\{ n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} m(z_i, z_j) - \mu - n^{-1} \sum_{i=1}^{n} [m_1(z_i) + m_2(z_i) - 2\mu] \right\}. \tag{8.6}$$

It will follow from U- and V-statistic theory that this remainder term is small.

A U-statistic has the form $\hat{U} = n^{-1}(n-1)^{-1} \sum_{i<j} a(z_i, z_j)$, where $a(z_1, z_2) = a(z_2, z_1)$. A V-statistic has the form $\hat{V} = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} m(z_1, z_j)$. A V-statistic is equal to a U-statistic plus an asymptotically negligible term, as in $\hat{V} = n^{-2} \sum_{i=1}^{n} m(z_i, z_i) + [(n-1)/n]\hat{U}$, where $a(z_i, z_j) = m(z_i, z_j) + m(z_j, z_i)$. The lead term, $n^{-2} \sum_{i=1}^{n} m(z_i, z_i)$ is a negligible "own observations" term, that converges in probability to zero at the rate $1/n$ as long as $E[m(z_i, z_i)]$ is finite.

The condition that the remainder term in eq. (8.6) have probability limit zero is known as the *projection* theorem for U- or V-statistics. For a U-statistic, $\bar{a}(z) = \int a(z, \tilde{z}) \mathrm{d}F_0(\tilde{z})$, and $E[\bar{a}(z)] = 0$, the projection theorem states if the data are i.i.d. and $a(z_1, z_2)$ has finite second moments, then $\sqrt{n}[\hat{U} - n^{-1} \sum_{i=1}^{n} \bar{a}(z_i)] \xrightarrow{p} 0$, where $n^{-1} \sum_{i=1}^{n} \bar{a}(z_i)$ is referred to as the projection of the U-statistic on the basic observations; see Serfling (1980). The V-statistic projection theorem states that the remainder in eq. (8.6) converges in probability to zero. The V-statistic projection theorem is implied by the U-statistic projection theorem, as can be shown in the following way. Let $a(z_1, z_2) = m(z_1, z_2) + m(z_2, z_1) - 2\mu$, so

$$n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} [m(z_i, z_j) - \mu] = n^{-2} \sum_{i=1}^{n} [m(z_i, z_i) - \mu] + [(n-1)/n]\hat{U}.$$

The first term following the equality should be negligible. The second term following the equality is a multiple of the U-statistic, where the multiplying constant converges to 1. Furthermore, $\bar{a}(z) = m_1(z) + m_2(z) - 2\mu$ in this case, so the projection of the U-statistic on the basic observations is $n^{-1} \sum_{i=1}^{n} [m_1(z_i) + m_2(z_i) - 2\mu]$. The U-statistic projection theorem then implies that the remainder in eq. (8.6) is small. Thus, it will follow from eq. (8.6) and the U-statistic projection theorem that condition (ii) of Theorem 8.1 is satisfied.

The previous discussion indicates that, for V-estimators, assumption (ii) follows from the V-statistic projection theorem. This projection result will also be important for assumption (ii) for kernel estimators, although in that case the V-statistic varies with the sample size. For this reason it is helpful to allow for $m(z_1, z_2)$ to depend on n when stating a precise result. Let $m_{n1}(z) = \int m_n(z, \tilde{z}) \mathrm{d}F_0(\tilde{z})$, $m_{n2}(z) = \int m_n(\tilde{z}, z) \mathrm{d}F_0(\tilde{z})$, and $Y_n = O_p(r_n)$ mean that $\| Y_n \| / r_n$ is bounded in probability for the Euclidean norm $\| \cdot \|$.

Lemma 8.4

If $z_1, z_2, \ldots$ are i.i.d. then $n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} m_n(z_i, z_j) - n^{-1} \sum_{i=1}^{n} [m_{n1}(z_i) + m_{n2}(z_i)] + \mu = O_p\{E[\| m_n(z_1, z_1) \|]/n + (E[\| m_n(z_1, z_2) \|^2])^{1/2}/n\}$.

The proof is technical, and so is postponed until Section 8.4. A consequence of this result is that condition (ii), the stochastic equicontinuity hypothesis, will be satisfied for U-estimators as long as $E[\|m(z_1, z_1, \theta_0)\|]$ and $E[\|m(z_1, z_2, \theta_0)\|^2]$ are finite. Lemma 8.4 actually gives a stronger result, that the convergence rate of the remainder is $1/\sqrt{n}$, but this result will not be used until Section 8.3.

With condition (ii) (finally) out of the way, one can consider conditions (iii) and (iv) for V-estimators. Assuming that $\mu = 0$, note that $\int G(z, \hat{\gamma} - \gamma_0) dF_0 = \int [\int m(z, \tilde{z}, \theta_0) \times dF_0(z)] d\tilde{F}(\tilde{z}) = \int \delta(z) d\tilde{F}(z)$ for $\delta(z) = m_2(z) = \int m(\tilde{z}, z, \theta_0) dF_0(z)$ and $\tilde{F}(z)$ equal to the empirical distribution. Thus, in this example conditions (iii) and (iv) are automatically satisfied because of the form of the estimator, giving all the assumptions of Theorem 8.1, with $g(z, \theta_0, \gamma_0) + \delta(z) = m_1(z) + m_2(z)$. An asymptotic normality result for V-estimators can then be stated by specifying conditions for uniform convergence of the Jacobian. The following condition is useful in this respect, and is also useful for showing the uniform convergence assumption of Theorem 2.1 and V-estimators.

Lemma 8.5

If $z_1, z_2, \ldots$ are i.i.d., $a(z_1, z_2, \theta)$, is continuous at each $\theta \in \Theta$ with probability one, $E[\sup_{\theta \in \Theta} \|a(z_1, z_1, \theta)\|] < \infty$, and $E[\sup_{\theta \in \Theta} \|a(z_1, z_2, \theta)\|] < \infty$, then $E[a(z_1, z_2, \theta)]$ is continuous in $\theta \in \Theta$, and $\sup_{\theta \in \Theta} \|n^{-2} \sum_{i=1}^n \sum_{j=1}^n a(z_i, z_j, \theta) - Ea(z_1, z_2, \theta)]\| \xrightarrow{p} 0$.

The proof is postponed until Section 8.4.

This result can be used to formulate conditions for asymptotic normality by adding a condition for convergence of the Jacobian.

Theorem 8.6

Suppose that $z_1, z_2, \ldots$ are i.i.d., $\hat{\theta} \xrightarrow{p} \theta_0$, (i) $E[m(z_1, z_2, \theta_0)] = 0$, $E[\|m(z_1, z_1, \theta_0)\|] < \infty$, $E[\|m(z_1, z_2, \theta_0)\|^2] < \infty$, (ii) $m(z_1, z_1, \theta)$ and $m(z_1, z_2, \theta)$ are continuously differentiable on a neighborhood of θ_0 with probability one, and there is a neighborhood $\mathcal{N}$ of θ_0, such that $E[\sup_{\theta \in \mathcal{N}} \|\nabla_\theta m(z_1, z_1, \theta)\|] < \infty$ and $E[\sup_{\theta \in \mathcal{N}} \|\nabla_\theta m(z_1, z_2, \theta)\|] < \infty$, (iii) $G_\theta = E[\nabla_\theta m(z_1, z_2, \theta_0)]$ is nonsingular. Then $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, G_\theta^{-1} \Omega G_\theta^{-1\prime})$ for $\Omega = \mathrm{Var}\{\int [m(z, \tilde{z}, \theta_0) + m(\tilde{z}, z, \theta_0)] dF_0(\tilde{z})\}$.

Proof

It follows by Lemma 8.4, assumption (i), and the preceding discussion that conditions (i)–(iv) of Theorem 8.1 are satisfied for $g(z, \gamma_0) + \delta(z) = \int [m(z, \tilde{z}, \theta_0) + m(\tilde{z}, z, \theta_0)] dF_0(\tilde{z})$, so it follows by the conclusion of Theorem 8.1 that $\sqrt{n} n^{-2} \sum_{i=1}^n \sum_{j=1}^n m(z_i, z_j, \theta_0) \xrightarrow{d} N(0, \Omega)$. Therefore, it suffices to show that $n^{-2} \sum_{i=1}^n \sum_{j=1}^n \nabla_\theta m(z_i, z_j, \bar{\theta}) \xrightarrow{p} G_\theta$ for any $\bar{\theta} \xrightarrow{p} \theta_0$. This condition follows by Lemma 8.5 and the triangle inequality. Q.E.D.

To use this result to make inferences about $\hat{\theta}$ it is useful to have an asymptotic variance estimator. Let $\hat{G}_\theta = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \nabla_\theta m(z_i, z_j, \hat{\theta})$ be a Jacobian estimator.

This estimator wll be consistent for G_θ under the conditions of Theorem 8.6. An estimator of $g(z,\theta_0,\gamma_0)+\delta(z)$ can be constructed by replacing θ_0 by $\hat\theta$ and F_0 by $\hat F$ in the expression given in Ω, to form

$$\hat u_i = n^{-1}\sum_{j=1}^{n}[m(z_i,z_j,\hat\theta)+m(z_j,z_i,\hat\theta)],\quad \hat\Omega = n^{-1}\sum_{i=1}^{n}\hat u_i\hat u_i'.$$

The following result is useful for showing consistency of this estimator. Let $m_n(z_1,z_2,\theta)$ depend on n and $m_{n1}(z)$ be as defined above.

Lemma 8.7

If $\sqrt{n}\|\hat\theta-\theta_0\|=O_p(1)$ then $n^{-1}\sum_{i=1}^{n}\|n^{-1}\sum_{j=1}^{n}m_n(z_i,z_j,\hat\theta)-m_{n1}(z_i)\|^2=O_p\{n^{-1}\times E[\sup_{\theta\in\mathcal{N}}\|m_n(z_1,z_1,\theta)\|^2+\sup_{\theta\in\mathcal{N}}\|\nabla_\theta m_n(z_1,z_2,\theta)\|^2+\|m_n(z_1,z_2,\theta_0)\|^2]\}$.

This result is proved in Section 8.4. Consistency of the variance estimator can now be shown, using Lemma 8.7.

Theorem 8.8

If the conditions of Theorem 8.6 are satisfied, $E[\sup_{\theta\in\mathcal{N}}\|m(z_1,z_1,\theta)\|^2]<\infty$ and $E[\sup_{\theta\in\mathcal{N}}\|\nabla_\theta m_n(z_1,z_2,\theta)\|^2]<\infty$ then $\hat G_\theta^{-1}\hat\Omega\hat G_\theta^{-1}\xrightarrow{p}G_\theta^{-1}\Omega G_\theta^{-1}$.

Proof

It follows by Lemmas 8.7 and 8.3 that $\hat\Omega\xrightarrow{p}\Omega$, and it follows as in the proof of Theorem 8.6 that $\hat G_\theta^{-1}\xrightarrow{p}G_\theta^{-1}$, so the conclusion follows by continuity of matrix multiplication. Q.E.D.

8.3. *First-step kernel estimation*

There are many examples of semiparametric two-step estimators that depend on kernel density or conditional expectations estimators. These include the estimators of Powell et al. (1989) and Robinson (1988b). Also, the nonparametric approximate consumer surplus estimator introduced earlier is of this form. For these estimators it is possible to formulate primitive assumptions for asymptotic normality, based on the conditions of Section 8.1.

Suppose that γ denotes a vector of functions of variables x, where x is an $r\times 1$ subvector of the data observation z. Let y denote another subvector of the data. The first-step estimator to be considered here will be the function of x with

$$\hat\gamma(x)=n^{-1}\sum_{i=1}^{n}y_iK_\sigma(x-x_i). \tag{8.7}$$

This is a kernel estimator of $f_0(x)E[y|x]$, where $f_0(x)$ is the marginal density of x. A kernel estimator of the density of x will be a component of $\hat{\gamma}(x)$ where the corresponding component of y is identically equal to 1. The nonparametric consumer surplus estimator depends on $\hat{\gamma}$ of this form, where $y_i = (1, q_i)'$.

Unlike V-estimators, two-step estimators that depend on the $\hat{\gamma}$ of eq. (8.6) will often be nonlinear in $\hat{\gamma}$. Consequently, the linearization condition (i) of Theorem 8.1 will be important for these estimators. For example, the nonparametric consumer surplus estimator depends on a ratio, with $g(z,\gamma) = \int_a^b [\gamma_2(x)/\gamma_1(x)]dx - \theta_0$. In this example the linearization $G(z, \gamma - \gamma_0)$ is obtained by expanding the ratio inside the integral. By $\tilde{a}/\tilde{b} - a/b = b^{-1}[1 - \tilde{b}^{-1}(\tilde{b} - b)][\tilde{a} - a - (a/b)(\tilde{b} - b)]$, the linearization of $\tilde{a}/\tilde{b}$ around a/b is $b^{-1}[\tilde{a} - a - (a/b)(\tilde{b} - b)]$. Therefore, the linear functional of assumption (i) is

$$G(z,\gamma) = \int_a^b f_0(x)^{-1}[-h_0(x), 1]\gamma(x)dx. \tag{8.8}$$

If $\gamma_{10}(x) = f_0(x)$ is bounded away from zero, $\gamma_{20}(x)$ is bounded, and $\gamma_1(x)$ is uniformly close to $\gamma_{10}(x)$ on $[a,b]$, then the remainder term will satisfy

$$\begin{aligned} &|g(z,\gamma) - g(z,\gamma_0) - G(z,\gamma - \gamma_0)| \\ &\quad \leqslant \int_a^b |\gamma_1(x)|^{-1} f_0(x)^{-1}[1 + |h_0(x)|][|\gamma_1(x) - f_0(x)|^2 + |\gamma_2(x) - \gamma_{20}(x)|^2]dx \\ &\quad \leqslant C \sup_{x\in[a,b]} \|\gamma(x) - \gamma_0(x)\|^2. \end{aligned} \tag{8.9}$$

Therefore assumption (i) of Theorem 8.1 will be satisfied if $\sqrt{n}\sup_{x\in[a,b]}\|\hat{\gamma}(x) - \gamma_0(x)\|^2 \xrightarrow{p} 0$.[47]

One feature of the consumer surplus example that is shared by other cases where conditional expectations are present is that the density in the denominator must be bounded away from zero in order for the remainder to be well behaved. This condition requires that the density only effects the estimator through its values on a bounded set, a "fixed trimming" condition, where the word trimming refers to limiting the effect of the density. In some examples, such as the consumer surplus one, this fixed trimming condition arises naturally, because the estimator only depends on x over a range of values. In other cases it may be necessary to guarantee that this condition holds by adding a weight function, as in the weighted average derivative example below. It may be possible to avoid this assumption, using results like those of Robinson (1988b), where the amount of trimming is allowed to decrease with sample size, but for simplicity this generalization is not considered here.

[47] In this case $\hat{\gamma}_1(x)$ will be uniformly close to $\gamma_{10}(x)$, and so will be bounded away from zero with probability approaching one if $\gamma_{10}(x)$ is bounded away from zero, on $[a,b]$.

In general, to check the linearization condition (i) of Theorem 8.1 it is necessary to specify a norm for the function γ. A norm that is quite convenient and applies to many examples is a supremum norm on a function and its derivatives. This norm does not give quite as sharp results as an integral norm, but it applies to many more examples, and one does not lose very much in working with a supremum norm rather than an integral norm.[48]

Let $\partial^j\gamma(x)/\partial x^j$ denote any vector consisting of all distinct jth-order partial derivatives of all elements of $\gamma(x)$. Also, let $\mathscr{X}$ denote a set that is contained in the support of x, and for some nonnegative integer d let

$$\|\gamma\| \equiv \max_{\ell \leqslant d} \sup_{x\in\mathscr{X}} \|\partial^\ell \gamma(x)/\partial x^\ell\|.$$

This type of norm is often referred to as a Sobolev norm.

With this norm the $n^{1/4}$ convergence rate of Theorem 8.1 will hold if the kernel estimator $\hat{\gamma}(x)$ and its derivatives converge uniformly on $\mathscr{X}$ at a sufficiently fast rate. To make sure that the rate is attainable it is useful to impose some conditions on the kernel, the true function $\gamma_0(x)$, the data vector y, and the bandwidth. The first assumption gives some useful conditions for the kernel.

Assumption 8.1

$K(u)$ is differentiable of order d, the derivatives of order d are bounded, $K(u)$ is zero outside a bounded set, $\int \mathscr{K}(u)\mathrm{d}u = 1$, there is a positive integer m such that for all $j < m$, $\int K(u)[\bigotimes_{\ell=1}^{j} u]\mathrm{d}u = 0$.

The existence of the dth derivative of the kernel means that $\|\hat{\gamma}\|$ will be well defined. The requirement that $K(u)$ is zero outside a bounded set could probably be relaxed, but is maintained here for simplicity. The other two conditions are important for controlling the bias of the estimator. They can be explained by considering an expansion of the bias of $\hat{\gamma}(x)$. For simplicity, suppose that x is a scalar, and note $E[\hat{\gamma}(x)] = \int E[y|\tilde{x}]f_0(\tilde{x})K_\sigma(x-\tilde{x})\mathrm{d}\tilde{x} = \int \gamma_0(\tilde{x})K_\sigma(x-\tilde{x})\mathrm{d}\tilde{x}$. Making the change of variables $u = (x-\tilde{x})/\sigma$ and expanding around $\sigma = 0$ gives

$$\begin{aligned} E[\hat{\gamma}(x)] &= \int \gamma_0(x-\sigma u)K(u)\mathrm{d}u \\ &= \sum_{0\leqslant j<m} \sigma^j \partial^j\gamma_0(x)/\partial x^j \int K(u)u^j\mathrm{d}u + \sigma^m \int \partial^m\gamma_0(x+\bar{\sigma}u)/\partial x^m K(u)u^m\mathrm{d}u \\ &= \gamma_0(x) + \sigma^m \int \partial^m\gamma_0(x+\bar{\sigma}u)/\partial x^m K(u)u^m\mathrm{d}u, \end{aligned} \tag{8.10}$$

[48] With an integral norm, the $\ln n$ term in the results below could be dropped. The other terms dominate this one, so that this change would not result in much improvement.

where $\bar{\sigma}$ is an intermediate value, assuming that derivatives up to order m of $\gamma_0(x)$ exist. The role of $\int K(u)du = 1$ is to make the coefficient of $\gamma_0(x)$ equal to 1, in the expansion. The role of the "zero moment" condition $\int K(u)u^j du = 0$, $(j < m)$, is to make all of the lower-order powers of σ disappear, so that the difference between $E[\hat{\gamma}(x)]$ and $\gamma_0(x)$ is of order σ^m. Thus, the larger m is, with a corresponding number of derivatives of $\gamma_0(x)$, the faster will be the convergence rate of $E[\hat{\gamma}(x)]$ to $\gamma_0(x)$. Kernels with this moment property will have to be negative when $j \geq 2$. They are often referred to as "higher-order" or "bias-reducing" kernels. Such higher-order kernels are used to obtain the $n^{1/4}$ convergence rate for $\hat{\gamma}$ and are also important for assumption (iv) of Theorem 8.1.

In order to guarantee that bias-reducing kernels have the desired effect, the function being estimated must be sufficiently smooth. The following condition imposes such smoothness.

Assumption 8.2

There is a version of $\gamma_0(x)$ that is continuously differentiable to order d with bounded derivatives on an open set containing $\mathscr{X}$.

This assumption, when combined with Assumption 8.1 and the expansion given above produce the following result on the bias of the kernel estimator $\hat{\gamma}$. Let $E[\hat{\gamma}]$ denote $E[\hat{\gamma}(x)]$ as a function of x.

Lemma 8.9

If Assumptions 8.1 and 8.2 are satisfied then $\| E[\hat{\gamma}] - \gamma \| = O(\sigma^m)$.

This result is a standard one on kernel estimators, as described in Härdle and Linton (1994), so its proof is omitted.

To obtain a uniform convergence rate for $\hat{\gamma}$ is also helpful to impose the following condition.

Assumption 8.3

There is $p \geq 4$ such that $E[\| y \|^p] < \infty$ and $E[\| y \|^p | x] f_0(x)$ is bounded.

Assumptions 8.1–8.3 can be combined to obtain the following result:

Lemma 8.10

If Assumptions 8.1–8.3 are satisfied and $\sigma = \sigma(n)$ such that $\sigma(n) \to 0$ and $n^{1-(2/p)}\sigma(n)^r / \ln n \to \infty$ then $\| \hat{\gamma} - \gamma_0 \| = O_p[(\ln n)^{1/2}(n\sigma^{r+2d})^{-1/2} + \sigma^m]$.

