Statistical Inference Based on Extremum Estimators

1 Introduction

Suppose θ^0, the true value of a p-dimensional parameter, is known to lie in some subset $\mathcal{S} \subset \mathbb{R}^p$. Often we choose to estimate θ^0 by minimizing (over $\theta \in \mathcal{S}$) some objective function $Q_n \equiv Q(\theta, y)$, where the random vector y represents the data from a sample of size n. Some examples are:

1. *Linear least squares regression,* where $Q_n = (y - X\theta)'(y - X\theta)$ and $X\theta^0$ is the conditional expectation of the random vector y given the regressor matrix X.

2. *Nonlinear least squares* (NLS), where $Q_n = [y - g(\theta, X)]'[y - g(\theta, X)]$, $g(\theta^0, X)$ is the conditional expectation of y given X, and g has a known functional form.

3. *Least absolute deviation* (LAD) regression where $Q_n = \sum_i |y_i - x_i'\theta|$ and the conditional median of y_i given x_i is equal to $x_i'\theta^0$.

4. *Maximum likelihood* (ML), where Q_n is minus the log of the joint probability density (or mass) function of the data in a fully specified parametric model.

5. *Generalized method of moments* (GMM), where $Q_n = m(\theta, y)'Wm(\theta, y)$, $m(\theta, y)$ is a q-dimensional vector such that plim $m(\theta^0, y) = 0$, and W is a $q \times q$ positive definite matrix.

2 Asymptotic Properties of Extremum Estimators

Clearly, not every function Q_n will lead to good estimates. We shall consider functions that have the following property: when n is large, $Q(\theta, y)/n$ (viewed as a function of θ) is with high probability very close to a nonrandom function $Q^*(\theta)$ that has a pronounced minimum at θ^0. Then, in large samples, it seems plausible that $\hat{\theta}$, the minimizer of $Q(\theta, y)$, should with high probability be close to θ^0, the minimizer of $Q^*(\theta)$. This intuition is made precise in the following consistency theorem:

> Suppose $Q(\theta,y)$ is a continuous function of θ in the compact parameter set $S \subset R^p$. Then, if $n^{-1}Q(\theta,y)$ converges in probability uniformly to a function which has a unique minimum at θ^0, the value $\hat{\theta}$ that minimizes $Q(\theta,y)$ converges in probability to θ^0.

Regularity assumptions on the exogenous variables and weak dependence across trials often imply that these conditions for consistency hold.

> EXAMPLE: In the linear model $y = X\theta^0 + u$, suppose the u's are i.i.d. with mean zero, variance σ^2 and independent of X. Assume further that $n^{-1}X'X$ converges in probability to a positive definite matrix B. For the least-squares criterion function, we have
>
> $$\frac{Q_n}{n} = \frac{(y - X\theta)'(y - X\theta)}{n} = \frac{u'u}{n} - \frac{2(\theta - \theta^0)X'u}{n} + \frac{(\theta - \theta^0)'X'X(\theta - \theta^0)}{n}$$
>
> But $n^{-1}u'u \xrightarrow{p} \sigma^2$ and $n^{-1}X'u \xrightarrow{p} 0$. Thus plim $n^{-1}Q_n = (\theta - \theta^0)'B(\theta - \theta^0)$ where the convergence is uniform in any compact set of parameter values. Since B is positive definite, this limiting function has a unique minimum at $\theta = \theta^0$.

Given consistency and employing linearization methods, we can often show that extremum estimators are approximately normal when the sample size is large. Suppose $Q(\theta, y)$ is twice differentiable in θ with first derivative vector $S_n(\theta)$ and continuous second derivative matrix $H_n(\theta)$. (S_n is often called the "score" and H_n the "hessian" for Q_n. Of course both will also depend on the sample data y, but this dependency will be surpressed to simplify the notation.) If $\hat{\theta}$ is a regular interior minimum of Q_n , then $S_n(\hat{\theta}) = 0$ and $H_n(\hat{\theta})$ is positive definite. Using the mean value theorem, we can write

$$0 = n^{-1/2} S_n(\hat{\theta}) = n^{-1/2} S_n(\theta^0) + n^{-1} H_n^* \sqrt{n}(\hat{\theta} - \theta^0)$$