This result is proved in Newey (1992b). Its proof is quite long and technical, and so is omitted. It follows from this result that $\sqrt{n}\| \hat{\gamma} - \gamma_0 \|^2 \xrightarrow{p} 0$, as required for as-

sumption (i) of Theorem 8.1, if $n^{1-(2/p)}\sigma(n)^r/\ln n \to \infty$, $\sqrt{n}\sigma^{2m} \to 0$, and $\sqrt{n}\ln n/(n\sigma^{r+2d}) \to 0$. These conditions will be satisfied for a range of bandwidth sequences $\sigma(n)$, if m and p are big enough, i.e. if the kernel is of "high-enough order", the true function $\gamma(x)$ is smooth enough, and there are enough moments of y. However, large values of m will be required if r is large.

For kernel estimators it turns out that assumption (ii) of Theorem 8.1 will follow from combining a V-statistic projection with a small bias condition. Suppose that $G(z,\gamma)$ is linear in γ, and let $\bar{\gamma} = E[\hat{\gamma}]$. Then $G(z,\hat{\gamma}-\gamma_0) = G(z,\hat{\gamma}-\bar{\gamma}) + G(z,\bar{\gamma}-\gamma_0)$. Let $m_n(z_i,z_j) = G[z_i, y_j K_\sigma(\cdot - x_j)]$, $m_{n2}(z) = \int m_n(\tilde{z},z)\,dF_0(\tilde{z}) = \int G[z, y_j K_\sigma(\cdot - x_j)]\,dF_0(z)$, and assume that $m_{n1}(z) = \int m_n(z,\tilde{z})\,dF_0(\tilde{z}) = G(z,\bar{\gamma})$, as should follow by the linearity of $G(z,\gamma)$. Then

$$\begin{aligned}
&\sqrt{n}\left\{\sum_{i=1}^n G(z_i,\hat{\gamma}-\bar{\gamma})/n - \int G(z,\hat{\gamma}-\bar{\gamma})\,dF_0(z)\right\} \\
&= \sqrt{n}\left\{\sum_{i=1}^n G(z_i,\hat{\gamma})/n - \sum_{i=1}^n G(z_i,\bar{\gamma})/n - \int G(z,\hat{\gamma})\,dF_0(z) + \int G(z,\bar{\gamma})\,dF_0(z)\right\} \\
&= \sqrt{n}\left\{n^{-2}\sum_{i=1}^n\sum_{j=1}^n m_n(z_i,z_j) - n^{-1}\sum_{i=1}^n m_{n1}(z_i) - n^{-1}\sum_{i=1}^n m_{n2}(z_i) + E[m_{n1}(z)]\right\},
\end{aligned} \tag{8.11}$$

where the second equality follows by linearity of $G(z,\gamma)$. The convergence in probability of this term to zero will follow by the V-statistic projection result of Lemma 8.4. The other term, $\sqrt{n}\{\sum_{i=1}^n G(z_i,\bar{\gamma}-\gamma_0)/n - \int G(z,\bar{\gamma}-\gamma)\,dF_0(z)\}$, will converge in probability to zero if $E[\|G(z,\bar{\gamma}-\gamma_0)\|^2] \to 0$, by Chebyshev's inequality, which should happen in great generality by $\bar{\gamma} \to \gamma_0$ as $\sigma \to 0$, as described precisely in the proof of Theorem 8.11 below. Thus, a V-statistic projection result when combined with a small bias condition that $E[\|G(z,\bar{\gamma}-\gamma_0)\|^2]$ goes to zero, gives condition (ii) of Theorem 8.1.

For kernel estimators, a simple condition for the mean-square differentiability assumption (iii) of Theorem 8.1 is that there is a conformable matrix $v(x)$ of functions of x such that

$$\int G(z,\gamma)\,dF_0 = \int v(x)\gamma(x)\,dx, \tag{8.12}$$

for some $v(x)$.This condition says $\int G(z,\gamma)\,dF_0$ can be represented as an integral, i.e. as an "average" over values of x. It leads to a simple form for $\delta(z)$. As previously discussed, in general $\delta(z)$ can be calculated by differentiating $\int G[z,\gamma(\eta)]\,dF_0$ with respect to the parameters η of a distribution of z, and finding $\delta(z)$ such that $\nabla_\eta \int G[z,\gamma(\eta)]\,dF_0 = E[\delta(z)S_\eta']$ for the score S_η and all sufficiently regular parametrizations. Let $E_\eta[\cdot]$ denote the expectation with respect to the distribution at this

parametrization. Here, the law of iterated expectations implies that

$$\int G[z,\gamma(\eta)]\,dF_0 = \int v(x)\gamma(x,\eta)\,dx = \int v(x)E_\eta[y|x]f(x|\eta)\,dx = E_\eta[v(x)y],$$

so differentiating gives $\nabla_\eta \int G[z,\gamma(\eta)]\,dF_0 = \nabla_\eta E_\eta[v(x)y] = E[v(x)yS'_\eta] = E[\delta(z)S'_\eta]$, for

$$\delta(z) = v(x)y - E[v(x)y]. \tag{8.13}$$

For example, for the consumer surplus estimator, by eq. (8.8), one has $v(x) = 1(a \leqslant x \leqslant b)f_0(x)^{-1}[-h_0(x),1]$ and $y = (1,q)$, so that $\delta(z) = 1(a \leqslant x \leqslant b)f_0(x)^{-1} \times [q - h_0(x)]$.

With a candidate for $\delta(z)$ in hand, it is easier to find the integral representation for assumption (iii) of Theorem 8.1. Partition z as $z = (x,w)$, where w are the components of z other than x. By a change of variables, $\int K_\sigma(x - x_i)\,dx = \int K(u)\,du = 1$, so that

$$\int G(z,\hat{\gamma} - \gamma_0)\,dF_0 = \int v(x)\hat{\gamma}(x)\,dx - \int v(x)\gamma_0(x)\,dx = n^{-1}\sum_{i=1}^{n}\int v(x)y_iK_\sigma(x - x_i)\,dx$$

$$- E[v(x)y] = n^{-1}\sum_{i=1}^{n}\int \delta(x,w_i)K_\sigma(x - x_i)\,dx = \int \delta(z)\,d\hat{F}, \tag{8.14}$$

where the integral of a function $a(z)$ over $d\hat{F}$ is equal to $n^{-1}\sum_{i=1}^{n}\int a(x,w_i)K_\sigma(x - x_i)dx$. The integral here will be the expectation over a distribution when $K(u) \geqslant 0$, but when $K(u)$ can be negative, as for higher-order kernels, then the integral cannot be interpreted as an expectation.

The final condition of Theorem 8.1, i.e. assumption (iv), will follow under straightforward conditions. To verify assumption (iv) of Theorem 8.1, it is useful to note that the integral in eq. (8.14) is close to the empirical measure, the main difference being that the empirical distribution of x has been replaced by a smoothed version with density $n^{-1}\sum_{i=1}^{n} K_\sigma(x - x_i)$ [for $K(u) \geqslant 0$]. Consequently, the difference between the two integrals can be interpreted as a smoothing bias term, with

$$\int \delta(z)\,d\hat{F} - \int \delta(z)\,d\tilde{F} = n^{-1}\sum_{i=1}^{n}\left[\int v(x)K_\sigma(x - x_i)\,dx - v(x_i)\right]y_i. \tag{8.15}$$

By Chebyshev's inequality, sufficient conditions for $\sqrt{n}$ times this term to converge in probability to zero are that $\sqrt{n}E[y_i\{\int v(x)K_\sigma(x - x_i)\,dx - v(x_i)\}] \to 0$ and that $E[\|y_i\|^2\|\int v(x)K_\sigma(x - x_i)\,dx - v(x_i)\|^2] \to 0$. As shown below, the bias-reducing kernel and smoothness parts of Assumptions 8.1–8.3 are useful in showing that the first

condition holds, while continuity of $v(x)$ at "most points" of $v(x)$ is useful for showing the second. In particular, one can show that the remainder term in eq. (8.15) is small, even when $v(x)$ is discontinuous, as is important in the consumer surplus example.

Putting together the various arguments described above leads to a result on asymptotic normality of the "score" $\sum_{i=1}^{n} g(z_i, \hat{\gamma})/\sqrt{n}$.

Theorem 8.11

Suppose that Assumptions 8.1–8.3 are satisfied, $E[g(z, \gamma_0)] = 0$, $E[\| g(z, \gamma_0) \|^2] < \infty$, $\mathcal{X}$ is a compact set, $\sigma = \sigma(n)$ with $n\sigma^{2r+4d}/(\ln n)^2 \to \infty$ and $n\sigma^{2m} \to 0$, and there is a vector of functionals $G(z, \gamma)$ that is linear in γ such that (i) for $\| \gamma - \gamma_0 \|$ small enough, $\| g(z, \gamma) - g(z, \gamma_0) - G(z, \gamma - \gamma_0) \| \leq b(z) \| \gamma - \gamma_0 \|^2$, $E[b(z)] < \infty$; (ii) $\| G(z, \gamma) \| \leq c(z) \| \gamma \|$ and $E[c(z)^2] < \infty$; (iii) there is $v(x)$ with $\int G(z, \gamma)\, dF_0(z) = \int v(x)\gamma(x)\, dx$ for all $\| \gamma \| < \infty$; (iv) $v(x)$ is continuous almost everywhere, $\int \| v(x) \|\, dx < \infty$, and there is $\varepsilon > 0$ such that $E[\sup_{\| v \| \leq \varepsilon} \| v(x + v) \|^4] < \infty$. Then for $\delta(z) = v(x)y - E[v(x)y]$, $\sum_{i=1}^{n} g(z_i, \hat{\gamma})/\sqrt{n} \xrightarrow{d} N\{0, \mathrm{Var}[g(z, \gamma_0) + \delta(z)]\}$.

Proof

The proof proceeds by verifying the conditions of Theorem 8.1. To show assumption (i) it suffices to show $\sqrt{n} \| \hat{\gamma} - \gamma_0 \|^2 \xrightarrow{p} 0$, which follows by the rate conditions on σ and Lemma 8.10. To show assumption (ii), note that by $K(u)$ having bounded derivatives of order d and bounded support, $\| G[z, yK_\sigma(\cdot - x)] \| \leq \sigma^{-r} c(z) \| y \|$. It then follows by Lemma 8.4 that the remainder term of eq. (8.11) is $O_p(n^{-1}\sigma^{-r} \times \{E[c(z_1) \| y_1 \|] + (E[c(z_1)^2 \| y_2 \|^2])^{1/2}\}) = o_p(1)$ by $n^{-1}\sigma^{-r} \to 0$. Also, the rate conditions imply $\sigma \to 0$, so that $E[\| G(z, \bar{\gamma} - \gamma_0) \|^2] \leq E[c(z)^2] \| \bar{\gamma} - \gamma_0 \|^2 \to 0$, so that the other remainder term for assumption (ii) also goes to zero, as discussed following eq. (8.11). Assumption (iii) was verified in the text, with $d\hat{F}$ as described there. To show assumption (iv), note that

$$
\begin{aligned}
&\left\| \sqrt{n} E\left[\left\{ \int v(x) K_\sigma(x - x_i)\, dx - v(x_i) \right\} y_i \right] \right\| \\
&\quad = \sqrt{n} \left\| \int \left\{ \int v(x + \sigma u) K(u)\, du \right\} \gamma_0(x)\, dx - \int v(x)\gamma_0(x)\, dx \right\| \\
&\quad = \sqrt{n} \left\| \int \int v(x) K(u) \gamma_0(x - \sigma u)\, du dx - \int v(x)\gamma_0(x)\, dx \right\| \\
&\quad = \sqrt{n} \left\| \int v(x) \left\{ \int [\gamma_0(x - \sigma u) - \gamma_0(x)] K(u)\, du \right\} dx \right\| \\
&\leq \sqrt{n} \int \| v(x) \| \left\| \int [\gamma_0(x - \sigma u) - \gamma_0(x)] K(u)\, du \right\| dx \leq C\sigma^m \int \| v(x) \|\, dx, \qquad (8.16)
\end{aligned}
$$

for some constant C. Therefore, $\|\sqrt{n}E[\{\int v(x)K_\sigma(x-x_i)\,dx - v(x_i)\}y_i]\| \leq C\sqrt{n}\sigma^m \to 0$. Also, by almost everywhere continuity of $v(x)$, $v(x+\sigma u)\to v(x)$ for almost all x and u. Also, on the bounded support of $K(u)$, for small enough σ, $v(x+\sigma u)\leq \sup_{\|v\|\leq\varepsilon} v(x+v)$, so by the dominated convergence theorem, $\int v(x+\sigma u)K(u)\,du \to \int v(x)K(u)\,du = v(x)$ for almost all x. Another application of the dominated convergence theorem, using boundedness of $K(u)$ gives $E[\|\int v(x)K_\sigma(x-x_i)\,dx - v(x_i)\|^4]\to 0$, so by the Cauchy–Schwartz inequality, $E[\|y_i\|^2\|\int v(x)K_\sigma(x-x_i)\,dx - v(x_i)\|^2]\to 0$. Condition (iv) then follows from the Chebyshev inequality, since the mean and variance of $n^{-1/2}\sum_{i=1}^n [\int v(x)K_\sigma(x-x_i)\,dx - v(x_i)]y_i$ go to zero. Q.E.D.

The assumptions of Theorem 8.11 can be combined with conditions for convergence of the Jacobian to obtain an asymptotic normality result with a first-step kernel estimator. As before, let $\Omega = \mathrm{Var}[g(z,\gamma_0)+\delta(z)]$.

Theorem 8.12

Suppose that $\hat\theta \xrightarrow{p} \theta_0 \in \mathrm{interior}(\Theta)$, the assumptions of Theorem 8.11 are satisfied, $E[g(z,\gamma_0)]=0$ and $E[\|g(z,\gamma_0)\|^2]<\infty$, for $\|\gamma-\gamma_0\|$ small enough, $g(z,\theta,\gamma)$ is continuously differentiable in θ on a neighborhood $\mathcal{N}$ of θ_0, there are $b(z)$, $\varepsilon>0$ with $E[b(z)]<\infty$, $\|\nabla_\theta g(z,\theta,\gamma)-\nabla_\theta g(z,\theta_0,\gamma_0)\| \leq b(z)[\|\theta-\theta_0\|^\varepsilon + \|\gamma-\gamma_0\|^\varepsilon]$, and $E[\nabla_\theta g(z,\theta_0,\gamma_0)]$ exists and is nonsingular. Then $\sqrt{n}(\hat\theta-\theta_0)\xrightarrow{d} N(0, G_\theta^{-1}\Omega G_\theta^{-1\prime})$.

Proof

It follows similarly to the proof of Theorem 8.2 that $\hat G_\theta^{-1}\xrightarrow{p} G_\theta^{-1}$, so the conclusion follows from Theorem 8.11 similarly to the proof of Theorem 8.2. Q.E.D.

As previously discussed, the asymptotic variance can be estimated by $\hat G_\theta^{-1}\hat\Omega\hat G_\theta^{-1\prime}$, where $\hat G_\theta = n^{-1}\sum_{i=1}^n \nabla_\theta g(z_i,\hat\theta,\hat\gamma)$ and $\hat\Omega = n^{-1}\sum_{i=1}^n \hat u_i\hat u_i'$ for $\hat u_i = g(z_i,\hat\theta,\hat\gamma)+\hat\delta(z_i)$. The main question here is how to construct an estimator of $\delta(z)$. Typically, the form of $\delta(z)$ will be known from assumption (iii) of Theorem 8.11, with $\delta(z)=\delta(z,\theta_0,\gamma_0)$ for some known function $\delta(z,\theta,\gamma)$. An estimator of $\delta(z)$ can then be formed by substituting $\hat\theta$ and $\hat\gamma$ for θ_0 and γ_0 to form

$$\hat\delta(z) = \delta(z,\hat\theta,\hat\gamma). \tag{8.17}$$

The following result gives regularity conditions for consistency of the corresponding asymptotic variance estimator.

Theorem 8.13

Suppose that the assumptions of Theorem 8.12 are satisfied and there are $b(z)$, $\varepsilon>0$, such that $E[b(z)^2]<\infty$ and for $\|\gamma-\gamma_0\|$ small enough, $\|g(z,\theta,\gamma)-g(z,\theta_0,\gamma)\|\leq b(z)\times[\|\theta-\theta_0\|^\varepsilon + \|\gamma-\gamma_0\|^\varepsilon]$ and $\|\delta(z,\theta,\gamma)-\delta(z,\theta_0,\gamma_0)\|\leq b(z)[\|\theta-\theta_0\|^\varepsilon + \|\gamma-\gamma_0\|^\varepsilon]$. Then $\hat G_\theta^{-1}\hat\Omega\hat G_\theta^{-1\prime}\xrightarrow{p} G_\theta^{-1}\Omega G_\theta^{-1\prime}$.

Proof

It suffices to show that the assumptions of Theorem 8.3 are satisfied. By the conditions of Theorem 8.12, $\|\hat{\theta} - \theta_0\| \xrightarrow{p} 0$ and $\|\hat{\gamma} - \gamma_0\| \xrightarrow{p} 0$, so with probability approaching one,

$$\sum_{i=1}^{n} \|g(z_i, \hat{\theta}, \hat{\gamma}) - g(z_i, \theta_0, \gamma_0)\|^2/n \leqslant \left(n^{-1}\sum_{i=1}^{n} b(z_i)^2\right)[\|\hat{\theta} - \theta_0\|^{\varepsilon} + \|\hat{\gamma} - \gamma_0\|^{\varepsilon}]^2 \xrightarrow{p} 0,$$

because $n^{-1}\sum_{i=1}^{n} b(z_i)^2$ is bounded in probability by the Markov inequality. It follows similarly that $\sum_{i=1}^{n} \|\hat{\delta}(z_i) - \delta(z_i)\|^2/n \xrightarrow{p} 0$, so the conclusion follows by Theorem 8.3. Q.E.D.

In some cases $\delta(z, \theta, \gamma)$ may be complex and difficult to calculate, making it hard to form the estimator $\delta(z, \hat{\theta}, \hat{\gamma})$. There is an alternative estimator, recently developed in Newey (1992b), that does not have these problems. It uses only the form of $g(z, \theta, \gamma)$ and the kernel to calculate the estimator. For a scalar ζ the estimator is given by

$$\hat{\delta}(z_i) = \nabla_{\zeta}\left[n^{-1}\sum_{j=1}^{n} g\{z_j, \hat{\theta}, \hat{\gamma} + \zeta y_i K_{\sigma}(\cdot - x_i)\}\right]\Big|_{\zeta=0}. \tag{8.18}$$

This estimator can be thought of as the influence of the ith observation through the kernel estimator. It can be calculated by either analytical or numerical differentiation. Consistency of the corresponding asymptotic variance estimator is shown in Newey (1992b).

It is helpful to consider some examples to illustrate how these results for first-step kernel estimates can be used.

Nonparametric consumer surplus continued: To show asymptotic normality, one can first check the conditions of Theorem 8.11. This estimator has $g(z, \gamma_0) = \int_a^b h_0(x)\,dx - \theta_0 = 0$, so the first two conditions are automatically satisfied. Let $\mathcal{X} = [a, b]$, which is a compact set, and suppose that Assumptions 8.1–8.3 are satisfied with $m = 2$, $d = 0$, and $p = 4$, so that the norm $\|\gamma\|$ is just a supremum norm, involving no derivatives. Note that $m = 2$ only requires that $\int uK(u)\,du = 0$, which is satisfied by many kernels. This condition also requires that $f_0(x)$ and $f_0(x)E[q|x]$ have versions that are twice continuously differentiable on an open set containing $[a, b]$, and that q have a fourth moment. Suppose that $n\sigma^2/(\ln n)^2 \to \infty$ and $n\sigma^4 \to 0$, giving the bandwidth conditions of Theorem 8.11, with $r = 1$ (here x is a scalar) and $d = 0$. Suppose that $f_0(x)$ is bounded away from zero on $[a, b]$. Then, as previously shown in eq. (8.9), assumption (i) is satisfied, with $b(z)$ equal to a constant and $G(z, \gamma) = \int_a^b f_0(x)^{-1}[-h_0(x), 1]\gamma(x)\,dx$. Assumption (ii) holds by inspection by $f_0(x)^{-1}$ and $h_0(x)$ bounded. As previously noted, assumption (iii) holds with $v(x) = 1(a \leqslant x \leqslant b) \times f_0(x)^{-1}[-h_0(x), 1]$. This function is continuous except at the points $x = a$ and $x = b$,

and is bounded, so that assumption (iv) is satisfied. Then by the conclusion of Theorem 8.11 it follows that

$$\sqrt{n}\left(\int_a^b \hat{h}(x) - \theta_0\right) \xrightarrow{d} N(0, E[1(a \leqslant x \leqslant b) f_0(x)^{-2}\{q - h_0(x)\}^2]), \tag{8.19}$$

an asymptotic normality result for a nonparametric consumer surplus estimator.

To estimate the asymptotic variance, note that in this example, $\delta(z) = 1(a \leqslant x \leqslant b) \times f_0(x)^{-1}[q - h_0(x)] = \delta(z, \gamma_0)$ for $\delta(z, \gamma) = 1(a \leqslant x \leqslant b)\gamma_1(x)^{-1}[q - \gamma_1(x)^{-1}\gamma_2(x)]$. Then for $\hat{\delta}(z) = \delta(z, \hat{\gamma})$, an asymptotic variance estimator will be

$$\hat{\Omega} = \sum_{i=1}^{n} \hat{\delta}(z_i)^2/n = n^{-1} \sum_{i=1}^{n} 1(a \leqslant x_i \leqslant b)\hat{f}(x_i)^{-2}[q_i - \hat{h}(x_i)]^2. \tag{8.20}$$

By the density bounded away from zero on $\mathcal{X} = [a, b]$, for $\|\gamma - \gamma_0\|$ small enough that $\gamma_1(x)$ is also bounded away from zero on $\mathcal{X}$, $|\delta(z_i, \gamma) - \delta(z_i, \gamma_0)| \leqslant C(1 + q_i)\|\gamma - \gamma_0\|$ for some constant C, so that the conditions of Theorem 8.13 are satisfied, implying consistency of $\hat{\Omega}$.

Weighted average derivative estimation: There are many examples of models where there is a dependent variable with $E[q|x] = \tau(x'\beta_0)$ for a parameter vector β_0, as discussed in Powell's chapter of this handbook. When the conditional expectation satisfies this "index" restriction, then $\nabla_x E[q|x] = \tau_v(x'\beta_0)\beta_0$, where $\tau_v(v) = d\tau(v)/dv$. Consequently, for any bounded function $w(x)$, $E[w(x)\nabla_x E[q|x]] = E[w(x)\tau_v(x'\beta_0)]\beta_0$, i.e. the weighted average derivative $E[w(x)\nabla_x E[q|x]]$ is equal to a scale multiple of the coefficients β_0. Consequently, an estimate of β_0 that is consistent up to scale can be formed as

$$\hat{\theta} = n^{-1} \sum_{i=1}^{n} w(x_i)\nabla_x \hat{h}(x_i), \quad \hat{h}(x) = \sum_{i=1}^{n} q_i K_\sigma(x - x_i) / \sum_{i=1}^{n} K_\sigma(x - x_i). \tag{8.21}$$

This is a weighted average derivative estimator.

This estimator takes the form given above where $\gamma_{10}(x) = f_0(x)$, $\gamma_{20}(x) = f_0(x) \times E[q|x]$, and

$$g(z, \theta, \gamma) = w(x)\nabla_x[\gamma_2(x)/\gamma_1(x)] - \theta. \tag{8.22}$$

The weight $w(x)$ is useful as a "fixed trimming" device, that will allow the application of Theorem 8.11 even though there is a denominator term in $g(z, \theta, \gamma)$. For this purpose, let $\mathcal{X}$ be a compact set, and suppose that $w(x)$ is zero outside $\mathcal{X}$ and bounded. Also impose the condition that $f_0(x) = \gamma_{10}(x)$ is bounded away from zero on $\mathcal{X}$. Suppose that Assumptions 8.1–8.3 are satisfied, $n\sigma^{2r+4}/(\ln n)^2 \to \infty$ and $n\sigma^{2m} \to 0$.

These conditions will require that $m > r + 2$, so that the kernel must be of the higher-order type, and $\gamma_0(x)$ must be differentiable of higher order than the dimension of the regressors plus 2. Then it is straightforward to verify that assumption (i) of Theorem 8.11 is satisfied where the norm $\|\gamma\|$ includes the first derivative, i.e. where $d = 1$, with a linear term given by

$$G(z,\gamma) = w(x)[a_0(x)\gamma(x) + \nabla_x\gamma(x)'b_0(x)],$$
$$a_0(x) = f_0(x)^{-1}[-h_{0x}(x) + h_0(x)s(x), -s(x)], \qquad b_0(x) = f_0(x)^{-1}[-h_0(x), 1]', \tag{8.23}$$

where an x subscript denotes a vector of partial derivatives, and $s(x) = f_{0x}(x)/f_0(x)$ is the score for the density of x. This result follows from expanding the ratio $\nabla_x[\gamma_2(x)/\gamma_1(x)]$ at each given point for x, using arguments similar to those in the previous example. Assumption (ii) also holds by inspection, by $f_0(x)$ bounded away from zero.

To obtain assumption (iii) in this example, an additional step is required. In particular, the derivatives $\nabla_x\gamma(x)$ have to be transformed to the function values $\gamma(x)$ in order to obtain the representation in assumption (iii). The way this is done is by integration by parts, as in

$$E[w(x)\nabla_x\gamma(x)'b_0(x)] = \int w(x)f_0(x)b_0(x)'[\nabla_x\gamma(x)]\,dx$$
$$= -\int \nabla_x[w(x)f_0(x)b_0(x)]'\gamma(x)\,dx,$$
$$\nabla_x[w(x)f_0(x)b_0(x)]' = w_x(x)[-h_0(x), 1] + w(x)[-h_{0x}(x), 0]$$

It then follows that $\int G(z,\gamma)\,dF_0 = \int v(x)\gamma(x)\,dx$, for

$$v(x) = -w_x(x)[-h_0(x), 1] - w(x)[-h_{0x}(x), 0] + w(x)a_0(x)$$
$$= -\{w_x(x) + w(x)s(x)\}[-h_0(x), 1] = \ell(x)[-h_0(x), 1],$$
$$\ell(x) = -w_x(x) - w(x)s(x). \tag{8.24}$$

By the assumption that $f_0(x)$ is bounded away from zero on $\mathscr{X}$ and that $\mathscr{X}$ is compact, the function $\ell(x)[-h_0(x), 1]$ is bounded, continuous, and zero outside a compact set, so that condition (iv) of Theorem 8.11 is satisfied. Noting that $\delta(z) = \ell(x)[q - h_0(x)]$, the conclusion of Theorem 8.11 then gives

$$\sqrt{n}\left[n^{-1}\sum_{i=1}^{n} w(x_i)\nabla_x\hat{g}(x_i) - \theta_0\right] \xrightarrow{d} N(0, \mathrm{Var}\{w(x)\nabla_x h_0(x) + \ell(x)[q - h_0(x)]\}). \tag{8.25}$$

The asymptotic variance of this estimator can be estimated as

$$\hat{\Omega} = n^{-1} \sum_{i=1}^{n} \hat{u}_i \hat{u}_i', \qquad \hat{u}_i = w(x_i)\nabla_x \hat{h}(x_i) - \hat{\theta} + \hat{\ell}(x_i)[q_i - \hat{h}(x_i)], \tag{8.26}$$

where $\hat{\ell}(x) = -w_x(x) - w(x)\hat{f}_x(x)/\hat{f}(x)$ for $\hat{f}(x) = n^{-1}\sum_{i=1}^{n} K(x - x_i)$. Consistency of this asymptotic variance estimator will follow analogously to the consumer surplus example.

One cautionary note due to Stoker (1991) is that the kernel weighted average derivative estimators tend to have large small sample biases. Stoker (1991) suggests a corrected estimate of $-[n^{-1}\sum_{i=1}^{n} \hat{\ell}(x_i)x_i']^{-1}\hat{\theta}$, and shows that this correction tends to reduce bias $\hat{\theta}$ and does not affect the asymptotic variance. Newey et al. (1992) suggest an alternative estimator $\hat{\theta} + n^{-1}\sum_{i=1}^{n} \hat{\ell}(x_i)[q_i - \hat{h}(x_i)]$, and show that this also tends to have smaller bias than $\hat{\theta}$. Newey et al. (1992) also show how to extend this correction to any two-step semiparametric estimator with a first-step kernel.

8.4. *Technicalities*

Proof of Lemma 8.4

Let $m_{ij} = m(z_i, z_j)$, $\bar{m}_{i\cdot} = m_1(z_i)$, and $\bar{m}_{\cdot i} = m_2(z_i)$. Note that $E[\|m_{11} - \mu\|] \leq E[\|m_{11}\|] + (E[\|m_{12}\|^2])^{1/2}$ and $(E[\|m_{12} - \mu\|^2])^{1/2} \leq 2(E[\|m_{12}\|^2])^{1/2}$ by the triangle inequality. Thus, by replacing $m(z_1, z_2)$ with $m(z_1, z_2) - \mu$ it can be assumed that $\mu = 0$. Note that $\|\sum_{ij} m_{ij}/n^2 - \sum_i(\bar{m}_{i\cdot} + \bar{m}_{\cdot i})/n\| = \|\sum_{ij}(m_{ij} - \bar{m}_{i\cdot} - \bar{m}_{\cdot j})/n^2\| \leq \|\sum_{i \neq j}(m_{ij} - \bar{m}_{i\cdot} - \bar{m}_{\cdot j})/n^2\| + \|\sum_i (m_{ii} - \bar{m}_{i\cdot} - \bar{m}_{\cdot i})/n^2\| \equiv T_1 + T_2$. Note $E[T_2] \leq (E[\|m_{11}\|] + 2 \times E[\|m_{12}\|])/n$. Also, for $i \neq j$, $k \neq \ell$ let $v_{ijk\ell} \equiv E[(m_{ij} - \bar{m}_{i\cdot} - \bar{m}_{\cdot j})'(m_{k\ell} - \bar{m}_{k\cdot} - \bar{m}_{\cdot \ell})]$. By i.i.d. observations, if neither k nor ℓ is equal to i or j, then $v_{ijk\ell} = 0$. Also for ℓ not equal to i or j, $v_{iji\ell} = E[(m_{ij} - \bar{m}_{i\cdot})'(m_{i\ell} - \bar{m}_{i\cdot})] = E[E[(m_{ij} - \bar{m}_{i\cdot})'(m_{i\ell} - \bar{m}_{i\cdot})|z_i, z_j]] = E[(m_{ij} - \bar{m}_{i\cdot})'(E[m_{i\ell}|z_i, z_j] - \bar{m}_{i\cdot})] = 0 = v_{ijj\ell}$. Similarly, $v_{ijk\ell} = 0$ if k equals neither i nor j. Thus,

$$\begin{aligned} E[T_1^2] &= \sum_{i \neq j} \sum_{k \neq \ell} v_{ijk\ell}/n^4 = \sum_{i \neq j} (v_{iji} + v_{ijji})/n^4 \\ &= 2(n^2 - n)E[\|m_{12} - \bar{m}_{1\cdot} - \bar{m}_{\cdot 2}\|^2]/n^4 = E[\|m_{12} - \bar{m}_{1\cdot} - \bar{m}_{\cdot 2}\|^2]O(n^{-2}), \end{aligned}$$

and $T_1 = O_p(\{E[\|m_{12} - \bar{m}_{1\cdot} - \bar{m}_{\cdot 2}\|^2]\}^{1/2} n^{-1}) = O_p(\{E[\|m_{12}\|^2]\}^{1/2} n^{-1})$. The conclusion then follows by the triangle inequality. Q.E.D.

Proof of Lemma 8.5

Continuity of $a(z_1, z_2, \theta)$ follows by the dominated convergence theorem. Without changing notation let $a(z_1, z_2, \theta) = a(z_1, z_2, \theta) - E[a(z_1, z_2, \theta)]$. This function satisfies the same dominance conditions as $a(z_1, z_2, \theta)$, so it henceforth suffices to assume that

$E[a(z_1, z_2, \theta)] = 0$ for all θ. Let $\hat{U}(\theta) = n^{-1}(n-1)^{-1}\sum_{i \neq j} a(z_i, z_j, \theta)$, and note that $\sup_{\theta \in \Theta} \| n^{-2}\sum_{i,j} a(z_i, z_j, \theta) - \hat{U}(\theta) \| \xrightarrow{p} 0$. Then by well known results on U-statistics as in Serfling (1980), for each θ, $\hat{U}(\theta) \xrightarrow{p} 0$. It therefore suffices to show stochastic equicontinuity of θ. The rest of the proof proceeds as in the proof of Lemma 2.4, with $\hat{\Delta}_{ij}(\theta, \delta) = \sup_{\|\tilde{\theta} - \theta\| \leq \delta} \| a(z_i, z_j, \tilde{\theta}) - a(z_i, z_j, \theta) \|$ replacing $\hat{\Delta}_i(\theta, \delta)$, $\sum_{i \neq j}$ replacing $\sum_{i=1}^n$, and the U-statistic convergence result $n^{-1}(n-1)^{-1}\sum_{i \neq j} \hat{\Delta}_{ij}(\theta, \delta) \xrightarrow{p} E[\hat{\Delta}_{12}(\theta, \delta)]$ replacing the law of large numbers. Q.E.D

Proof of Lemma 8.7

Let $\hat{m}_{ij} = m_n(z_i, z_j, \hat{\theta})$, $m_{ij} = m_n(z_i, z_j, \theta_0)$, and $m_{1i} = m_{n1}(z_i)$. By the triangle inequality, we have

$$n^{-1}\sum_{i=1}^n \| n^{-1}\sum_{j=1}^n \hat{m}_{ij} - m_{1i} \|^2 \leq Cn^{-2}\sum_{i=1}^n \| \hat{m}_{ii} \|^2$$

$$+ Cn^{-1}\sum_{i=1}^n \| n^{-1}\sum_{j \neq i} (\hat{m}_{ij} - m_{ij}) \|^2 + Cn^{-1}\sum_{i=1}^n \| (n-1)^{-1}\sum_{j \neq i} (m_{ij} - m_{1i}) \|^2$$

$$+ Cn^{-2}\sum_{i=1}^n \| m_{1i} \|^2 = R_1 + R_2 + R_3 + R_4.$$

for some positive constant C. Let $b(z_i) = \sup_{\theta \in \mathcal{N}} \| m_n(z_i, z_i, \theta) \|$ and $b(z_i, z_j) = \sup_{\theta \in \mathcal{N}} \| \nabla_\theta m_n(z_i, z_j, \theta) \|$. With probability approaching one, $R_1 \leq Cn^{-2}\sum_{i=1}^n b(z_i)^2 = O_p\{n^{-1}E[b(z_1)^2]\}$. Also, $R_2 \leq Cn^{-1}\sum_{i=1}^n \| n^{-1}\sum_{j \neq i} b(z_i, z_j) \|^2 \| \hat{\theta} - \theta_0 \|^2 \leq Cn^{-2} \times \sum_{i \neq j} b(z_i, z_j)^2 \| \hat{\theta} - \theta_0 \|^2 = O_p\{n^{-1}E[b(z_1, z_2)^2]\}$. Also, by the Chebyshev and Cauchy–Schwartz inequalities, $E[R_3] \leq CE[\| m_{12} \|^2]/n$ and $E[R_4] \leq CE[\| m_{12} \|^2]/n$. The conclusion then follows by the Markov and triangle inequalities. Q.E.D.

9. Hypothesis testing with GMM estimators

This section outlines the large sample theory of hypothesis testing for GMM estimators. The trinity of Wald, Lagrange multiplier, and likelihood ratio test statistics from maximum likelihood estimation extend virtually unchanged to this more general setting. Our treatment provides a unified framework that specializes to both classical maximum likelihood methods and traditional linear models estimated on the basis of orthogonality restrictions.

Suppose data z are generated by a process that is parametrized by a $k \times 1$ vector θ. Let $\ell(z, \theta)$ denote the log-likelihood of z, and let θ_0 denote the true value of θ in the population. Suppose there is an $m \times 1$ vector of functions of z and θ, denoted $g(z, \theta)$, that have zero expectation in the population if and only if θ equals θ_0:

$$g(\theta) \underset{m \times 1}{\equiv} E\underset{m \times 1}{g}(\underset{k \times 1}{z, \theta}) \equiv \int g(z, \theta) e^{\ell(z, \theta_0)} dz = 0, \quad \text{if } \theta = \theta_0.$$

Then, $Eg(z, \theta)$ are *moments*, and the analogy principle suggests that an estimator of θ_0 can be obtained by solving for θ that makes the sample analogs of the population moments small. Identification normally requires that $m \geqslant k$. If the inequality is strict, and the moments are not degenerate, then there are *overidentifying* moments that can be used to improve estimation efficiency and/or test the internal consistency of the model.

In this set-up, there are several alternative interpretations of z. It may be the case that z is a complete description of the data and $\ell(z, \theta)$ is the "full information" likelihood. Alternatively, some components of observations may be margined out, and $\ell(z, \theta)$ may be a marginal "limited information" likelihood. Examples are the likelihood for one equation in a simultaneous equations system, or the likelihood for continuous observations that are classified into discrete categories. Also, there may be "exogenous" variables (covariates), and the full or limited information likelihood above may be written conditioning on the values of these covariates. From the standpoint of statistical analysis, variables that are conditioned out behave like constants. Then, it does not matter for the discussion of hypothesis testing that follows which interpretation above applies, except that when regularity conditions are stated it should be understood that they hold almost surely with respect to the distribution of covariates.

Several special cases of this general set-up occur frequently in applications. First, if $\ell(z, \theta)$ is a full or limited information likelihood function, and $g(z, \theta) = \nabla_\theta \ell(z, \theta)$ is the score vector, then we obtain maximum likelihood estimation.[49] Second, if $z = (y, x, w)$ and $g(z, \theta) = w'(y - x\theta)$ asserts orthogonality in the population between *instruments* w and regression *disturbances* $\varepsilon = y - x\theta_0$, then GMM specializes to 2SLS, or in the case that $w = x$, to OLS. These linear regression set-ups generalize immediately to nonlinear regression orthogonality conditions based on the form $g(z, \theta) = w'[y - h(x, \theta)]$.

Suppose an i.i.d. sample $z_1, \ldots, z_n$ is obtained from the data generation process. A GMM estimator of θ_0 is the vector $\hat{\theta}_n$ that minimizes the generalized distance of the sample moments from zero, where this generalized distance is defined by the quadratic form

$$-Q_n(\theta) = \tfrac{1}{2}\hat{g}_n(\theta)' \Omega_n^{-1} \hat{g}_n(\theta),$$

with $\hat{g}_n(\theta) \equiv (1/n)\sum_{t=1}^n g(z_t, \theta)$ and Ω_n an $m \times m$ positive definite symmetric matrix that defines a "distance metric". Define the covariance matrix of the moments, $\underset{m \times m}{\Omega} \equiv Eg(z, \theta_0)g(z, \theta_0)'$. Efficient weighting of a given set of m moments requires that Ω_n converge to Ω as $n \to \infty$.[50] Also, define the Jacobian matrix $\underset{m \times k}{G} \equiv E\nabla_\theta g(z, \theta_0)$, and

[49] If the sample score has multiple roots, we assume that a root is selected that achieves a global maximum of the likelihood function.

[50] This weighting is efficient in that it minimizes the asymptotic covariance matrix in the class of all estimators obtained by setting to zero k linear combinations of the m moment conditions. Obviously, if there are exactly k moments, then the weighting is irrelevant. It is often useful to obtain initial consistent asymptotically normal GMM estimators employing an inefficient weighting that reduces computation, and then apply the one-step theorem to get efficient estimators.

let G_n denote an array that approaches G as $n \to \infty$. The arrays Ω_n and G_n may be functions of (preliminary) estimates $\tilde{\theta}_n$ of θ_0. When it is necessary to make this dependence explicit, write $\Omega_n(\tilde{\theta}_n)$ and $G_n(\tilde{\theta}_n)$.