where H_n^* is the hessian H_n with each element evaluated at a θ value somewhere between $\hat{\theta}$ and θ^0. Suppose that $n^{-1/2} S_n(\theta^0)$ converges in distribution to a normal random vector having mean zero and $p \times p$ covariance matrix A. Then, if $n^{-1} H_n^*$ converges in probability to a $p \times p$ positive definite matrix B, it follows that $\sqrt{n}(\hat{\theta} - \theta^0)$ has the same limiting distribution as $-B^{-1} n^{-1/2} S_n(\theta^0)$ and hence

$$\sqrt{n}(\hat{\theta} - \theta^0) \xrightarrow{d} N(0, B^{-1} A B^{-1}). \tag{1}$$

In large samples, one might then act as though $\hat{\theta}$ were normal with mean θ^0 and variance matrix $n^{-1} \widehat{B}^{-1} \widehat{A} \widehat{B}^{-1}$, where $\widehat{A}$ and $\widehat{B}$ are consistent estimates of A and B.

To prove that (1) is valid, we must show that $n^{-1/2} S_n(\theta^0)$ tends to a normal random variable and that $n^{-1} H_n^*$ tends to a nonrandom, full-rank matrix. In many problems $S_n(\theta^0)$ is the sum of n independent (or weakly dependent) mean-zero random variables; standard central limit theorems can then be employed. Demonstrating the convergence of $n^{-1} H_n^*$ is usually more difficult since we typically do not have a closed form expression for H_n^*. However, continuity arguments and the law of large numbers often can be employed to show that $n^{-1} H_n^*$ converges in probability to $B = \lim \mathrm{E} n^{-1} H_n(\theta^0)$. Rigorous proofs of the asymptotic normality of extremal estimators will not be attempted here. The LAD case where Q is nondifferentiable is particularly tricky since the linearization no longer can be obtained by the mean value theorem. In contrast, the OLS case where Q is quadratic is simple since then H_n does not depend on θ.

3 Computing Extremum Estimates

In practice, the computational problem of finding the minimum of $Q(\theta, y)$ is nontrivial. Computer intensive iterative methods are often successful. If $Q(\theta, y)$ is twice differentiable in θ, the quadratic Taylor series approximation around a point $\theta = a^0$ is given by

$$Q(\theta, y) \approx Q(a^0, y) + (a^0 - \theta)' S^0 + \frac{1}{2}(a^0 - \theta)' H_0 (a^0 - \theta)$$

where S^0 is the gradient of Q_n and H_0 is the matrix of second derivatives, both evaluated at $\theta = a^0$. The Newton-Raphson algorithm for minimizing Q_n starts with a trial value a^0 and, if H_0 is positive definite, minimizes the quadratic approximation obtaining $a^1 = a^0 - H_0^{-1} S^0$. [If H_0 is not definite, one can replace it with $H + cI$ for some small number c.] The next step is to evaluate S and H at $\theta = a^1$ and repeat the calculation. Using the recursion $a^{r+1} = a^r - H_r^{-1} S^r$, one continues until $a^{r+1} \approx a^r$. Then, as long as H_r is positive definite, one uses a^r as the estimate $\hat{\theta}$ since $S(a^r) \approx 0$. Of course, this yields at best only a *local* minimum; one must try various starting values a^0 to be confident that one has truly minimized Q_n.

The Gauss-Newton algorithm is a variant of Newton-Raphson for the special case where Q_n can be written as $e'e/2$ and the elements of the n-dimensional vector e are nonlinear functions of θ. Defining Z to be the $n \times p$ matrix $\partial e / \partial \theta'$, the gradient vector S can be

written as $Z'e$. Moreover, the hessian H is equal to $Z'Z$ plus a term that tends to be smaller. The G-N algorithm drops this smaller term and uses the recursion $a^{r+1} = a^r - (Z_r'Z_r)^{-1}Z_r'e_r$, where Z and e are evaluated at the previous estimate a^r. Note that the G-N algorithm can be implemented by a sequence of least squares regressions.