Theorems 2.6, 3.4, and 4.5 for consistency, asymptotic normality, and asymptotic covariance matrix estimation, guarantee that the unconstrained GMM estimator $\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} Q_n(\theta)$ is consistent and asymptotically normal, with $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, B^{-1})$; where $B \equiv G'\Omega^{-1}G$. Further, from Theorem 4.5, the asymptotic covariance matrix can be estimated using

$$G_n = \frac{1}{n}\sum_{t=1}^{n} \nabla_\theta g(z_t, \tilde{\theta}_n) \xrightarrow{p} G,$$

$$\Omega_n = \frac{1}{n}\sum_{t=1}^{n} g(z_t, \tilde{\theta}_n) g(z_t, \tilde{\theta}_n)' \xrightarrow{p} \Omega,$$

where $\tilde{\theta}_n$ is any $\sqrt{n}$-consistent estimator of θ_0 [i.e., $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is stochastically bounded]. A practical procedure for estimation is to first estimate θ using the GMM criterion with an arbitrary Ω_n, such as $\Omega_n = I$. This produces an initial $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$. Then use the formulae above to estimate the asymptotically efficient Ω_n, and use the GMM criterion with this distance metric to obtain the final estimator $\hat{\theta}_n$. Equation (5.1) establishes that $\Gamma \equiv -Eg(z, \theta_0)\nabla_\theta \ell(z, \theta_0)' \equiv E\nabla_\theta g(z, \theta_0) \equiv G$. It will sometimes be convenient to estimate G by

$$\Gamma_n = -\frac{1}{n}\sum_{t=1}^{n} g(z_t, \tilde{\theta}_t)\nabla_\theta \ell(z_t, \tilde{\theta}_n)'.$$

In the maximum likelihood case $g = \nabla_\theta \ell$, one has $\Omega = \Gamma = G$, and the asymptotic covariance matrix of the unconstrained estimator simplifies to Ω^{-1}.

9.1. *The null hypothesis and the constrained GMM estimator*

Suppose there is an r-dimensional null hypothesis on the data generation process,

$$H_0: \underset{r \times 1}{a}(\theta_0) = 0.$$

We will consider alternatives to the null of the form

$$H_1: a(\theta_0) \neq 0,$$

or *asymptotically local* alternatives of the form

$$H_{1n}: a(\theta_0) = \delta/\sqrt{n} \neq 0.$$

Assume that $\underset{r\times k}{A} \equiv \nabla_\theta a(\theta_0)$ has rank r. The null hypothesis may be linear or nonlinear. A particularly simple case is $H_0: \theta = \theta^0$, or $a(\theta) \equiv \theta - \theta^0$, so the parameter vector θ is completely specified under the null. More generally, there will be $k - r$ parameters to be estimated when one imposes the null. One can define a *constrained* GMM estimator by optimizing the GMM criterion subject to the null hypothesis:

$$\bar{\theta}_n = \operatorname{argmax}_{\theta\in\Theta} Q_n(\theta), \quad \text{subject to } a(\theta) = 0.$$

Define a Lagrangian for $\bar{\theta}_n$: $\mathscr{L}_n(\theta, \gamma) = Q_n(\theta) - \underset{1\times r}{a}(\theta)' \underset{r\times 1}{\gamma}$. In this expression, γ is the vector of undetermined Lagrangian multipliers; these will be nonzero when the constraints are binding. The first-order conditions for solution of this problem are

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \sqrt{n}\nabla_\theta Q_n(\bar{\theta}_n) - \nabla_\theta a(\bar{\theta}_n)'\sqrt{n}\bar{\gamma}_n \\ -a(\bar{\theta}_n) \end{bmatrix}.$$

A first result establishes that $\bar{\theta}_n$ is consistent under the null or local alternatives:

Theorem 9.1

Suppose the hypotheses of Theorem 2.6. Suppose $a(\theta_0) = \delta/\sqrt{n}$, including the null when $\delta = 0$, with a continuously differentiable and A of rank r. Then $\bar{\theta} \xrightarrow{p} \theta_0$.

Proof

Let θ_{0n} minimize $[E\hat{g}_n(\theta)]'\Omega^{-1}[E\hat{g}_n(\theta)]$ subject to $a(\theta) = \delta/\sqrt{n}$. Continuity of this objective function and the uniqueness of its minimum imply $\theta_{0n} \to \theta_0$. Then $Q_n(\bar{\theta}_n) \leqslant Q_n(\theta_{0n}) \xrightarrow{p} 0$, implying $Q_n(\bar{\theta}_n) \xrightarrow{p} 0$. But Q_n converges uniformly to $[E\hat{g}_n(\theta)]'\Omega^{-1} \times [E\hat{g}_n(\theta)]$, so the argument of Theorem 2.6 implies $\bar{\theta}_n \xrightarrow{p} \theta_0$. Q.E.D.

The consistency of $\bar{\theta}_n$ implies

$$\nabla_\theta Q_n(\bar{\theta}_n) \xrightarrow{p} -G'\Omega^{-1}Eg(z, \theta_0) = 0,$$

$$\nabla_\theta a(\bar{\theta}_n) \xrightarrow{p} A \Rightarrow A'\bar{\gamma}_n = -\nabla_\theta Q_n(\bar{\theta}_n) + o_p \xrightarrow{p} 0,$$

and since A is of full rank, $\bar{\gamma}_n \xrightarrow{p} 0$. A central limit theorem implies

$$-\Omega^{-1/2}\sqrt{n}\hat{g}_n(\theta_0) \equiv \underset{m\times 1}{\mathscr{U}_n} \xrightarrow{d} \underset{m\times 1}{\mathscr{U}} \sim N(0, I). \tag{9.1}$$

A Taylor's expansion of the sample moments about θ_0 gives

$$\sqrt{n}\hat{g}_n(\theta) = \sqrt{n}\hat{g}_n(\theta_0) + G_n\sqrt{n}(\theta - \theta_0), \tag{9.2}$$

with G_n evaluated at points between θ and θ_0. Substituting this expression for the final term in the unconstrained first-order condition $0 = \sqrt{n}\nabla_\theta Q_n(\hat{\theta}_n) \equiv -G_n'\Omega_n^{-1} \times \hat{g}_n(\hat{\theta}_n)$ and using the consistency of $\hat{\theta}_n$ and uniform convergence of $G_n(\theta)$ yields

$$\begin{aligned} 0 &= -G'\Omega^{-1/2}\mathscr{U}_n + B\sqrt{n}(\hat{\theta}_n - \theta_0) + o_p \\ \Rightarrow \sqrt{n}(\hat{\theta}_n - \theta_0) &= B^{-1}G'\Omega^{-1/2}\mathscr{U}_n + o_p. \end{aligned} \tag{9.3}$$

Similarly, substituting $\sqrt{n}\hat{g}_n(\bar{\theta}_n) = \sqrt{n}\hat{g}_n(\theta_0) + G_n\sqrt{n}(\bar{\theta}_n - \theta_0) = -G'\Omega^{-1/2}\mathscr{U}_n + G\sqrt{n}(\bar{\theta}_n - \theta_0) + o_p$, and $\sqrt{n}a(\bar{\theta}_n) = \sqrt{n}a(\theta_0) + A\sqrt{n}(\bar{\theta}_n - \theta_0) + o_p \equiv \delta + A\sqrt{n}(\bar{\theta}_n - \theta_0) + o_p$ in the first-order conditions for $\bar{\theta}_n$ yields

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} G'\Omega^{-1/2}\mathscr{U}_n \\ -\delta \end{bmatrix} - \begin{bmatrix} B & A' \\ A & 0 \end{bmatrix}\begin{bmatrix} \sqrt{n}(\bar{\theta}_n - \theta_0) \\ \sqrt{n}\bar{\gamma}_n \end{bmatrix} + o_p. \tag{9.4}$$

From the formula for partitioned inverses,

$$\begin{bmatrix} B & A' \\ A & 0 \end{bmatrix}^{-1} = \begin{bmatrix} B^{-1/2}MB^{-1/2} & B^{-1}A'(AB^{-1}A')^{-1} \\ (AB^{-1}A')^{-1}AB^{-1} & -(AB^{-1}A')^{-1} \end{bmatrix}, \tag{9.5}$$

where $M = I - B^{-1/2}A'(AB^{-1}A')^{-1}AB^{-1/2}$ is a $k \times k$ idempotent matrix of rank $k - r$. Applying this to eq. (9.4) yields

$$\begin{bmatrix} \sqrt{n}(\bar{\theta}_n - \theta_0) \\ \sqrt{n}\bar{\gamma}_n \end{bmatrix} = \begin{bmatrix} -B^{-1}A'(AB^{-1}A')^{-1} \\ (AB^{-1}A')^{-1} \end{bmatrix}\delta + \begin{bmatrix} B^{-1/2}MB^{-1/2} \\ (AB^{-1}A')^{-1}AB^{-1} \end{bmatrix}G'\Omega^{-1/2}\mathscr{U}_n + o_p. \tag{9.6}$$

Then, the asymptotic distribution of $\sqrt{n}(\bar{\theta}_n - \theta_0)$ under a local alternative, or the null with $\delta = 0$, is $N[-B^{-1}A'(AB^{-1}A')^{-1}\delta, B^{-1/2}MB^{-1/2}]$.

Writing out $M = I - B^{-1/2}A'(AB^{-1}A')^{-1}AB^{-1/2}$ yields

$$\begin{aligned} \sqrt{n}(\bar{\theta}_n - \theta_0) = {} & B^{-1}G'\Omega^{-1/2}\mathscr{U}_n - B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}\mathscr{U}_n \\ & - B^{-1}A'(AB^{-1}A')^{-1}\delta + o_p. \end{aligned} \tag{9.7}$$

The first terms on the right-hand side of eq. (9.7) and the right-hand side of eq. (9.3) are identical, to order o_p. Then, they can be combined to conclude that

$$\sqrt{n}(\hat{\theta}_n - \bar{\theta}_n) = B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}\mathscr{U}_n + B^{-1}A'(AB^{-1}A')^{-1}\delta + o_p, \tag{9.8}$$

so that $\sqrt{n}(\hat{\theta}_n - \bar{\theta}_n)$ is asymptotically normal with mean $B^{-1}A'(AB^{-1}A')^{-1}\delta$ and

Table 1

Statistic	Formula	Asymptotic covariance matrix
$\sqrt{n}(\hat{\theta}_n - \theta_0)$	$B^{-1}G'\Omega^{-1/2}\mathscr{U}_n + o_p$	$B^{-1} \equiv C$
$\sqrt{n}(\bar{\theta}_n - \theta_0)$	$-B^{-1}A'(AB^{-1}A')^{-1}\delta + B^{-1/2}MB^{-1/2}G'\Omega^{-1/2}\mathscr{U}_n + o_p$	$B^{-1/2}MB^{-1/2}$
$\sqrt{n}(\hat{\theta}_n - \bar{\theta}_n)$	$B^{-1}A'(AB^{-1}A')^{-1}\delta + B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}\mathscr{U}_n + o_p$	$B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}$
$\sqrt{n}\bar{\gamma}_n$	$(AB^{-1}A')^{-1}\delta + (AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}\mathscr{U}_n + o_p$	$(AB^{-1}A')^{-1}$
$\sqrt{n}a(\hat{\theta}_n)$	$\delta + AB^{-1}G'\Omega^{-1/2}\mathscr{U}_n + o_p$	$AB^{-1}A'$
$\sqrt{n}\nabla_\theta Q_n(\bar{\theta}_n)$	$A'(AB^{-1}A')^{-1}\delta + A'(AB^{-1}A')^{-1}AB^{-1}G'\Omega^{-1/2}\mathscr{U}_n + o_p$	$A'(AB^{-1}A')^{-1}A$

covariance matrix $B^{-1/2}(I - M)B^{-1/2} \equiv B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}$. Note that the asymptotic covariance matrices satisfy $\operatorname{acov}(\hat{\theta}_n - \bar{\theta}_n) = \operatorname{acov}\hat{\theta}_n - \operatorname{acov}\bar{\theta}_n$, or *the variance of the difference equals the difference of the variances.* This proposition is familiar in a maximum likelihood context where the variance in the deviation between an efficient estimator and any other estimator equals the difference of the variances. We see here that it also applies to *relatively* efficient GMM estimators that use available moments and constraints optimally.

The results above and some of their implications are summarized in Table 1. Each statistic is distributed asymptotically as a linear transformation of a common standard normal random vector $\mathscr{U}$. Recall that $B = G'\Omega^{-1}G$ is a positive definite $k \times k$ matrix, and let $C = B^{-1} \equiv \operatorname{acov}\hat{\theta}_n$. Recall that $M = I - B^{-1/2}A'(AB^{-1}A')^{-1} \times AB^{-1/2}$ is a $k \times k$ idempotent matrix of rank $k - r$.

9.2. *The test statistics*

The test statistics for the null hypothesis fall into three major classes, sometimes called the *trinity*. *Wald statistics* are based on deviations of the unconstrained estimates from values consistent with the null. *Lagrange multiplier* (LM) or *score statistics* are based on deviations of the constrained estimates from values solving the unconstrained problem. *Distance metric statistics* are based on differences in the GMM criterion between the unconstrained and constrained estimators. In the case of maximum likelihood estimation, the distance metric statistic is asymptotically equivalent to the *likelihood ratio statistic*. There are several variants for Wald statistics in the case of the general nonlinear hypothesis; these reduce to the same expression in the simple case where the parameter vector is completely determined under the null. The same is true for the LM statistic. There are often significant computational advantages to using one member or variant of the trinity rather than another. On the other hand, they are all *asymptotically equivalent*. Thus, at least to first-order asymptotic approximation, there is no statistical reason to choose be-

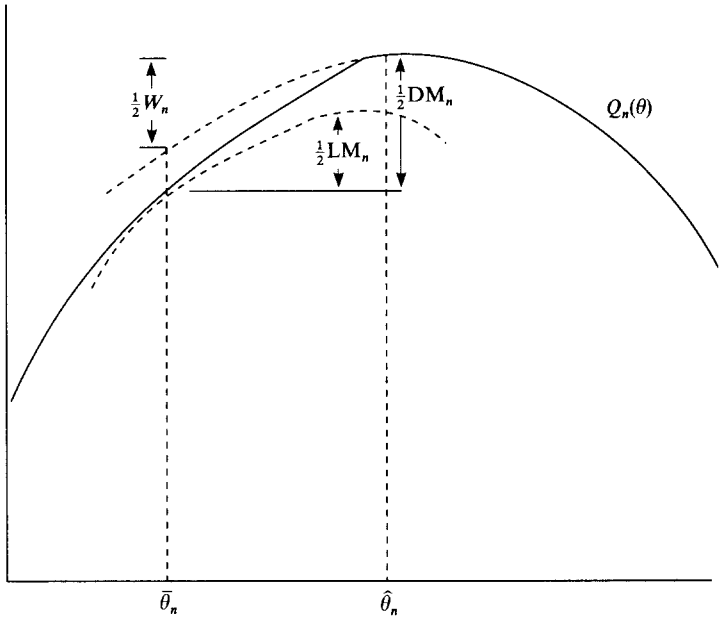

Figure 3. GMM tests.

tween them. This pattern of first-order asymptotic equivalence for GMM estimates is exactly the same as for maximum likelihood estimates.

Figure 3 illustrates the relationship between distance metric (DM), Wald (W), and score (LM) tests. In the case of maximum likelihood estimation, the distance metric criterion is replaced by the likelihood ratio.

The arguments $\hat{\theta}_n$ and $\bar{\theta}_n$ are the unconstrained GMM estimator and the GMM estimator subject to the null hypothesis, respectively. The GMM criterion function is plotted, along with quadratic approximations to this function through the respective arguments $\hat{\theta}_n$ and $\bar{\theta}_n$. The Wald statistic (W) can be interpreted as twice the difference in the criterion function at the two estimates, using a quadratic approximation to the criterion function at $\hat{\theta}_n$. The Lagrange multiplier (LM) statistic can be interpreted as twice the difference in the criterion function of the two estimates, using a quadratic approximation at $\bar{\theta}_n$. The distance metric (DM) statistic is twice the difference in the distance metric between the unconstrained and constrained estimators.

We develop the test statistics initially for the general nonlinear hypothesis $a(\theta_0) = 0$; the various statistics we consider are given in Table 2. In this table, recall that $\operatorname{acov} \hat{\theta}_n = B$ and $\operatorname{acov} \bar{\theta}_n = B^{-1/2} M B^{-1/2}$. In the following section, we consider the important special cases, including maximum likelihood and nonlinear least squares.

Table 2

	Test statistics
Wald statistics	
W_{1n}	$na(\hat{\theta}_n)'[AB^{-1}A']^{-1}a(\hat{\theta}_n)$
W_{2n}	$n(\hat{\theta}_n - \bar{\theta}_n)'\{\text{acov}(\hat{\theta}_n) - \text{acov}(\bar{\theta}_n)\}^{-}(\hat{\theta}_n - \bar{\theta}_n)$ $= n(\hat{\theta}_n - \bar{\theta}_n)'B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}(\hat{\theta}_n - \bar{\theta}_n)$
W_{3n}	$n(\hat{\theta}_n - \bar{\theta}_n)'\,\text{acov}(\hat{\theta}_n)^{-1}(\hat{\theta}_n - \bar{\theta}_n)$
Lagrange multiplier statistics	
LM_{1n}	$n\bar{\gamma}_n'AB^{-1}A'\bar{\gamma}_n$
LM_{2n}	$n\nabla_\theta Q_n(\bar{\theta}_n)'\{A'(AB^{-1}A')^{-1}A\}^{-}\nabla_\theta Q_n(\bar{\theta}_n)$ $= n\nabla_\theta Q_n(\bar{\theta}_n)'B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}\nabla_\theta Q_n(\bar{\theta}_n)$
LM_{3n}	$n\nabla_\theta Q_n(\bar{\theta}_n)'B^{-1}\nabla_\theta' Q_n(\bar{\theta}_n)$
Distance metric statistic	
DM_n	$-2n[Q_n(\bar{\theta}_n) - Q_n(\hat{\theta}_n)]$

In particular, when the hypothesis is that a subset of the parameters are constants, there are some simplifications of the statistics, and some versions are indistinguishable.

The following theorem gives the large sample distributions of these statistics:

Theorem 9.2

Suppose the conditions of Theorems 2.6, 3.4, and 4.5 are satisfied, and $a(\theta)$ is continuously differentiable with A of rank r. The test statistics in Table 2 are asymptotically equivalent under the null or under local alternatives. Under the null, the statistics converge in distribution to a chi-square with r degrees of freedom. Under a local alternative $a(\theta_0) = \delta/\sqrt{n}$, the statistics converge in distribution to a noncentral chi-square with r degrees of freedom and a noncentrality parameter $\delta'(AB^{-1}A')^{-1}\delta$.

Proof

All of the test statistics are constructed from the expressions in Table 1. If q is an expression from the table with asymptotic covariance matrix $R = \text{acov}\, q$ and asymptotic mean $R\lambda$ under local alternatives to the null, then the statistic will be of the form $q'R^+q$, where R^+ is any symmetric matrix that satisfies $RR^+R = R$. The matrix R^+ will be the ordinary inverse R^{-1} if R is nonsingular, and may be the Moore–Penrose generalized inverse R^- if R is singular. Section 9.8 defines generalized inverses, and Lemma 9.7 in that section shows that if q is a normal random vector with covariance matrix R of rank r and mean $R\lambda$, then $q'R^+q$ is distributed noncentral chi-square with r degrees of freedom and noncentrality parameter $\lambda'R\lambda$ under local alternatives to the null.

Consider W_{1n}. Under the local alternative $a(\theta_0) = \delta/\sqrt{n}$, row five of Table 1 gives $q = \delta + AB^{-1}G'\Omega^{-1/2}\mathscr{U}$ normal with mean δ and a nonsingular covariance matrix $R = AB^{-1}A'$. Let $\lambda = R^{-1}\delta$. Then Lemma 9.7 implies the result with noncentrality parameter $\lambda' R\lambda = \delta' R^{-1}\delta \equiv \delta'(AB^{-1}A')^{-1}\delta$.

Consider W_{2n}. The generalized inverse R^- of $R = \operatorname{acov}\hat{\theta}_n - \operatorname{acov}\bar{\theta}_n$ can be written as:

$$R^- \equiv \{B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}\}^- = B^{+1/2}\{B^{-1/2}A'(AB^{-1}A')^{-1}AB^{-1/2}\}^- B^{+1/2}$$
$$= B^{+1/2}\{B^{-1/2}A(AB^{-1}A)^{-1}AB^{-1/2}\}B^{+1/2} = A'(AB^{-1}A')^{-1}A.$$

The first identity substitutes the covariance formula from row 2 of Table 1. The second and third equalities follow from Section 9.8, Lemma 9.5, (5) and (4), respectively. One can check that $\lambda = R^-B^{-1}A'(AB^{-1}A')^{-1}\delta$ satisfies $R\lambda = B^{-1}A' \times (AB^{-1}A')^{-1}\delta$, so that $\lambda' R\lambda = \delta'(AB^{-1}A')^{-1}\delta$.

The statistic W_{3n} is obtained by noting that for $R = B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}$, the matrix $R^+ = B$ satisfies $RR^+R = R$ and $\lambda = R^+B^{-1}A'(AB^{-1}A')^{-1}\delta$ satisfies $R\lambda = B^{-1}A'(AB^{-1}A')^{-1}\delta$.

Similar arguments establish the properties of the LM statistics. In particular, the second form of the statistic LM_{2n} follows from previous argument that $A'(AB^{-1}A')^{-1}A'$ and $B^{-1}A'(AB^{-1}A')^{-1}AB^{-1}$ are generalized inverses, and the statistic LM_{3n} is obtained by noting that $R = A'(AB^{-1}A')^{-1}A$ has $RR^+R = R$ when $R^+ = B^{-1}$.

To demonstrate the asymptotic equivalence of DM_n to the earlier statistics, make a Taylor's expansion of the sample moments for $\bar{\theta}_n$ about $\hat{\theta}_n$, $\sqrt{n}\hat{g}_n(\bar{\theta}_n) = \sqrt{n}\hat{g}_n(\hat{\theta}_n) + G_n\sqrt{n}(\bar{\theta}_n - \hat{\theta}_n) + o_p$, and substitute this in the expression for DM_n to obtain

$$\mathrm{DM}_n = -2n\{Q_n(\bar{\theta}_n) - Q_n(\hat{\theta}_n)\}$$
$$= 2\sqrt{n}(\bar{\theta}_n - \hat{\theta}_n)'G_n'\Omega_n^{-1}\sqrt{n}\hat{g}_n(\hat{\theta}_n) + \sqrt{n}(\bar{\theta}_n - \hat{\theta}_n)'G_n'\Omega_n^{-1}G_n\sqrt{n}(\bar{\theta}_n - \hat{\theta}_n) + o_p$$
$$= n(\bar{\theta}_n - \hat{\theta}_n)'B(\bar{\theta}_n - \hat{\theta}_n) + o_p \equiv W_{3n} + o_p,$$

with the last equality holding since $G_n'\Omega_n^{-1}\sqrt{n}\hat{g}_n(\hat{\theta}_n) = 0$. Q.E.D.

The Wald statistic W_{1n} asks how close are the unconstrained estimators to satisfying the constraints; i.e., how close to zero is $a(\hat{\theta}_n)$? This variety of the test is particularly useful when the unconstrained estimator is available and the matrix A is easy to compute. For example, when the null is that a subvector of parameters equal constants, then A is a selection matrix that picks out the corresponding rows and columns of B^{-1}, and this test reduces to a quadratic form with the deviations of the estimators from their hypothesized values in the wings, and the inverse of their asymptotic covariance matrix in the center. In the special case H_0: $\theta = \theta^0$, one has $A = I$.

The Wald test W_{2n} is useful if both the unconstrained and constrained estimators are available. Its first version requires only the readily available asymptotic covariance matrices of the two estimators, but for $r < k$ requires calculation of a generalized inverse. Algorithms for this are available, but are often not as numerically stable as classical inversion algorithms because near-zero and exact-zero characteristic roots are treated very differently. The second version involves only ordinary inverses, and is potentially quite useful for computation in applications.

The Wald statistic W_{3n} treats the constrained estimators *as if they were constants with a zero asymptotic covariance matrix.* This statistic is particularly simple to compute when the unconstrained and constrained estimators are available, as no matrix differences or generalized inverses are involved, and the matrix A need not be computed. The statistic W_{2n} is in general larger than W_{3n} in finite samples, since the center of the second quadratic form is $(\operatorname{acov} \hat{\theta}_n)^{-1}$ and the center of the first quadratic form is $(\operatorname{acov} \hat{\theta}_n - \operatorname{acov} \bar{\theta}_n)^{-}$, while the tails are the same. Nevertheless, the two statistics are asymptotically equivalent.

The approach of Lagrange multiplier or score tests is to calculate the constrained estimator $\bar{\theta}_n$, and then to base a statistic on the discrepancy from zero at this argument of a condition that would be zero if the constraint were not binding. The statistic LM_{1n} asks how close the Lagrangian multipliers $\bar{\gamma}_n$, measuring the degree to which the hypothesized constraints are binding, are to zero. This statistic is easy to compute if the constrained estimation problem is actually solved by Lagrangian methods, and the multipliers are obtained as part of the calculation. The statistic LM_{2n} asks how close to zero is the gradient of the distance criterion, evaluated at the constrained estimator. This statistic is useful when the constrained estimator is available and it is easy to compute the gradient of the distance criterion, say using the algorithm to seek minimum distance estimates. The second version of the statistic avoids computation of a generalized inverse.

The statistic LM_{3n} bears the same relationship to LM_{2n} that W_{3n} bears to W_{2n}. This flavor of the test statistic is particularly convenient to calculate, as it can be obtained by auxiliary regressions starting from the constrained estimator $\bar{\theta}_n$:

Theorem 9.3

LM_{3n} can be calculated by a 2SLS regression:
(a) Regress $\nabla_\theta \ell(z_t, \bar{\theta}_n)'$ on $g(z_t, \bar{\theta}_n)$, and retrieve fitted values $\nabla_\theta \hat{\ell}(z_t, \bar{\theta}_n)'$.
(b) Regress 1 on $\nabla_\theta \hat{\ell}(z_t, \bar{\theta}_n)$, and retrieve fitted values $\hat{y}_t$. Then $\mathrm{LM}_{3n} = \sum_{t=1}^{n} \hat{y}_t^2$.
For MLE, $g = \nabla_\theta \ell$, and this procedure reduces to OLS.

Proof

Let y be an n-vector of 1's, X an $n \times k$ array whose rows are $\nabla_\theta \ell'$, Z an $n \times m$ array whose rows are g'. The first regression yields $\hat{X} = Z(Z'Z)^{-1}Z'X$, and the second regression yields $\hat{y} = \hat{X}(\hat{X}'\hat{X})^{-1}\hat{X}'y$. Then, $(1/n)Z'Z = \Omega_n$, $(1/n)Z'X = \Gamma_n$, $(1/n)Z'y =$

$\hat{g}_n(\bar{\theta}_n)$, and

$$\hat{y}'\hat{y} = y'\hat{X}(\hat{X}'\hat{X})^{-1}\hat{X}'y = y'Z(Z'Z)^{-1}Z'X[X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'y.$$

Note that $\nabla_\theta Q_n(\bar{\theta}_n) = -G_n'\Omega_n^{-1}\hat{g}_n(\bar{\theta}_n) = -\Gamma_n'\Omega_n^{-1}\hat{g}_n(\bar{\theta}_n)$. Substituting terms, $\hat{\gamma}'\hat{y} = \mathrm{LM}_{3n}$. Q.E.D.

Another form of the auxiliary regression for computing LM_{3n} arises in the case of nonlinear instrumental variable regression. Consider the model $y_t = h(x_t, \theta_0) + \varepsilon_t$ with $E(\varepsilon_t | w_t) = 0$ and $E(\varepsilon_t^2 | w_t) = \sigma^2$, where w_t is a vector of instruments. Define $z_t = (y_t, x_t, w_t)$ and $g(z_t, \theta) = w_t[y_t - h(x_t, \theta)]$. Then $Eg(z, \theta_0) = 0$ and $Eg(z, \theta_0)g(z, \theta_0)' = \sigma^2 Ew_t w_t'$. The GMM criterion $Q_n(\theta)$ for this model is

$$\frac{-1}{2\sigma^2}\left[\frac{1}{n}\sum_{t=1}^{n} w_t\{y_t - h(x_t, \theta)\}\right]'\left(\frac{1}{n}\sum_{t=1}^{n} w_t w_t'\right)^{-1}\left[\frac{1}{n}\sum_{t=1}^{n} w_t\{y_t - h(x_t, \theta)\}\right];$$

the scalar σ^2 does not affect the optimization of this function. Consider the hypothesis $a(\theta_0) = 0$, and let $\bar{\theta}_n$ be the GMM estimator obtained subject to this hypothesis. One can compute LM_{3n} by the following method:

(a) Regress $\nabla_\theta h(x_t, \bar{\theta}_n)$ on w_t, and retrieve the fitted values $\nabla_\theta \hat{h}_t$.

(b) Regress the residual $u_t = y_t - h(x_t, \bar{\theta}_n)$ on $\nabla_\theta \hat{h}_t$, and retrieve the fitted values $\hat{u}_t$.

Then $\mathrm{LM}_{3n} = n\sum_{t=1}^n \hat{u}_t^2 / \sum_{t=1}^n u_t^2 \equiv nR^2$, with R^2 the *uncentered* multiple correlation coefficient. Note that this is not in general the same as the standard R^2 produced by OLS, since the denominator of that definition is the sum of squared deviations of the dependent variable about its mean. When the dependent variable has mean zero (e.g. if the nonlinear regression has an additive intercept term), the centered and uncentered definitions coincide.

The approach of the distance metric test is based on the discrepancy between the value of the distance metric, evaluated at the constrained estimate, and the minimum attained by the unconstrained estimate. This estimator is particularly convenient when both the unconstrained and constrained estimators can be computed, and the estimation algorithm returns the goodness-of-fit statistics. In the case of linear or nonlinear least squares, this is the familiar test statistic based on the sum of squared residuals from the constrained and unconstrained regressions.

The tests based on GMM estimation with an optimal weight matrix can be extended to any extremum estimator. Consider such an estimator, satisfying eq. (1.1). Also, let $\bar{\theta}$ be a restricted estimator, maximizing $\hat{Q}_n(\theta)$ subject to $a(\theta) = 0$. Suppose that the equality $H = -\Sigma$ is satisfied, for the Hessian matrix H and the asymptotic variance Σ [of $\sqrt{n}\nabla_\theta \hat{Q}_n(\theta_0)$] from Theorem 3.1. This property is a generalization of the information matrix equality to any extremum estimator. For GMM estimation with optimal weight matrix, this equality is satisfied if the objective function is normalized by $\frac{1}{2}$, i.e. $\hat{Q}_n(\theta) = \frac{1}{2}\hat{g}_n(\theta)'\hat{\Omega}^{-1}\hat{g}_n(\theta)$. Let $\hat{\Sigma}$ denote an estimator

of Σ based on $\hat{\theta}$ and $\bar{\Sigma}$ an estimator based on $\bar{\theta}$. Consider the following test statistics:

$$W = na(\hat{\theta})'[\hat{A}\hat{\Sigma}^{-1}\hat{A}']^{-1}a(\hat{\theta}), \qquad \hat{A} = \nabla_\theta a(\hat{\theta}),$$
$$\text{LM} = n\nabla_\theta \hat{Q}_n(\bar{\theta})'\bar{\Sigma}^{-1}\nabla_\theta \hat{Q}_n(\bar{\theta}),$$
$$\text{DM} = 2n[\hat{Q}_n(\hat{\theta}) - \hat{Q}_n(\bar{\theta})].$$

The statistic W is analogous to the first Wald statistic in Table 2 and the statistic LM to the third LM statistic in Table 2. We could also give analogs of the other statistics in Table 2, but for brevity we leave these extensions to the reader. Under the conditions of Theorems 2.1, 3.1, and 4.1, $H = -\Sigma$ and the same conditions on $a(\theta)$ previously given, these three test statistics will all have an asymptotic chi-squared distribution, with degrees of freedom equal to the number of components of $a(\theta)$.

As we have discussed, optimal GMM estimation provides one example of these statistics. The MLE also provides an example, as does optimal CMD estimation. Nonlinear least squares also fits this framework, if homoskedasticity holds and the objective function is normalized in the right way. Suppose that $\text{Var}(y|x) = \sigma^2$, a constant. Consider the objective function $\hat{Q}_n(\theta) = (2\hat{\sigma}^2)^{-1}\sum_{i=1}^n [y_i - h(x_i, \theta)]^2$, where $\hat{\sigma}^2$ is an estimator of σ^2. Then it is straightforward to check that, because of the normalization of dividing by $2\hat{\sigma}^2$, the condition $H = -\Sigma$ is satisfied. In this example, the DM test statistic will have a familiar squared residual form.

There are many examples of estimators where $H = -\Sigma$ is not satisfied. In these cases, the Wald statistic can still be used, but $\hat{\Sigma}^{-1}$ must be replaced by a consistent estimator of the asymptotic variance of $\hat{\theta}$. There is another version of the LM statistic that will be asymptotically equivalent to the Wald statistic in this case, but for brevity we do not describe it here. Furthermore, the DM statistic will not have a chi-squared distribution. These results are further discussed for quasi-maximum likelihood estimation by White (1982a), and for the general extremum estimator case by Gourieroux et al. (1983).

9.3. *One-step versions of the trinity*

Calculation of Wald or Lagrange multiplier test statistics in finite samples requires estimation of G, Ω, and/or A. Any convenient consistent estimates of these arrays will do, and will preserve the asymptotic equivalence of the tests under the null and local alternatives. In particular, one can evaluate terms entering the definitions of these arrays at $\hat{\theta}_n$, $\bar{\theta}_n$, or any other consistent estimator of θ_0. In sample analogs that converge to these arrays by the law of large numbers, one can freely substitute sample and population terms that leave the probability limits unchanged. For example, if $z_t = (y_t, x_t)$ and $\tilde{\theta}_n$ is any consistent estimator of θ_0, then Ω can be estimated by (1) an analytic expression for $Eg(z, \theta)g(z, \theta)'$, evaluated at $\tilde{\theta}_n$, (2) a sample average $(1/n)\sum_{t=1}^n g(z_t, \tilde{\theta}_n)g(z_t, \tilde{\theta}_n)'$, or (3) a sample average of conditional

expectations $(1/n)\sum_{t=1}^{n} E_{y|x_t} g(y, x_t, \tilde{\theta}_n) g(y, x_t, \tilde{\theta}_n)'$. These first-order efficiency equivalences do *not* hold in finite samples, or even to higher orders of $\sqrt{n}$. Thus, there may be clear choices between these when higher orders of approximation are taken into account.

The next result is an application of the one-step theorem in Section 3.4, and shows how one can start from any initial $\sqrt{n}$-consistent estimator of θ_0, and in one iteration obtain versions of the trinity that are asymptotically equivalent to versions obtained when the exact estimators $\hat{\theta}_n$ and $\bar{\theta}_n$ are used. Further, the required iterations can usually be cast as regressions, so their computation is relatively elementary. Consider the GMM criterion $Q_n(\theta)$. Suppose $\tilde{\theta}_n$ is any consistent estimator of θ_0 such that $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is stochastically bounded. Let $\hat{\theta}_n$ be the unconstrained maximizer of Q, and $\bar{\theta}_n$ be the maximizer of Q subject to the constraint $a(\theta) = 0$. Suppose the null hypothesis, or a local alternative, $a(\theta_0) = \delta/\sqrt{n}$, is true. The unconstrained one-step estimator from eq. (3.11), $\hat{\tilde{\theta}}_n = \tilde{\theta}_n - (G_n' \Omega_n^{-1} G_n)^{-1} G_n' \times \Omega_n^{-1} \hat{g}_n(\tilde{\theta}_n)$, satisfies $\sqrt{n}(\hat{\tilde{\theta}}_n - \hat{\theta}_n) \xrightarrow{p} 0$. Similarly, define one-step constrained estimators from the Lagrangian first-order conditions:

$$\begin{bmatrix} \bar{\tilde{\theta}}_n \\ \bar{\tilde{\gamma}}_n \end{bmatrix} = \begin{bmatrix} \tilde{\theta}_n \\ 0 \end{bmatrix} - \begin{bmatrix} B & A' \\ A & 0 \end{bmatrix}^{-1} \begin{bmatrix} \nabla_\theta Q_n(\tilde{\theta}_n) \\ -a(\tilde{\theta}_n) \end{bmatrix}.$$

Note in this definition that $\gamma = 0$ is a trivial initially consistent estimator of the Lagrangian multipliers under the null or local alternatives, and that the arrays B and A can be estimated at $\tilde{\theta}_n$. The one-step theorem again applies, yielding $\sqrt{n}(\bar{\tilde{\theta}}_n - \bar{\theta}_n) \xrightarrow{p} 0$ and $\sqrt{n}(\bar{\tilde{\gamma}}_n - \bar{\gamma}_n) \xrightarrow{p} 0$. Then, these one-step equivalents can be substituted in any of the test statistics of the trinity without changing their asymptotic distribution.

A regression procedure for calculating the one-step expressions is often useful for computation. The adjustment from $\tilde{\theta}_n$ yielding the one-step unconstrained estimator is obtained by a two-stage least squares regression of the constant one on $\nabla_\theta \ell(z_t, \tilde{\theta}_n)$, with $g(z_t, \tilde{\theta}_n)$ as instruments; i.e.

(a) Regress each component of $\nabla_\theta \ell(z_t, \tilde{\theta}_n)$, on $g(z_t, \tilde{\theta}_n)$ in the sample $t = 1, \ldots, n$, and retrieve fitted values $\nabla_\theta \hat{\ell}(z_t, \tilde{\theta}_n)$.

(b) Regress 1 on $\nabla_\theta \hat{\ell}(z_t, \tilde{\theta}_n)$; and adjust $\tilde{\theta}_n$ by the amounts of the fitted coefficients.

Step (a) yields $\nabla_\theta \hat{\ell}(z_t, \tilde{\theta}_n)' = g(z_t, \tilde{\theta}_n) \Omega_n^{-1} \Gamma_n$, and step (b) yields coefficients

$$\Delta = \left[\sum_{t=1}^{n} [\nabla_\theta \hat{\ell}(z_t, \tilde{\theta}_n)][\nabla_\theta \hat{\ell}(z_t, \tilde{\theta}_n)]' \right]^{-1} \sum_{t=1}^{n} \nabla_\theta \hat{\ell}(z_t, \tilde{\theta}_n)$$

$$= (\Gamma_n' \Omega_n^{-1} \Gamma_n)^{-1} \Gamma_n' \Omega_n^{-1} \hat{g}_n(\tilde{\theta}_n).$$

This is the adjustment indicated by the one-step theorem.

Computation of one-step constrained estimators is conveniently done using the formulae

$$\begin{aligned}\bar{\tilde{\theta}}_n &= \hat{\tilde{\theta}}_n - B^{-1}A'(AB^{-1}A')^{-1}a(\hat{\tilde{\theta}}_n)\\ &\equiv \tilde{\theta}_n + \Delta - B^{-1}A'(AB^{-1}A')^{-1}[a(\tilde{\theta}_n) + A\Delta],\\ \hat{\tilde{\gamma}}_n &= -(AB^{-1}A')^{-1}a(\hat{\tilde{\theta}}_n) \equiv -(AB^{-1}A')^{-1}[a(\tilde{\theta}_n) + A\Delta],\end{aligned}$$

with A and B evaluated at $\tilde{\theta}_n$. To derive these formulae from the first-order conditions for the Lagrangian problem, replace $\nabla_\theta Q_n(\tilde{\theta})$ by the expression $-(\Gamma_n\Omega_n^{-1}\Gamma_n') \times (\hat{\tilde{\theta}}_n - \tilde{\theta}_n)$ from the one-step definition of the unconstrained estimator, replace $a(\tilde{\theta}_n)$ by $a(\hat{\tilde{\theta}}_n) + A(\hat{\tilde{\theta}}_n - \tilde{\theta}_n)$, and use the formula for a partitioned inverse.

9.4. *Special cases*

Maximum likelihood. We have noted that maximum likelihood estimation can be treated as GMM estimation with moments equal to the score, $g = \nabla_\theta \ell$. The statistics in Table 2 remain the same, with the simplification that $B = \Omega (= G = \Gamma)$. The likelihood ratio statistic $2n[L_n(\hat{\theta}_n) - L_n(\bar{\theta}_n)]$, where $L_n(\theta) = (1/n)\sum_{t=1}^n \ell(z_t, \theta)$, is shown by a Taylor's expansion about $\hat{\theta}_n$ to be asymptotically equivalent to the Wald statistic W_{3n}, and hence to all the statistics in Table 2.