4 Two-step Estimators

We often encounter problems where $Q_n(\theta)$ is difficult to minimize because some elements of θ enter in a complicated way. In those cases, a two-step estimation procedure is often employed. Suppose the parameter vector is partitioned into two parts (θ_1, θ_2) and that θ_1 enters into Q in a simple way so that, if θ_2 were known, Q could easily be minimized over θ_1.That is, defining the gradient of Q with respect to θ_1 by S_1 , we assume that the equation $S_1(\theta_1, \theta_2) = 0$ can easily by solved as $\theta_1 = h(\theta_2, y)$. If one could find a simple estimate say $\tilde{\theta}_2$, one might estimate θ_1 by $\widetilde{\theta}_1 = h(\widetilde{\theta}_2, y)$. That is, we would estimate θ_1 by minimizing $Q(\theta_1, \widetilde{\theta}_2)$. If both the (computationally difficult) true extremal estimator $\widehat{\theta}$ and the simple estimator $\widetilde{\theta}_2$ are consistent and jointly asymptotically normal, then it can be shown that the estimator $\widetilde{\theta}_1$ is also consistent and asymptotically normal. Its asymptotic variance will typicaly depend on the asymptotic variance of $\widetilde{\theta}_2$ and can be computed by linearizing the first-order condition $S_1(\widetilde{\theta}_1, \widetilde{\theta}_2) = 0$. If, for example,

$$n^{-1/2}S_1(\widetilde{\theta}_1, \widetilde{\theta}_2) = n^{-1/2}S_1(\theta_1^0, \theta_2^0) + B_{11}\sqrt{n}(\widetilde{\theta}_1 - \theta_1^0) + B_{12}\sqrt{n}(\widetilde{\theta}_2 - \theta_2^0) + o_p(1)$$

then the asymptotic variance of $\widetilde{\theta}_1$ can be calculated from

$$\sqrt{n}(\widetilde{\theta}_1 - \theta_1^0) = -B_{11}^{-1}[n^{-1/2}S_1(\theta_1^0, \theta_2^0) + B_{12}\sqrt{n}(\widetilde{\theta}_2 - \theta_2^0)] + o_p(1)$$

as long as one knows the joint limiting distribution of $n^{-1/2}S_1(\theta_1^0, \theta_2^0)$ and $\sqrt{n}(\widetilde{\theta}_2 - \theta_2^0)$.

5 Asymptotic Tests Based on Extremum Estimators

Consider the null hypothesis that the true parameter value θ^0 satisfies the equation $g(\theta^0) = 0$, where g is a vector of q smooth functions with continuous $q \times p$ Jacobian matrix $G(\theta) = \partial g/\partial\theta$. For notational convenience we define $G \equiv G(\theta^0)$. Suppose we have estimated θ by minimizing some objective function $Q(\theta, y)$ as in section 2 and that the standardized estimator $\sqrt{n}(\hat{\theta} - \theta^0)$ is asymptotically $N(0, V)$ where $V = B^{-1}AB^{-1}$. (A and B are defined in that section.) Then, using the delta method, we find

$$\sqrt{n}[g(\widehat{\theta}) - g(\theta^0)] \approx G\sqrt{n}(\widehat{\theta} - \theta^0) \xrightarrow{d} N(0, GVG').$$

Under the null hypothesis, $ng(\hat{\theta})'(GVG')^{-1}g(\hat{\theta})$ is asymptotically $\chi^2(q)$. Replacing G with the consistent estimate $\widehat{G} \equiv G(\widehat{\theta})$ does not affect the asymptotic distribution. Hence, for some consistent estimate $\widehat{V}$, we might reject the hypothesis that $g(\theta^0) = 0$ if the *Wald statistic*

$$W \equiv ng(\widehat{\theta})'(\widehat{G}\widehat{V}\widehat{G}')^{-1}g(\widehat{\theta}) \tag{2}$$

is larger than the 95% quantile of a $\chi^2(q)$ distribution.