Suppose one sets up an estimation problem in terms of a maximum likelihood criterion, but that one does not in fact have the true likelihood function. Suppose that in spite of this misspecification, optimization of the selected criterion yields consistent estimates. One place this commonly arises is when panel data observations are serially correlated, but one writes down the *marginal* likelihoods of the observations ignoring serial correlation. These are sometimes called *pseudo-likelihood* criteria. The resulting estimators can be interpreted as GMM estimators, so that hypotheses can be tested using the statistics in Table 2. Note however that now $G \neq \Omega$, so that $B = G'\Omega^{-1}G$ must be estimated in full, and one cannot do tests using a likelihood ratio of the pseudo-likelihood function.

Least squares. Consider the nonlinear regression model $y = h(x, \theta) + \varepsilon$, and suppose $E(y|x) = h(x, \theta)$ and $E[\{y - h(x, \theta)\}^2|x] = \sigma^2$. Minimizing the least squares criterion $Q_n(\theta) = \sum_{t=1}^n [y_t - h(z_t, \theta)]^2$ is asymptotically equivalent to GMM estimation with $g(z, \theta) = [y - h(x, \theta)]\nabla_\theta h(x, \theta)$ and a distance metric $\Omega_n = (\sigma^2/n)\sum_{t=1}^n [\nabla_\theta h(x, \theta_0)] \times [\nabla_\theta h(x, \theta_0)]'$. For this problem, $B = \Omega = G$. If $h(z_t, \theta) = z_t\theta$ is linear, one has $g(z_t, \theta) = u_t(\theta)z_t$, where $u_t(\theta) = y_t - z_t\theta$ is the regression residual, and $\Omega_n = (\sigma^2/n)\sum_{t=1}^n z_t z_t'$.

Instrumental variables. Consider the regression model $y_t = h(z_t, \theta_0) + \varepsilon_t$ where ε_t may be correlated with $\nabla_\theta h(z_t, \theta_0)$. Suppose there are *instruments* w such that $E(\varepsilon_t|w_t) = 0$. For this problem, one has the moment conditions $g(y_t, z_t, w_t, \theta) = [y_t - h(z_t, \theta)]f(w_t)$ satisfying $Eg(y_t, z_t, w_t, \theta_0) = 0$ for any vector of functions $f(w)$ of

the instruments, so the GMM criterion becomes

$$Q_n(\theta) = \frac{1}{2}\left[\frac{1}{n}\sum_{t=1}^{n}\{y_t - h(z_t,\theta)\}f(w_t)\right]' \Omega_n^{-1}\left[\frac{1}{n}\sum_{t=1}^{n}\{y_t - h(z_t,\theta)\}f(w_t)\right]$$

with $\Omega_n = (\sigma^2/n)\sum_{t=1}^{n} f(w_t)f(w_t)'$. Suppose that it were feasible to construct the conditional expectation of the gradient of the regression function conditioned on w, $q_t = E[\nabla_\theta h(z_t,\theta_0)|w_t]$. This is the optimal vector of functions of the instruments, in the sense that the GMM estimator based on $f(w) = q$ will yield estimators with an asymptotic covariance matrix that is smaller in the positive definite sense than any other distinct vector of functions of w. A feasible GMM estimator with good efficiency properties may then be obtained by first obtaining a preliminary $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$ employing a simple practical distance metric, second, regressing $\nabla_\theta h(z_t,\tilde{\theta}_n)$ on a flexible family of functions of w_t, such as low-order polynomials in w, and, third, using fitted values from this regression as the vector of functions $f(w_t)$ in a final GMM estimation. Note that only one Newton–Raphson step is needed in the last stage. Simplifications of this problem result when $h(z,\theta) = z\theta$ is linear in θ; in this case, the feasible procedure above is simply 2SLS, and no iteration is needed.

Simple hypotheses. An important practical case of the general nonlinear hypothesis $a(\theta_0) = 0$ is that a subset of the parameters are zero. (A hypothesis that parameters equal constants other than zero can be reduced to this case by reparametrization.) Assume $\theta' = \underset{1\times(k-r)}{(\alpha'}, \underset{1\times r}{\beta'})$ and H_0: $\beta = 0$. The first-order conditions for solution of this problem are $0 = \sqrt{n}\nabla_\alpha Q_n(\bar{\theta}_n)$, $0 = \sqrt{n}\nabla_\beta Q_n(\bar{\theta}_n) + \sqrt{n}\bar{\gamma}_n$, and $0 = \bar{\beta}_n$, implying $\bar{\gamma}_n = -\nabla_\beta Q_n(\bar{\theta}_n)$, and $A = \left[\underset{r\times(k-r)}{0} \middle| \underset{r\times r}{I_r}\right]$. Let $C \equiv B^{-1}$ be the asymptotic covariance matrix of $\sqrt{n}(\hat{\theta}_n - \theta_0)$, and $AB^{-1}A' = C_{\beta\beta}$ the submatrix of C for β. Taylor's expansions about $\hat{\theta}_n$ of the first-order conditions imply $\sqrt{n}(\hat{\alpha}_n - \bar{\alpha}_n) = -B_{\alpha\alpha}B_{\alpha\beta}\sqrt{n}\hat{\beta}_n + o_p$ and $\sqrt{n}\bar{\gamma}_n = [B_{\beta\beta} - B_{\beta\alpha}B_{\alpha\alpha}^{-1}B_{\alpha\beta}]\sqrt{n}\hat{\beta}_n + o_p = \hat{\beta}_n' C_{\beta\beta}^{-1}\hat{\beta}_n + o_p$. Then the Wald statistics are

$$W_{1n} = n\hat{\beta}' C_{\beta\beta}^{-1}\hat{\beta},\ W_{2n} = n\begin{bmatrix}\hat{\alpha}_n - \bar{\alpha}_n\\ \hat{\beta}_n\end{bmatrix}'\begin{bmatrix}B_{\alpha\beta}\\ B_{\beta\beta}\end{bmatrix}C_{\beta\beta}^{-1}[B_{\beta\alpha}\ B_{\beta\beta}]\begin{bmatrix}\hat{\alpha}_n - \bar{\alpha}_n\\ \hat{\beta}_n\end{bmatrix},$$

$$W_{3n} = n\begin{bmatrix}\hat{\alpha}_n - \bar{\alpha}_n\\ \hat{\beta}_n\end{bmatrix}'B\begin{bmatrix}\hat{\alpha}_n - \bar{\alpha}_n\\ \hat{\beta}_n\end{bmatrix}.$$

One can check the asymptotic equivalence of these statistics by substituting the expression for $\sqrt{n}(\hat{\alpha}_n - \bar{\alpha}_n)$. The LM statistic, in any version, becomes $\mathrm{LM}_n = n\nabla_\beta Q_n(\bar{\theta}_n)' C_{\beta\beta}\nabla_\beta Q_n(\bar{\theta}_n)$. Recall that B, hence C, can be evaluated at any consistent estimator of θ_0. In particular, the constrained estimator is consistent under the null

or under local alternatives. The LM testing procedure for this case is then to (a) compute the constrained estimator $\bar{\alpha}_n$ subject to the condition $\beta = 0$, (b) calculate the gradient and Hessian of Q_n with respect to the full parameter vector, evaluated at $\bar{\alpha}_n$ and $\beta = 0$, and (c) form the quadratic form above for LM_n from the β part of the gradient and the β submatrix of the inverse of the Hessian. Note that this does not require any iteration of the GMM criterion with respect to the full parameter vector.

It is also possible to carry out the calculation of the LM_n test statistic using auxiliary regressions. This could be done using the auxiliary regression technique introduced earlier for the calculation of LM_{3n} in the case of any nonlinear hypothesis, but a variant is available for this case that reduces the size of the regressions required. The steps are as follows:

(a) Regress $\nabla_\alpha \ell(z_t, \bar{\theta}_n)'$ and $\nabla_\beta \ell(z_t, \bar{\theta}_n)'$ on $g(z_t, \bar{\theta}_n)$, and retrieve the fitted values $\nabla_\alpha \hat{\ell}(z_t, \bar{\theta}_n)$ and $\nabla_\beta \hat{\ell}(z_t, \bar{\theta}_n)$.

(b) Regress $\nabla_\beta \hat{\ell}(z_t, \bar{\theta}_n)$ on $\nabla_\alpha \hat{\ell}(z_t, \bar{\theta}_n)$, and retrieve the *residual* $u(z_t, \bar{\theta}_n)$.

(c) Regress the constant 1 on the residual $u(z_t, \bar{\theta}_n)$, and calculate the sum of squares of the *fitted* values of 1. This quantity is LM_n.

To justify this method, start from the gradient of the GMM criterion,

$$0 = \nabla_\alpha Q_n(\bar{\alpha}_n, 0) = -G_{n\alpha}\Omega_n^{-1}\hat{g}_n(\bar{\alpha}_n, 0),$$

$$\nabla_\beta Q_n(\bar{\alpha}_n, 0) = -G_{n\beta}\Omega_n^{-1}\hat{g}_n(\bar{\alpha}_n, 0),$$

where G_n is partitioned into its α and β submatrices. From the formula for the partitioned inverses, one has for $C = B^{-1}$ the expression

$$C_{\beta\beta} = [\Gamma_\beta \Omega^{-1}\Gamma_\beta' - \Gamma_\beta \Omega^{-1}\Gamma_\alpha'(\Gamma_\alpha \Omega^{-1}\Gamma_\alpha')^{-1}\Gamma_\alpha \Omega^{-1}\Gamma_\beta']^{-1}.$$

The fitted values from step (a) satisfy

$$\nabla_\beta \hat{\ell}(z_t, \bar{\theta}_n)' = g(z_t, \bar{\theta}_n)\Omega_n^{-1}G_{n\beta}',$$

and

$$\nabla_\alpha \hat{\ell}(z_t, \bar{\theta}_n)' = g(z_t, \bar{\theta}_n)\Omega_n^{-1}G_{n\alpha}'.$$

Then the residuals from step (b) satisfy

$$u(z_t, \bar{\theta}_n) = g(z_t, \bar{\theta}_n)\Omega_n^{-1}G_{n\beta}' - g(z_t, \bar{\theta}_n)\Omega_n^{-1}G_{n\alpha}'(G_{n\alpha}\Omega_n^{-1}G_{n\alpha}')^{-1}G_{n\alpha}\Omega_n^{-1}G_{n\beta}'.$$

Then

$$\frac{1}{n}\sum_{t=1}^{n} u(z_t, \bar{\theta}_n)' = \nabla_\beta Q_n(\bar{\alpha}_n, 0)' - \nabla_\alpha Q_n(\bar{\alpha}_n, 0)'(G_{n\alpha}\Omega_n^{-1}G_{n\alpha}')^{-1}G_{n\alpha}\Omega_n^{-1}G_{n\beta}'$$

$$\equiv \nabla_\beta Q_n(\bar{\alpha}_n, 0)',$$

and

$$\frac{1}{n}\sum_{t=1}^{n} u(z_t, \bar{\theta}_n)u(z_t, \bar{\theta}_n)' = C_{n\beta\beta}^{-1}.$$

Then, the step (c) regression yields LM_n. In the case of maximum likelihood estimation, step (a) is redundant and can be omitted.

9.5. *Tests for overidentifying restrictions*

Consider the GMM estimator based on moments $g(z_t, \theta)$, where g is $m \times 1$, θ is $k \times 1$, and $m > k$, so there are *overidentifying moments*. The criterion

$$Q_n(\theta) = -\tfrac{1}{2}\hat{g}_n(\theta)'\Omega_n^{-1}\hat{g}_n(\theta),$$

evaluated at its maximizing argument $\hat{\theta}_n$ for any $\Omega_n \xrightarrow{\mathrm{p}} \Omega$, has the property that $-2n\hat{Q}_n \equiv -2nQ_n(\hat{\theta}_n) \xrightarrow{\mathrm{d}} \chi^2_{m-k}$ under the null hypothesis that $Eg(z, \theta_0) = 0$. This statistic then provides a specification test for the overidentifying moments in g. It can also be used as an indicator for convergence in numerical search for $\hat{\theta}_n$.

To demonstrate this result, recall from eqs. (9.1) and (9.2) that $-\Omega^{-1/2}\sqrt{n}\hat{g}_n(\theta_0) = \mathscr{U}_n \xrightarrow{\mathrm{d}} \mathscr{U} \sim N(0, I)$ and $\sqrt{n}(\hat{\theta}_n - \theta_0) = B^{-1}G'\Omega^{-1/2}\mathscr{U}_n + o_{\mathrm{p}}$. Then, a Taylor's expansion yields

$$\sqrt{n}\hat{g}_n(\hat{\theta}_n) = -\Omega_n^{1/2}\mathscr{U}_n + G_n(G_n'\Omega_n^{-1}G_n)^{-1}G_n'\Omega_n^{-1/2}\mathscr{U}_n + o_{\mathrm{p}} = -\Omega_n^{1/2}R_n\mathscr{U}_n + o_{\mathrm{p}},$$

where $R_n = I - \Omega_n^{-1/2}G_n(G_n'\Omega_n^{-1}G_n)^{-1}G_n'\Omega_n^{-1/2}$ is idempotent of rank $m-k$. Then

$$-2nQ_n(\hat{\theta}_n) = \mathscr{U}_n'R_n\mathscr{U}_n + o_{\mathrm{p}} \xrightarrow{\mathrm{d}} \chi^2_{m-k}.$$

Suppose that instead of estimating θ using the full list of moments, one uses a linear combination $Lg(z, \theta)$, where L is $r \times m$ with $k \leqslant r < m$. In particular, L may select a subset of the moments. Let $\bar{\theta}_n$ denote the GMM estimator obtained from these moment combinations, and assume the identification conditions are satisfied so $\bar{\theta}_n$ is $\sqrt{n}$-consistent. Then the statistic $S = n\hat{g}_n(\bar{\theta}_n)'\Omega_n^{-1/2}R_n\Omega_n^{-1/2}\hat{g}_n(\bar{\theta}_n) \xrightarrow{\mathrm{d}} \chi^2_{m-k}$ under H_0, and this statistic is asymptotically equivalent to the statistic $-2nQ_n(\hat{\theta}_n)$. This result holds for any $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$ of θ_0, not necessarily the optimal GMM estimator for the moments $Lg(z, \theta)$, or even an initially consistent estimator based on only these moments. The distance metric in the center of the quadratic form S does not depend on L, so that the formula for the statistic is invariant with respect to the choice of the initially consistent estimator. This implies in particular that the test statistics S for overidentifying restrictions, starting from

different subsets of the moment conditions, are all asymptotically equivalent. However, the presence of the idempotent matrix R_n in the center of the quadratic form S is critical to its statistical properties. Only the GMM distance metric criterion using all moments, evaluated at $\hat{\theta}_n$, is asymptotically equivalent to S. Substitution of another $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$ in place of $\hat{\theta}_n$ yields an asymptotically equivalent version of S, but $-2nQ_n(\tilde{\theta}_n)$ is not asymptotically chi-square distributed.

These results are a simple corollary of the one-step theorem. Starting from $\tilde{\theta}_n$, the one-step estimator of $\hat{\theta}_n$ is $\sqrt{n}(\hat{\tilde{\theta}}_n - \tilde{\theta}_n) = -(G_n'\Omega_n^{-1}G_n)^{-1}G_n'\Omega_n^{-1}\hat{g}_n(\tilde{\theta}_n)$. Then, one has a one-step estimator $\sqrt{n}\hat{g}_n(\hat{\tilde{\theta}}_n) = \sqrt{n}\hat{g}_n(\tilde{\theta}_n) + G_n\sqrt{n}(\hat{\tilde{\theta}}_n - \tilde{\theta}_n) = \Omega_n^{-1/2}R_n\Omega_n^{-1/2} \times \sqrt{n}\hat{g}_n(\tilde{\theta}_n)$. Substituting this expression in the formula for $-2nQ_n(\hat{\theta}_n)$ yields the statistic S.

The test for overidentifying restrictions can be recast as an LM test by artificially embedding the original model in a richer model. Partition the moments

$$g(z,\theta) = \begin{bmatrix} g^1(z,\theta) \\ g^2(z,\theta) \end{bmatrix},$$

where g^1 is $k \times 1$ with $G_1 = E\nabla_\theta g^1(z,\theta_0)$ of rank k, and g^2 is $(m-k)\times 1$ with $G_2 = E\nabla_\theta g^2(z,\theta_0)$. Embed this in the model

$$\tilde{g}(z,\theta,\psi) = \begin{bmatrix} g^1(z,\theta) \\ g^2(z,\theta) + \psi \end{bmatrix},$$

where ψ is an $(m-k)$ vector of additional parameters. The first-order condition for GMM estimation of this expanded model is

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} G_{1n} & G_{2n} \\ 0 & I_{m-k} \end{bmatrix}\begin{bmatrix} \Omega_n^{-1} & 0 \\ 0 & I_{m-k} \end{bmatrix}\begin{bmatrix} \hat{g}_n^1(\bar{\theta}_n) \\ \hat{g}_n^2(\bar{\theta}_n) - \bar{\psi}_n \end{bmatrix}.$$

The second block of conditions are satisfied by $\bar{\psi}_n = \hat{g}_n^2(\bar{\theta}_n)$, no matter what $\bar{\theta}_n$, so $\bar{\theta}_n$ is determined by $0 = G_{1n}'\Omega_n^{-1}\hat{g}_n^1(\bar{\theta}_n)$. This is simply the estimator obtained from the first block of moments, and coincides with the earlier definition of $\bar{\theta}_n$. Thus, *unconstrained* estimation of the *expanded* model coincides with *restricted* estimation of the original model.

Next consider GMM estimation of the expanded model subject to H_0: $\psi = 0$. This constrained estimation obviously coincides with GMM estimation using all moments in the original model, and yields $\hat{\theta}_n$. Thus, *constrained* estimation of the *expanded* model coincides with *unrestricted* estimation of the original model.

The distance metric test statistic for the constraint $\psi = 0$ in the expanded model is $\mathrm{DM}_n = -2n[\tilde{Q}_n(\hat{\theta}_n, 0) - \tilde{Q}_n(\bar{\theta}_n, \bar{\psi}_n)] \equiv -2nQ_n(\hat{\theta}_n)$, where $\tilde{Q}$ denotes the criterion as a function of the expanded parameter list. One has $\tilde{Q}_n(\hat{\theta}_n, 0) \equiv Q_n(\hat{\theta}_n)$ from the coincidence of the constrained expanded model estimator and the unrestricted

original model estimator, and one has $\tilde{Q}_n(\bar{\theta}_n, \bar{\psi}_n) = 0$ since the number of moments equals the number of parameters. Then, the test statistic $-2nQ_n(\hat{\theta}_n)$ for overidentifying restrictions is identical to a distance metric test in the expanded model, and hence asymptotically equivalent to any of the trinity of tests for H_0: $\psi = 0$ in the expanded model.

We give four examples of econometric problems that can be formulated as tests for overidentifying restrictions:

Example 9.1

If $y = x\beta + \varepsilon$ with $E(\varepsilon|x) = 0$, $E(\varepsilon^2|x) = \sigma^2$, then the moments

$$g^1(z, \beta) = \begin{bmatrix} x(y - x\beta) \\ (y - x\beta)^2 - \sigma^2 \end{bmatrix}$$

can be used to estimate β and σ^2. If ε is normal, then these GMM estimators are MLE. Normality can be tested via the additional moments that give skewness and kurtosis,

$$g^2(x, \beta) = \begin{bmatrix} (y - x\beta)^3/\sigma^3 \\ (y - x\beta)^4/\sigma^4 - 3 \end{bmatrix}.$$

Example 9.2

In the linear model $y = xb + \varepsilon$ with $E(\varepsilon|x) = 0$ and $E(\varepsilon_t\varepsilon_s|x) = 0$ for $t \neq s$, but with possible heteroskedasticity of unknown form, one gets the OLS estimates b of β and $V(b) = s^2(X'X)^{-1}$ under the null hypothesis of homoskedasticity. A test for homoskedasticity can be based on the population moments $0 = E\ \text{vecu}[x'x(\varepsilon^2 - \sigma^2)]$, where "vecu" means the vector formed from the upper triangle of the array. The sample value of this moment vector is

$$\text{vecu}\left[\frac{1}{n}\sum_{t=1}^{n} x_t'x_t\{(y_t - x_t\beta)^2 - s^2\}\right],$$

the difference between the White robust estimator and the standard OLS estimator of $\text{vecu}[X'\Omega X]$.

Example 9.3

If $\ell(z, \theta)$ is the log-likelihood of an observation, and $\hat{\theta}_n$ is the MLE, then an additional moment condition that should hold if the model is specified correctly is the information matrix equality

$$0 = E\nabla_{\theta\theta}\ell(z, \theta_0) + E\nabla_\theta\ell(z, \theta_0)\nabla_\theta\ell(z, \theta_0)'.$$

The sample analog is White's information matrix test, which then can be interpreted as a GMM test for overidentifying restrictions.

Example 9.4

In the nonlinear model $y = h(x, \theta) + \varepsilon$ with $E(\varepsilon|x) = 0$, and $\bar{\theta}_n$ a GMM estimator based on moments $w(x)[y - h(x, \theta)]$, where $w(x)$ is some vector of functions of x, suppose one is interested in testing the stronger assumption that ε is *independent* of x. A necessary and sufficient condition for independence is $E[w(x) - Ew(x)] \times f[y - h(x, \theta_0)] = 0$ for every function f and vector of functions w for which the moments exist. A specification test can be based on a selection of such moments.

9.6. *Specification tests in linear models*[51]

GMM tests for overidentifying restrictions have particularly convenient forms in linear models; see Newey and West (1988) and Hansen and Singleton (1982). Three standard specification tests will be shown to have this interpretation. We summarize a few properties of projections that will be used in the following discussion. Let $\mathscr{P}_X = X(X'X)^- X$ denote the *projection matrix* from $\mathbb{R}^n$ onto the linear subspace $\mathbb{X}$ *spanned* by an $n \times p$ array X. (We use a Moore–Penrose generalized inverse in the definition of $\mathscr{P}_X$ to handle the possibility that X is less than full rank; see Section 9.8.) Let $\mathscr{Q}_X = I - \mathscr{P}_X$ denote the projection matrix onto the linear subspace orthogonal to $\mathbb{X}$. Note that $\mathscr{P}_X$ and $\mathscr{Q}_X$ are idempotent. If $\mathbb{X}$ is a subspace generated by an array X and $\mathbb{W}$ is a subspace generated by an array $W = [X\ Z]$ that contains X, then $\mathscr{P}_X\mathscr{P}_W = \mathscr{P}_W\mathscr{P}_X = \mathscr{P}_X$; i.e. a projection onto a subspace is left invariant by a further projection onto a larger subspace, and a two-stage projection onto a large subspace followed by a projection onto a smaller one is the same as projecting directly onto the smaller one. The subspace of $\mathbb{W}$ that is orthogonal to $\mathbb{X}$ is generated by $\mathscr{Q}_X W$; i.e., it is the set of linear combinations of the residuals, orthogonal to $\mathbb{X}$, obtained by regressing W on X. Any y in $\mathbb{R}^n$ has a unique decomposition $y = \mathscr{P}_X y + \mathscr{Q}_X\mathscr{P}_W y + \mathscr{Q}_W y$ into the sum of projections onto $\mathbb{X}$, the subspace of $\mathbb{W}$ orthogonal to $\mathbb{X}$, and the subspace orthogonal to $\mathbb{W}$. The projection $\mathscr{Q}_X\mathscr{P}_W$ can be rewritten $\mathscr{Q}_X\mathscr{P}_W = \mathscr{P}_W - \mathscr{P}_X = \mathscr{P}_W\mathscr{Q}_X = \mathscr{Q}_X\mathscr{P}_W\mathscr{Q}_X$, or since $\mathscr{Q}_X W = \mathscr{Q}_X[X\ Z] = [0\ \mathscr{Q}_X Z]$, $\mathscr{Q}_X\mathscr{P}_W = \mathscr{P}_{\mathscr{Q}_X W} = \mathscr{P}_{\mathscr{Q}_X Z} = \mathscr{Q}_X Z(Z'\mathscr{Q}_X Z)^- Z'\mathscr{Q}_X$. This implies that $\mathscr{Q}_X\mathscr{P}_W$ is idempotent since $(\mathscr{Q}_X\mathscr{P}_W)(\mathscr{Q}_X\mathscr{P}_W) = \mathscr{Q}_X(\mathscr{P}_W\mathscr{Q}_X)\mathscr{P}_W = \mathscr{Q}_X(\mathscr{Q}_X\mathscr{P}_W)\mathscr{P}_W = \mathscr{Q}_X\mathscr{P}_W$.

Omitted variables test: Consider the regression model $y = X\beta + \varepsilon$, where y is $n \times 1$, X is $n \times k$, $E(\varepsilon|X) = 0$, and $E(\varepsilon\varepsilon'|X) = \sigma^2 I$. Suppose one has the hypothesis H_0: $\beta_1 = 0$, where β_1 is a $p \times 1$ subvector of β. Define $u = y - Xb$ to be the residual associated with an estimator b of β. The GMM criterion is then $2nQ = u'X(X'X)^{-1} \cdot X'u/\sigma^2$. The *projection matrix* $\mathscr{P}_X \equiv X(X'X)^{-1}X'$ that appears in the center of this criterion can obviously be decomposed as $\mathscr{P}_X \equiv \mathscr{P}_{X_2} + (\mathscr{P}_X - \mathscr{P}_{X_2})$. Under H_0,

[51] Paul Ruud contributed substantially to this section.

$u = y - X_2 b_2$ and $X'u$ can be interpreted as $k = p + q$ overidentifying moments for the q parameters β_2. Then, the GMM test statistic for overidentifying restrictions is the minimum value $-2n\hat{Q}_n$ in b_2 of $u'\mathscr{P}_X u/\sigma^2$. But $\mathscr{P}_X u = \mathscr{P}_{X_2} u + (\mathscr{P}_X - \mathscr{P}_{X_2})y$ and $\min_{b_2} u'\mathscr{P}_{X_2} u = 0$ (at the OLS estimator under H_0 that makes u orthogonal to X_2). Then $-2n\hat{Q}_n = y'(\mathscr{P}_X - \mathscr{P}_{X_2})y/\sigma^2$. The unknown variance σ^2 in this formula can be replaced by any consistent estimator s^2, in particular the estimated variance of the disturbance from either the restricted or the unrestricted regression, without altering the asymptotic distribution, which is χ^2_q under the null hypothesis.

The statistic $-2n\hat{Q}_n$ has three alternative interpretations. First,

$$-2n\hat{Q}_n = y'\mathscr{P}_X y/\sigma^2 - y'\mathscr{P}_{X_2} y/\sigma^2 = \frac{\mathrm{SSR}_{X_2} - \mathrm{SSR}_X}{\sigma^2},$$

which is the difference of the sum of squared residuals from the restricted regression under H_0 and the sum of squared residuals from the unrestricted regression, normalized by σ^2. This is a large sample version of the usual finite sample F-test for H_0. Second, note that the fitted value of the dependent variable from the restricted regression is $\hat{y}_0 = \mathscr{P}_{X_2} y$, and from the unrestricted regression is $\hat{y}_u = \mathscr{P}_X y$, so that

$$-2n\hat{Q}_n = (\hat{y}_0'\hat{y}_0 - \hat{y}_u'\hat{y}_u)/\sigma^2 = (\hat{y}_0 - \hat{y}_u)'(\hat{y}_0 - \hat{y}_u)/\sigma^2 = \|\hat{y}_0 - \hat{y}_u\|^2/\sigma^2.$$

Then, the statistic is calculated from the distance between the fitted values of the dependent variable with and without H_0 imposed. Note that this computation requires no covariance matrix calculations. Third, let b_0 denote the GMM estimator restricted by H_0 and b_u denote the unrestricted GMM estimator. Then, b_0 consists of the OLS estimator for β_2 and the hypothesized value 0 for β_1, while b_u is the OLS estimator for the full parameter vector. Note that $\hat{y}_0 = Xb_0$ and $\hat{y}_u = Xb_u$, so that $\hat{y}_0 - \hat{y}_u = X(b_0 - b_u)$. Then

$$-2n\hat{Q}_n = (b_0 - b_u)'(X'X/\sigma^2)(b_0 - b_u) = (b_0 - b_u)'V(b_u)^{-1}(b_0 - b_u).$$

This is the Wald statistic W_{3n}. From the equivalent form W_{2n} of the Wald statistic, this can also be written as a quadratic form $-2n\hat{Q}_n = b_{1,u}'V(b_{1,u})^{-1}b_{1,u}$, where $b_{1,u}$ is the subvector of unrestricted estimates for the parameters that are zero under the null hypothesis.

The Hausman exogeneity test: Consider the regression $y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon$, and the null hypothesis that X_1 is exogenous, where X_2 is known to be exogenous, and X_3 is known to be endogenous. Suppose N is an array of instruments, including X_2, that are sufficient to identify the coefficients when the hypothesis is false. Let $W = [N\ X_1]$ be the full set of instruments available when the null hypothesis is true. Then the best instruments under the null hypothesis are $\hat{X}_0 = \mathscr{P}_W X \equiv [X_1\ X_2\ \hat{X}_3]$, and the best instruments under the alternative are $\hat{X}_u = \mathscr{P}_N X \equiv [\hat{X}_1\ X_2\ \hat{X}_3]$. The test statistic for overidentifying restrictions is $-2n\hat{Q}_n = y'(\mathscr{P}_{\hat{X}_0} - \mathscr{P}_{\hat{X}_u})y/\sigma^2$, as in the previous case. This can be written $-2nQ_n = (\mathrm{SSR}_{\hat{X}_u} - \mathrm{SSR}_{\hat{X}_0})/\sigma^2$,

with the numerator the difference in sum of squared residuals from an OLS regression of y on $\hat{X}_u$ and an OLS regression of y on $\hat{X}_0$. Also, $-2n\hat{Q}_n = \| \hat{y}_{\hat{X}_0} - \hat{y}_{\hat{X}_u} \|^2/\sigma^2$, the difference between the fitted values of y from a regression on $\hat{X}_u$ and a regression on $\hat{X}_0$. Finally,

$$-2n\hat{Q}_n = (b_{2SLS_0} - b_{2SLS_u})'[V(b_{2SLS_u}) - V(b_{2SLS_0})]^{-}(b_{2SLS_0} - b_{2SLS_u}),$$

an extension of the Hausman–Taylor exogeneity test to the problem where some variables are suspect and others are known to be exogenous. Newey and West (1988) show that the matrix in the center of this quadratic form has rank equal to the rank of X_1, and that the test statistic can be written equivalently as a quadratic form in the subvector of differences of the 2SLS estimates for the X_1 coefficients, with the ordinary inverse of the corresponding submatrix of differences of variances in the center of the quadratic form.

Testing for overidentifying restrictions in a structural system: Consider an equation $y = X\beta + \varepsilon$ from a system of simultaneous equations, and let W denote the array of instruments (exogenous and predetermined variables) in the system. Let $\hat{X} = \mathcal{P}_W X$ denote the fitted values of X obtained from OLS estimation of the reduced form. The equation is *overidentified* if the number of instruments W exceeds the number of right-hand-side variables X. The GMM test statistic for overidentification is the minimum in β of

$$-2nQ_n(\beta) = u'\mathcal{P}_W u/\sigma^2 = u'\mathcal{P}_{\hat{X}} u/\sigma^2 + u'(\mathcal{P}_W - \mathcal{P}_{\hat{X}})u/\sigma^2,$$

where $u = y - X\beta$. As before, $-2n\hat{Q}_n = y'(\mathcal{P}_W - \mathcal{P}_{\hat{X}})y/\sigma^2$. Under H_0, this statistic is asymptotically chi-squared distributed with degrees of freedom equal to the difference in ranks of W and $\hat{X}$. This statistic can be interpreted as the difference in the sum of squared residuals from the 2SLS regression of y on X and the sum of squared residuals from the reduced form regression of y on W, normalized by σ^2. A computationally convenient equivalent form is $-2n\hat{Q}_n = \| \hat{y}_W - \hat{y}_{\hat{X}} \|^2/\sigma^2$, the sum of squares of the difference between the reduced form fitted values and the 2SLS fitted values of y, normalized by σ^2. Finally, $-2n\hat{Q}_n = y'\mathcal{P}_{\hat{X}}\mathcal{P}_W\mathcal{P}_{\hat{X}}y/\sigma^2 = nR^2/\sigma^2$, where R^2 is the multiple correlation coefficient from regressing the 2SLS residuals on all the instruments; this result follows from the equivalent formulae for the projection onto the subspace of $\mathbb{W}$ orthogonal to the subspace spanned by $\hat{X}$. This test statistic does *not* have a version that can be written as a quadratic form with the wings containing a difference of coefficient estimates from the 2SLS and reduced form regressions.

9.7. *Specification testing in multinomial models*

As applications of GMM testing, we consider hypotheses arising in the context of analysis of discrete response data. The first example is a test for omitted variables

in multinomial data, which extends to various tests of functional specification by introduction of appropriate omitted variables. The second example tests for the presence of random effects in discrete panel data.

Example 9.5

Suppose J multinomial outcomes are indexed $C=\{1,\dots,J\}$. Define $z=(d_1,\dots,d_J,x)$, where d_j is one if outcome j is observed, and zero otherwise. The x are exogenous variables. The log-likelihood of an observation is

$$\ell(z,\theta)=\sum_{i\in C} d_i \log P_C(i,x,\theta),$$

where $P_C(i,x,\theta)$ is the probability that i is observed from C, given x. Suppose $\theta=(\alpha,\beta)$, and the null hypothesis H_0: $\beta=0$. We derive an LM test starting from the maximum likelihood estimates of α under the constraint $\beta=0$. Define

$$u_i=[d_i-P_C(i,x,\bar\theta_n)]P_C(i,x,\bar\theta_n)^{-1/2},$$
$$q_i=P_C(i,x,\bar\theta_n)^{1/2}\nabla_\theta \log P_C(i,x,\bar\theta_n).$$

Then, in a sample $t=1,\dots,n$, one has $(1/n)\sum_{t=1}^n \nabla_\theta \ell(z_t,\bar\theta_n)\equiv(1/n)\sum_{t=1}^n\sum_{i\in C} q_{it}u_{it}$. Also, $(1/n)\sum_{t=1}^n\sum_{i\in C} q_i q_i' \xrightarrow{\text{p}} \Omega$ since

$$\Omega=-E\nabla_{\theta\theta'}\ell\equiv -E\nabla_\theta \sum_{i\in C}[d_i-P_C(i,x,\theta_0)]\nabla_\theta \log P_C(i,x,\theta_0)$$
$$=E\sum_{i\in C} P_C(i,x,\theta_0)[\nabla_\theta \log P(i,x,\bar\theta)][\nabla_\theta \log P(i,x,\bar\theta)]'.$$

Then,

$$\mathrm{LM}_{3n}=n\left[\frac{1}{n}\sum_{t=1}^n\sum_{i\in C} q_{it}u_{it}\right]'\left[\frac{1}{n}\sum_{t=1}^n\sum_{i\in C} q_{it}q_{it}'\right]^{-1}\left[\frac{1}{n}\sum_{t=1}^n\sum_{i\in C} q_{it}u_{it}\right].$$

This statistic can be computed from the sum of squares of the fitted values of u_{it} from an auxiliary regression over i and t of u_{it} on q_{it}. If R^2 is the multiple correlation coefficient from this regression, and $\bar u$ is the sample mean of the u_{it}, then $\mathrm{LM}_{3n}=n(J-1)R^2+(1-R^2)\bar u^2$.

McFadden (1987) shows for the multinomial logit model that the Hausman and McFadden (1984) test for the independence from irrelevant alternatives property of this model can be calculated as an omitted variable test of the form above, where the omitted variables are interactions of the original variables and dummy variables for subsets of C where nonindependence is suspected. Similarly, Lagrange multiplier tests of the logit model against nested logit alternatives can be cast as omitted

variable tests where the omitted variables are interactions of dummy variables for suspect subsets A of C and variables of the form $\log[P_C(i,x,\bar{\theta}_n)/\sum_{j\in A} P_C(i,x,\bar{\theta}_n)]$.

Example 9.6

We develop a Lagrange multiplier test for unobserved heterogeneity in discrete panel data. A case is observed to be either in state $d_t = +1$ or $d_t = -1$ in periods $t = 1,\ldots,T$. A probability model for these observations that allows unobserved heterogeneity is

$$P(d_1,\ldots,d_n|x_1,\ldots,x_T,\beta_1,\ldots,\beta_n,\delta) = \int_{-\infty}^{+\infty} \prod_{t=1}^{T} F[d_t(x_t\beta_t + \sqrt{\delta}v)]h(v)\,dv,$$

where $x_1,\ldots,x_T$ are exogenous, $\beta_1,\ldots,\beta_T$ and δ are parameters, F is a cumulative distribution function for a density that is symmetric about zero, and v is an unobserved "case effect" heterogeneity. The density $h(v)$ is normalized so that $Ev = 0$ and $Ev^2 = 1$.

When $\delta = 0$, this model reduces to a series of independent Bernoulli trials,

$$P(d_1,\ldots,d_T|x_1,\ldots,x_T,\beta_1,\ldots,\beta_T,0) = \prod_{t=1}^{T} F(d_t x_t\beta_t),$$

and is easily estimated. For example, F normal yields binary probits, and F logistic yields binary logits. A Lagrange multiplier test for $\delta = 0$ will detect the presence of unobserved heterogeneity across cases. Assume a sample of n cases, drawn randomly from the population. The LM test statistic is

$$\mathrm{LM} = \frac{\left[\sum \nabla_\delta \ell\right]^2}{n\left\{\sum (\nabla_\delta \ell)^2/n - \left[\sum (\nabla_\delta \ell)(\nabla_\beta \ell)'/n\right]\left[\sum (\nabla_\beta \ell)(\nabla_\beta \ell)'/n\right]^{-1}\left[\sum (\nabla_\beta \ell)(\nabla_\delta \ell)'/n\right]\right\}},$$

where ℓ is the log-likelihood of the case, $\nabla_\beta \ell = (\nabla_{\beta_1}\ell,\ldots,\nabla_{\beta_T}\ell)$, and all the derivatives are evaluated at $\delta = 0$ and the Bernoulli model estimates of β. The β derivatives are straightforward,

$$\ell_{\beta_t} = d_t x_t f(d_t x_t\beta_t)/F(d_t x_t\beta_t),$$

where f is the density of F. The δ derivative is more delicate, requiring use of l'Hôpital's rule:

$$\ell_\delta = \frac{1}{2}\left\{\sum_{t=1}^{T}\left[\frac{d_t f'(d_t x_t\beta_t)}{F(d_t x_t\beta_t)} - \frac{f(d_t x_t\beta_t)^2}{F(d_t x_t\beta_t)^2}\right] + \left[\sum_{t=1}^{n}\frac{d_t f(d_t x_t\beta_t)}{F(d_t x_t\beta_t)}\right]^2\right\}.$$

The reason for introducing δ in the form above, so $\sqrt{\delta}v$ appeared in the probability, was to get a statistic where $\sum \nabla_\delta \ell$ was not identically zero. The alternative would have been to develop the test statistic in terms of the first non-identically zero higher derivative; see Lee and Chesher (1986).

The LM statistic can be calculated by regressing the constant 1 on $\nabla_\delta \ell$ and $\nabla_{\beta_1}\ell, \ldots, \nabla_{\beta_n}\ell$, where all these derivatives are evaluated at $\delta = 0$ and the Bernoulli model estimates, and then forming the sum of squares of the fitted values. Note that the LM statistic is independent of the shape of the heterogeneity distribution $h(v)$, and is thus a "robust" test against heterogeneity of any form.

9.8. *Technicalities*

Some test statistics are conveniently defined using generalized inverses. This section gives a constructive definition of a generalized inverse, and lists some of its properties. A matrix $\underset{k\times m}{A^-}$ is a *Moore–Penrose generalized inverse* of a matrix $\underset{m\times k}{A}$ if it has three properties:

(i) $AA^-A = A$,
(ii) $A^-AA^- = A^-$,
(iii) AA^- and A^-A are symmetric.

There are other generalized inverse definitions that have some, but not all, of these properties; in particular A^+ will denote any matrix that satisfies (i).

First, a method for constructing a generalized inverse is described, and then some of the implications of the definition are developed. The construction is called the *singular value decomposition* (SVD) of a matrix, and is of independent interest as a tool for finding the eigenvalues and eigenvectors of a symmetric matrix, and for calculation of inverses of moment matrices of data with high multicollinearity; see Press et al. (1986) for computational algorithms and programs.