Sometimes solving for $\widehat{\theta}$ is computationally difficult and one seeks a way to test the hypothesis that $g(\theta^0) = 0$ that avoids this computation. Suppose there exists an easy-to-calculate estimate $\tilde{\theta}$ that satisfies $g(\tilde{\theta}) = 0$. Suppose further that, *when the null hypothesis is true,* $\tilde{\theta}$ is consistent, asymptotically normal and the usual linear approximations are valid:

$$\sqrt{n}[g(\hat{\theta}) - g(\tilde{\theta})] = G\sqrt{n}(\hat{\theta} - \tilde{\theta}) + o_p(1)$$

$$n^{-1/2}[S_n(\hat{\theta}) - S_n(\tilde{\theta})] = B\sqrt{n}(\hat{\theta} - \tilde{\theta}) + o_p(1).$$

Assuming B is a continuous function of θ^0, define $\tilde{G} = G(\tilde{\theta})$ and $\tilde{B} = B(\tilde{\theta})$. Since $g(\tilde{\theta})$ and $S_n(\hat{\theta})$ are both zero vectors, we see that W is asymptotically equivalent under the null hypothesis to

$$n(\widehat{\theta} - \tilde{\theta})'\widehat{G}'(\widehat{G}\widehat{V}\widehat{G}')^{-1}\widehat{G}(\widehat{\theta} - \tilde{\theta}) \tag{3}$$

and to

$$n^{-1}S_n(\tilde{\theta})'\tilde{B}^{-1}\tilde{G}'[\tilde{G}\tilde{V}\tilde{G}']^{-1}\tilde{G}\tilde{B}^{-1}S_n(\tilde{\theta}). \tag{4}$$

Note that (4) can be computed without solving the minimization problem.

In the linear regression model where $Q_n = \frac{1}{2}(y - X\theta)'(y - X\theta)$, we find $H = X'X$ and $\text{var}[S_n(\theta)] = \sigma^2 X'X$. Suppose the null hypothesis is the linear constraint $G\theta^0 = 0$. Estimating A by $s^2X'X/n$ and B by $X'X/n$, we obtain

$$W = \widehat{\theta}'G'[G(X'X)^{-1}G']^{-1}G\widehat{\theta}/s^2$$

which is q times the usual F-statistic. This feature of LS regression (that the variance matrix for the score is proportional to the expectation of the hessian) holds in many estimation problems. When it occurs, not only do (2), (3) and (4) simplify but also some additional asymptotically equivalent expressions for the Wald statistic are available.

Consider the problem of minimizing $Q(\theta, y)$ subject to the constraint $g(\theta) = 0$. The solution can be used for our $\tilde{\theta}$. The first order condition for an interior minimum is $S_n(\tilde{\theta}) = \tilde{G}'\lambda$ which implies

$$\lambda = (\tilde{G}B^{-1}\tilde{G}')^{-1}\tilde{G}'B^{-1}S_n(\tilde{\theta}) \quad \text{and} \quad S_n(\tilde{\theta}) = \tilde{G}'(\tilde{G}B^{-1}\tilde{G}')^{-1}\tilde{G}'B^{-1}S_n(\tilde{\theta}).$$

If $A = cB$ and $\tilde{\theta}$ is the constrained minimizer of Q_n, the test statistics (3) and (4) simplify to

$$n(\hat{\theta} - \tilde{\theta})'\widehat{B}(\hat{\theta} - \tilde{\theta})/\hat{c} \tag{3'}$$

and

$$n^{-1}S_n(\tilde{\theta})'\tilde{B}^{-1}S_n(\tilde{\theta})/\hat{c}\ . \tag{4'}$$

Furthermore, a Taylor's series expansion of the statistic

$$2[Q_n(\hat{\theta}) - Q_n(\tilde{\theta})]/\hat{c} \tag{5}$$

shows that it too is asymptotically equivalent to (3') and hence to W.

Thus, if $A = cB$ for some nonzero scalar c, plim $\hat{c} = c$, and $\tilde{\theta}$ is the constrained minimizer of Q_n, we have four aymptotically equivalent statistics that might be used for testing $g(\theta^0) = 0$:

(a) a quadratic form in $g(\widehat{\theta})$:
$$ng(\widehat{\theta})'(\widehat{G}\widehat{B}^{-1}\widehat{G}')^{-1}g(\widehat{\theta})/\hat{c}$$

(b) a quadratic form in the estimator difference $\tilde{\theta} - \widehat{\theta}$
$$n(\widehat{\theta} - \widetilde{\theta})'\widehat{B}(\widehat{\theta} - \widetilde{\theta})/\hat{c}$$

(c) a quadratic form in the score (which is the same as a quadratic form in the Lagrange multiplier λ)
$$n^{-1}S(\widetilde{\theta})'\widetilde{B}^{-1}S(\widetilde{\theta})/\hat{c} \equiv n^{-1}\lambda'\widetilde{G}\widetilde{B}^{-1}\widetilde{G}'\lambda/\hat{c}$$

(d) the difference in constrained and unconstrained minimized objective function (multiplied by $2/\hat{c}$).
$$2[Q(\widetilde{\theta}, y) - Q(\widehat{\theta}, y)]/\hat{c}.$$

Although our discussion has concerned only the null distribution of these tests, the asymptotic equivalence holds also under nearby alternatives. Exact equivalence occurs in the linear regression case because there H and G are nonrandom and do not depend on the unknown θ^0. In the general case where H may be random and both may depend on θ^0, we find only asymptotic equivalence.

There are many different ways to consistently estimate B, including the hessians $n^{-1}H_n(\widehat{\theta})$ and $n^{-1}H_n(\tilde{\theta})$ as well as the analytic expression $n^{-1}EH_n(\theta)$ evaluated at $\widehat{\theta}$ or $\tilde{\theta}$. Thus there are really lots of asymptotically equivalent test statistics available.

Computational convenience is often used as a basis for choice among asymptotically equivalent tests. However, if p and q are large, the asymptotic approximations are sometimes poor. It is usually wise to perform some simulations to verify that the chosen test statistic has approximately the correct small sample rejection probability under the null hypothesis.

When Q is minus the log likelihood function we find that $A = B$ since A and B are then alternative expressions for the limiting information matrix; (d) is then the likelihood ratio statistic and (c) is the score statistic. When the correct weighting matrix is used in the quadratic form defining GMM, we also have A proportional to B and a choice of tests.

6 Score Tests as Diagnostics

The following type of problem often occurs in econometrics. We postulate a fairly simple probability model for the data but entertain the possibility that a more complicated model may be needed. Let $Q_n(\theta_1, \theta_2)$ be the objective function assuming the complicated model is correct; θ_1 is the p-dimensional parameter vector in the simple model and θ_2 is the q-dimensional vector of additional parameters need to cope with the complication. The simple model being correct is equivalent to the parametric hypothesis that $\theta_2 = 0$. Thus, before publishing estimates of θ_1 based on the simple model, one might want to check that θ_2 is really close to zero. A Wald or likelihood ratio test would require actually estimating the complicated model. The score test does not and is therefore commonly used as a diagnostic.

Assuming that $A = B$, the score test statistic for $\theta_2 = 0$ is $n^{-1}S(\widetilde{\theta})'\widetilde{B}^{-1}S(\widetilde{\theta})$, where $\widetilde{\theta}$ minimizes Q_n subject to the constraint. The score vector S can be partitioned into two subvectors, say S_1 and S_2 . But, when evaluated at the constrained estimate $\widetilde{\theta}$, S_1 must be zero (since that is the first-order condition for minimizing Q subject to $\theta_2 = 0$.) Thus the test statistic can be written as $n^{-1}S_2(\widetilde{\theta})'CS_2(\widetilde{\theta})$, where C is the $q \times q$ lower right hand block of B^{-1} . Using $n^{-1}H(\widetilde{\theta})$ as an estimate of B and partitioning conformably, a natural estimate of C is $n(\widetilde{H}_{22} - \widetilde{H}_{21}\widetilde{H}_{11}^{-1}\widetilde{H}_{12})^{-1}$. In other words, to test the hypothesis that the simple model is valid: first, compute $\widetilde{\theta}_1$, the MLE for the parameters of the simple model; second, evaluate the score S_2 at $\widetilde{\theta} = (\widetilde{\theta}_1, 0)$; third, compute the test statistic $S_2(\widetilde{\theta})'(\widetilde{H}_{22} - \widetilde{H}_{21}\widetilde{H}_{11}^{-1}\widetilde{H}_{12})^{-1}S_2(\widetilde{\theta})$.

In many examples, the information matrix is block diagonal implying plim $n^{-1}H_{12} = 0$; then the test statistic can be simplified to $S_2(\widetilde{\theta})'\widetilde{H}_{22}^{-1}S_2(\widetilde{\theta})$. In other words, when the information matrix is block diagonal, one can test $\theta_2 = 0$ by pretending that θ_1 were known and equal to the estimate $\widetilde{\theta}_1$.