Lemma 9.4

Every real $m \times k$ matrix A of rank r can be decomposed into a product

$$\underset{m\times k}{A} = \underset{m\times r}{U}\ \underset{r\times r}{D}\ \underset{r\times k}{V'},$$

where D is a diagonal matrix with positive nonincreasing elements down the diagonal, and U and V are column-orthonormal; i.e. $U'U = I_r = V'V$.

Proof

The $m \times m$ matrix AA' is symmetric and positive semi-definite. Then, there exists an $m \times m$ orthonormal matrix W, partitioned $W = [W_1\ W_2]$ with W_1 of dimension $m \times r$, such that $W_1'(AA')W_1 = G$ is diagonal with positive, nonincreasing diagonal

elements, and $W_2'(AA')W_2 = 0$, implying $A'W_2 = 0$. Define D from G by replacing the diagonal elements of G by their positive square roots. Note that $W'W = I = WW' \equiv W_1W_1' + W_2W_2'$. Define $U = W_1$ and $V' = D^{-1}U'A$. Then, $U'U = I_r$ and $V'V = D^{-1}U'AUD^{-1} = D^{-1}GD^{-1} = I_r$. Further, $A = (I_m - W_2W_2')A = UU'A = UDV'$. This establishes the decomposition. Q.E.D.

Note that if A is symmetric, then U is the array of eigenvectors of A corresponding to the nonzero roots, so that $A'U = UD_1$, with D_1 the $r \times r$ diagonal matrix with the nonzero eigenvalues in descending magnitude down the diagonal. In this case, $V = A'UD^{-1} = UD_1D^{-1}$. Since the elements of D_1 and D are identical except possibly for sign, the columns of U and V are either equal (for positive roots) or reversed in sign (for negative roots).

Lemma 9.5

The Moore–Penrose generalized inverse of an $m \times k$ matrix A is the matrix $A^- = \underset{k\times r}{V}\ \underset{r\times r}{D^{-1}}\ \underset{r\times m}{U'}$. Let A^+ denote any matrix, including A^-, that satisfies $AA^+A = A$. These matrices satisfy:

(1) $A^+ = A^{-1}$ if A is square and nonsingular.
(2) The system of equations $Ax = y$ has a solution if and only if $y = AA^+y$, and the linear subspace of all solutions is the set of vectors $x = A^+y + [I - A^+A]z$ for all $z \in \mathbb{R}^k$.
(3) AA^+ and A^+A are idempotent.
(4) If A is idempotent, then $A = A^-$.
(5) If $A = BCD$ with B and D nonsingular, then $A^- = D^{-1}C^-B^{-1}$, and any matrix $A^+ = D^{-1}C^+B^{-1}$ satisfies $AA^+A = A$.

Proof

Elementary; see Pringle and Rayner (1971).

Lemma 9.6

If A is square, symmetric, and positive semi-definite of rank r, then

(1) There exist Q positive definite and R idempotent of rank r such that $A = QRQ$ and $A^- = Q^{-1}RQ^{-1}$.
(2) There exists $\underset{k\times r}{U}$ column-orthonormal such that $U'AU = D$ is nonsingular diagonal and $A^- = U(U'AU)^{-1}U'$.
(3) A has a symmetric square root $B = A^{1/2}$, and $A^- = B^-B^-$.

Proof

Let $W = [U\ W_2]$ be an orthogonal matrix diagonalizing A. Then, $U'AU = D$, a diagonal matrix of positive eigenvalues, and $AW_2 = 0$. Define $Q = W\begin{bmatrix} D^{1/2} & 0 \\ 0 & I_{m-r} \end{bmatrix}W'$, $R = W\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}W'$, and $B = UD^{1/2}U'$. Q.E.D.

Lemma 9.7

If $y \sim N(A\lambda, A)$, with A of rank r, and A^+ is any symmetric matrix satisfying $AA^+A = A$, then $y'A^+y$ is noncentral chi-square distributed with r degrees of freedom and noncentrality parameter $\lambda'A\lambda$.

Proof

Let $W = [U\ W_2]$ be an orthonormal matrix that diagonalizes A, as in the proof of Lemma 9.6, with $U'AU = D$, a positive diagonal $r \times r$ matrix, and $W'AW_2 = 0$, implying $AW_2 = 0$. Then, the nonsingular transformation $z = \begin{bmatrix} D^{-1/2} & 0 \\ 0 & I \end{bmatrix} W'y$ has mean $\begin{bmatrix} D^{-1/2}U'A\lambda \\ 0 \end{bmatrix}$ and covariance matrix $\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$, so that $z_1 = D^{-1/2}U'y$ is distributed $N(D^{-1/2}U'A\lambda, I_r)$, $z_2 = W_2y = 0$, implying $W'y = [D^{1/2}z_1\ \ 0]$. It is standard that $z'z$ has a noncentral chi-square distribution with r degrees of freedom and noncentrality parameter $\lambda'AUD^{-1}U'A\lambda = \lambda'A\lambda$. The condition $A = AA^+A$ implies $U'AU = U'AWW'A^+WW'AU$, or

$$D = [D\,0]W'A^+W[D\,0]' = D(U'A^+U)D.$$

Hence, $U'A^+U = D^{-1}$. Then

$$\begin{aligned} y'A^+y &= y'WW'A^+WW'y = [z_1'D^{1/2}\ 0](W'A^+W)[D^{1/2}z_1'\ 0]' \\ &= z_1'D^{1/2}(U'A^+U)D^{1/2}z_1 = z_1'z_1. \end{aligned}$$

Q.E.D.

References

Ait-Sahalia, Y. (1993) "Asymptotic Theory for Functionals of Kernel Estimators", MIT Ph.D. thesis.

Amemiya, T. (1973) "Regression Analysis When the Dependent Variable is Truncated Normal". *Econometrica*, 41, 997–1016.

Amemiya, T. (1974) "The Nonlinear Two-Stage Least-Squares Estimator", *Journal of Econometrics*, 2, 105–110.

Amemiya, T. (1985) *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Andersen, P.K. and R.D. Gill (1982) "Cox's Regression Model for Counting Processes: A Large Sample Study", *The Annals of Statistics*, 19, 1100–1120.

Andrews, D.W.K. (1990) "Asymptotics for Semiparametric Econometric Models: I. Estimation and Testing", *Cowles Foundation Discussion Paper* No. 908R.

Andrews, D.W.K. (1992) "Generic Uniform Convergence", *Econometric Theory*, 8, 241–257.

Andrews, D.W.K. (1994) "Empirical Process Methods in Econometrics", in: R. Engle and D. McFadden, eds., *Handbook of Econometrics, Vol. 4*, Amsterdam: North-Holland.

Barro, R.J. (1977) "Unanticipated Money Growth and Unemployment in the United States", *American Economic Review*, 67, 101–115.

Bartle, R.G. (1966) *The Elements of Integration*, New York: John Wiley and Sons.
Bates, C.E. and H. White (1992) "Determination of Estimators with Minimum Asymptotic Covariance Matrices", preprint, University of California, San Diego.
Berndt, E.R., B.H. Hall, R.E. Hall and J.A. Hausman (1974) "Estimation and Inference in Nonlinear Structural Models", *Annals of Economic and Social Measurement*, 3, 653–666.
Bickel, P. (1982) "On Adaptive Estimation," *Annals of Statistics*, 10, 647–671.
Bickel, P., C.A.J. Klaassen, Y. Ritov and J.A. Wellner (1992) "Efficient and Adaptive Inference in Semiparametric Models" Forthcoming monograph, Baltimore, MD: Johns Hopkins University Press.
Billingsley, P. (1968) *Convergence of Probability Measures*, New York: Wiley.
Bloomfeld, P. and W.L. Steiger (1983) *Least Absolute Deviations: Theory, Applications, and Algorithms*, Boston: Birkhauser.
Brown, B.W. (1983) "The Identification Problem in Systems Nonlinear in the Variables", *Econometrica*, 51, 175–196.
Burguete, J., A.R. Gallant and G. Souza (1982) "On the Unification of the Asymptotic Theory of Nonlinear Econometric Models", *Econometric Reviews*, 1, 151–190.
Carroll, R.J. (1982) "Adapting for Heteroskedasticity in Linear Models", *Annals of Statistics*, 10, 1224–1233.
Chamberlain, G. (1982) "Multivariate Regression Models for Panel Data", *Journal of Econometrics*, 18, 5–46.
Chamberlain, G. (1987) "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions", *Journal of Econometrics*, 34, 305–334.
Chesher, A. (1984) "Testing for Neglected Heterogeneity", *Econometrica*, 52, 865–872.
Chiang, C.L. (1956) "On Regular Best Asymptotically Normal Estimates", *Annals of Mathematical Statistics*, 27, 336–351.
Daniels, H.E. (1961) "The Asymptotic Efficiency of a Maximum Likelihood Estimator", in: *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 151–163, Berkeley: University of California Press.
Davidson, R. and J. MacKinnon (1984) "Convenient Tests for Probit and Logit Models", *Journal of Econometrics*, 25, 241–262.
Eichenbaum, M.S., L.P. Hansen and K.J. Singleton (1988) "A Time Series Analysis of Representative Agent Models of Consumption and Leisure Choice Under Uncertainty", *Quarterly Journal of Economics*, 103, 51–78.
Eicker, F. (1967) "Limit Theorems for Regressions with Unequal and Dependent Errors", in: L.M. LeCam and J. Neyman, eds., *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press.
Fair, R.C. and D.M. Jaffee (1972) "Methods of Estimation for Markets in Disequilibrium", *Econometrica*, 40, 497–514.
Ferguson, T.S. (1958) "A Method of Generating Best Asymptotically Normal Estimates with Application to the Estimation of Bacterial Densities", *Annals of Mathematical Statistics*, 29, 1046–1062.
Fisher, F.M. (1976) *The Identification Problem in Econometrics*, New York: Krieger.
Fisher, R.A. (1921) "On the Mathematical Foundations of Theoretical Statistics", *Philosophical Transactions, A*, 222, 309–368.
Fisher, R.A. (1925) "Theory of Statistical Estimation", *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.
Gourieroux, C., A. Monfort and A. Trognon (1983) "Testing Nested or Nonnested Hypotheses", *Journal of Econometrics*, 21, 83–115.
Gourieroux, C., A. Monfort and A. Trognon (1984) "Psuedo Maximum Likelihood Methods: Theory", *Econometrica*, 52, 681–700.
Hajek, J. (1970) "A Characterization of Limiting Distributions of Regular Estimates", *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 14, 323–330.
Hansen, L.P. (1982) "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50, 1029–1054.

Hansen, L.P. (1985a) "A Method for Calculating Bounds on the Asymptotic Covariance Matrices of Generalized Method of Moments Estimators", *Journal of Econometrics*, 30, 203–238.

Hansen, L.P. (1985b) "Notes on Two Step GMM Estimators", Discussion, December meetings of the Econometric Society.

Hansen, L.P. and K.J. Singleton (1982) "Generalized Instrumental Variable Estimation of Nonlinear Rational Expectations Models", *Econometrica*, 50, 1269–1286.

Hansen, L.P., J. Heaton and R. Jagannathan (1992) "Econometric Evaluation of Intertemporal Asset Pricing Models Using Volatility Bounds", mimeo, University of Chicago.

Härdle, W. (1990) *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.

Härdle, W. and O. Linton (1994) "Nonparametric Regression", in: R. Engle and D. McFadden, eds., *Handbook of Econometrics, Vol. 4*, Amsterdam: North-Holland.

Hausman, J.A. (1978) "Specification Tests in Econometrics", *Econometrica*, 46, 1251–1271.

Hausman, J.A. and D. McFadden (1984) "Specification Tests for the Multinomial Logit Model", *Econometrica*, 52, 1219–1240.

Heckman, J.J. (1976) "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models", *Annals of Economic and Social Measurement*, 5, 475–492.

Honoré, B.E. (1992) "Timmed LAD and Least Squares Estimation of Truncated and Censored Models with Fixed Effects", *Econometrica*, 60, 533–565.

Honoré, B.E. and J.L. Powell (1992) "Pairwise Difference Estimators of Linear, Censored, and Truncated Regression Models", mimeo, Northwestern University.

Huber, P.J. (1964) "Robust Estimation of a Location Parameter", *Annals of Mathematical Statistics*, 35, 73–101.

Huber, P. (1967) "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions", in: L.M. LeCam and J. Neyman, eds., *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press.

Huber, P. (1981) *Robust Statistics*, New York: Wiley.

Ibragimov, I.A. and R.Z. Has'minskii (1981) *Statistical Estimation: Asymptotic Theory*, New York: Springer-Verlag.

Jennrich (1969), "Asymptotic Properties of Nonlinear Least Squares Estimators", *Annals of Mathematical Statistics*, 20, 633–643.

Koenker, R. and G. Bassett (1978) "Regression Quantiles", *Econometrica*, 46, 33–50.

LeCam, L. (1956) "On the Asymptotic Theory of Estimation and Testing Hypotheses", in: L.M. LeCam and J. Neyman, eds., *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 129–156, Berkeley: University of California Press.

Lee, L. F. and A. Chesher (1986) "Specification Testing when the Score Statistics are Identically Zero", *Journal of Econometrics*, 31, 121–149.

Maasoumi, E. and P.C.B. Phillips (1982) "On the Behavior of Inconsistent Instrumental Variables Estimators", *Journal of Econometrics*, 19, 183–201.

Malinvaud, E. (1970) "The Consistency of Nonlinear Regressions", *Annals of Mathematical Statistics*, 41, 956–969.

Manski, C. (1975) "Maximum Score Estimation of the Stochastic Utility Model of Choice", *Journal of Econometrics*, 3, 205–228.

McDonald, J.B. and W.K. Newey (1988) "Partially Adaptive Estimation of Regression Models Via the Generalized T Distribution", *Econometric Theory*, 4, 428–457.

McFadden, D. (1987) "Regression-Based Specification Tests for the Multinomial Logit Model", *Journal of Econometrics*, 34, 63–82.

McFadden, D. (1989) "A Method of Simulated Moments for Estimation of Multinomial Discrete Response Models Without Numerical Integration", *Econometrica*, 57, 995–1026.

McFadden, D. (1990) "An Introduction to Asymptotic Theory: Lecture Notes for 14.381", mimeo, MIT.

Newey, W.K. (1984) "A Method of Moments Interpretation of Sequential Estimators", *Economics Letters*, 14, 201–206.
Newey, W.K. (1985) "Generalized Method of Moments Specification Testing", *Journal of Econometrics*, 29, 229–256.
Newey, W.K. (1987) "Asymptotic Properties of a One-Step Estimator Obtained from an Optimal Step Size", *Econometric Theory*, 3, 305.
Newey, W.K. (1988) "Interval Moment Estimation of the Truncated Regression Model", mimeo, Department of Economics, MIT.
Newey, W.K. (1989) "Locally Efficient, Residual-Based Estimation of Nonlinear Simultaneous Equations Models", mimeo, Department of Economics, Princeton University.
Newey, W.K. (1990) "Semiparametric Efficiency Bounds", *Journal of Applied Econometrics*, 5, 99–135.
Newey, W.K. (1991a) "Uniform Convergence in Probability and Stochastic Equicontinuity", *Econometrica*, 59, 1161–1167.
Newey, W.K. (1991b) "Efficient Estimation of Tobit Models Under Conditional Symmetry", in: W. Barnett, J. Powell and G. Tauchen, eds., *Semiparametric and Nonparametric Methods in Statistics and Econometrics*, Cambridge: Cambridge University Press.
Newey, W.K. (1992a) "The Asymptotic Variance of Semiparametric Estimators", MIT Working Paper.
Newey, W.K. (1992b) "Partial Means, Kernel Estimation, and a General Asymptotic Variance Estimator", mimeo, MIT.
Newey, W.K. (1993) "Efficient Two-Step Instrumental Variables Estimation", mimeo, MIT.
Newey, W.K. and J.L. Powell (1987) "Asymmetric Least Squares Estimation and Testing", *Econometrica*, 55, 819–847.
Newey, W.K. and K. West (1988) "Hypothesis Testing with Efficient Method of Moments Estimation", *International Economic Review*, 28, 777–787.
Newey, W.K., F. Hsieh and J. Robins (1992) "Bias Corrected Semiparametric Estimation", mimeo, MIT.
Olsen, R.J. (1978) "Note on the Uniqueness of the Maximum Likelihood Estimator for the Tobit Model", *Econometrica*, 46, 1211–1216.
Pagan, A.R. (1984) "Econometric Issues in the Analysis of Regressions with Generated Regressors", *International Economic Review*, 25, 221–247.
Pagan, A.R. (1986) "Two Stage and Related Estimators and Their Applications", *Review of Economic Studies*, 53, 517–538.
Pakes, A. (1986) "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks", *Econometrica*, 54, 755–785.
Pakes, A. and D. Pollard (1989) "Simulation and the Asymptotics of Optimization Estimators", *Econometrica*, 57, 1027–1057.
Pierce, D.A. (1982) "The Asymptotic Effect of Substituting Estimators for Parameters in Certain Types of Statistics", *Annals of Statistics*, 10, 475–478.
Pollard, D. (1985) "New Ways to Prove Central Limit Theorems", *Econometric Theory*, 1, 295–314.
Pollard, D. (1989) *Empirical Processes: Theory and Applications*, CBMS/NSF Regional Conference Series Lecture Notes.
Powell, J.L. (1984) "Least Absolute Deviations Estimation for the Censored Regression Model", *Journal of Econometrics*, 25, 303–325.
Powell, J.L. (1986) "Symmetrically Trimmed Least Squares Estimation for Tobit Models", *Econometrica*, 54, 1435–1460.
Powell, J.L., J.H. Stock and T.M. Stoker (1989) "Semiparametric Estimation of Index Coefficients", *Econometrica*, 57, 1403–1430.
Pratt, J.W. (1981) "Concavity of the Log Likelihood", *Journal of the American Statistical Association*, 76, 103–106.
Press, W.H., B.P. Flannery, S.A. Tenkolsky and W.T. Vetterling (1986) *Numerical Recipes*, Cambridge University Press.

Pringle, R. and A. Rayner (1971) *Generalized Inverse Matrices*, London: Griffin.
Robins, J. (1991) "Estimation with Missing Data", preprint, Epidemiology Department, Harvard School of Public Health.
Robinson, P.M. (1988a) "The Stochastic Difference Between Econometric Statistics", *Econometrica*, 56, 531–548.
Robinson, P. (1988b) "Root-N-Consistent Semiparametric Regression", *Econometrica*, 56, 931–954.
Rockafellar, T. (1970) *Convex Analysis*, Princeton: Princeton University Press.
Roehrig, C.S. (1989) "Conditions for Identification in Nonparametric and Parametric Models", *Econometrica*, 56, 433–447.
Rothenberg, T.J. (1971) "Identification in Parametric Models", *Econometrica*, 39, 577–592.
Rothenberg, T. J. (1973) *Efficient Estimation with a priori Information*, Cowles Foundation Monograph 23, New Haven: Yale University Press.
Rothenberg, T.J. (1984) "Approximating the Distributions of Econometric Estimators and Test Statistics", Ch. 15 in: Z. Griliches and M.D. Intriligator, eds., *Handbook of Econometrics, Vol 2*, Amsterdam, North-Holland.
Rudin, W. (1976) *Principles of Mathematical Analysis*, New York: McGraw-Hill.
Sargan, J.D. (1959) "The Estimation of Relationships with Autocorrelated Residuals by the Use of Instrumental Variables", *Journal of the Royal Statistical Society Series B*, 21, 91–105.
Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics*, New York: Wiley.
Stoker, T. (1991) "Smoothing Bias in the Measurement of Marginal Effects", MIT Sloan School Working Paper, WP3377-91-ESA.
Stone, C. (1975) "Adaptive Maximum Likelihood Estimators of a Location Parameter", *Annals of Statistics*, 3, 267–284.
Tauchen, G.E. (1985) "Diagnostic Testing and Evaluation of Maximum Likelihood Models", *Journal of Econometrics*, 30, 415–443.
Van der Vaart, A. (1991) "On Differentiable Functionals", *Annals of Statistics*, 19, 178–204.
Wald (1949) "Note on the Consistency of the Maximum Likelihood Estimate", *Annals of Mathematical Statistcs*, 20, 595–601.
White, H. (1980) "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity", *Econometrica*, 48, 817–838.
White, H. (1982a) "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, 50, 1–25.
White, H. (1982b) "Consequences and Detection of Misspecified Linear Regression Models", *Journal of the American Statistical Association*, 76, 419–433.