10/12/83
correc. 10/13/83

MAXIMUM LIKELIHOOD STEP SIZE AND THE ONE STEP THEOREM

Whitney K. Newey*
Princeton University

Econometric Research Program
Research Memorandum No. 308

October 1983

We investigate the statistical properties of an estimator which is obtained on one Newton-Raphson step from an initial consistent estimator using a stepsize which maximizes the likelihood function. We show that the maximum likelihood stepsize has a probability limit of one and, consequently that the resulting estimator is asymptotically equivalent to the maximum likelihood estimator. This result is useful in hypothesis testing applications and as a crude diagnostic check of the validity of a one-step procedure.

*Discussions with Daniel McFadden led to this note and Jerry Hausman has provided useful comments.

I. Introduction

In spite of recent strong advances in computing technology, the use of one Newton-Raphson (or Gauss-Newton) iteration starting from an initial consistent estimator can still provide a useful alternative to full maximum likelihood in large problems or problems where the likelihood function is not well behaved.[1] It is well known that when the step size on this iteration is chosen to be one the resulting estimator is asymptotically equivalent to the maximum likelihood estimator (MLE).[1a] An alternative choice of step size is to choose the step size which maximizes the likelihood in a one dimensional search. For example, if the initial estimator is a maximum likelihood estimator subject to parameter restrictions, and one iteration is used to obtain an unrestricted estimator, the use of the maximum likelihood step size can guarantee that the resulting likelihood ratio statistic is positive.[2] We show that, under appropriate regularity conditions, the maximum likelihood step size converges in probability to one, so that the resulting estimator is asymptotically equivalent to the maximum likelihood estimator.

II. The Maximum Likelihood Step Size

Let θ be a $q \times 1$ vector of parameters, and z_t be a $1 \times p$ vector of random variables, $t=1, \ldots, T$. Let $L(\theta) = \frac{1}{T} \ln f(z_1, \ldots, z_T|\theta)$ be the normalized log-likelihood function. We will assume that $L(\theta)$ is twice continuously differentiable and will denote the $q \times 1$ gradient vector by $L_\theta(\theta)$ and the Hessian matrix by $L_{\theta\theta}(\theta)$. The following assumption specifies the data generating process.

Assumption 1: For each T the random vector $z_1, \ldots, z_T$ has probability density function $f(z_1, \ldots, z_T|\theta_o)$, where θ_o lies in the interior of some compact set S.

Consider an estimator $\tilde{\theta}$ of θ_o. We are interested in the properties of an estimator $\bar{\theta}$ obtained from the iteration.

$$\bar{\theta} = \tilde{\theta} + \lambda_T D_T L_\theta(\tilde{\theta}) \tag{1}$$

where D_T is a $q \times q$ matrix and λ_T is a suitably chosen step size. In order to obtain the asymptotic properties we make the following assumptions concerning $L(\theta)$ and $\tilde{\theta}$.

Assumption 2: For any θ_T such that plim $\theta_T = \theta_o$,

$$\text{plim } L_{\theta\theta}(\theta_T) = -J \tag{2}$$

where J is positive definite.

Assumption 3: The maximum likelihood estimator $\hat{\theta}$ satisfies $L(\hat{\theta}) = \max_{\theta \text{ in } S} L(\theta)$ and

$$\text{plim } \hat{\theta} = \theta_o \tag{3}$$

where θ_o is contained in the interior of the compact set S.

Assumption 4: For some $a > o$, $\tilde{\theta}$ satisfies

$$T^a(\hat{\theta} - \tilde{\theta}) = o_p(1) , \tag{4}$$

$$\lim_{b \to o} \lim_{T \to \infty} \sup P(T^{2a}(\hat{\theta} - \tilde{\theta})'(\hat{\theta} - \tilde{\theta}) < b) = 0 \tag{5}$$

where b is restricted to be greater than zero. The matrix J of Assumption 2 is the information matrix, and this Assumption specifies that $L_{\theta\theta}(\Theta)$ has a uniform convergence property. Assumption 3 specifies that the MLE is consistent for θ_o. For simplicity we do not give fundamental regularity conditions which imply that Assumptions 2 and 3 are satisfied. Assumption 5 states that $T^a(\hat{\theta} - \tilde{\theta})$ is bounded in probability and does not have a limiting point mass at zero. This assumption will be satisfied for $a = 1/2$ if both $\sqrt{T}\ (\hat{\theta} - \theta_o)$ and $\sqrt{T}\ (\tilde{\theta} - \theta_o)$ have limiting normal distributions and $\tilde{\theta}$ is not asymptotically equivalent to $\hat{\theta}$. A case where Assumption 5 might be satisfied with, say, $a = 3/2$ is if $\tilde{\theta}$ is itself obtained from an iteration like equation (1), so that $\tilde{\theta}$ is first-order equivalent to $\hat{\theta}$ but not second order.

To obtain our result we make the following assumption concerning D_T and the step size λ_T.

Assumption 5: The matrix D_T satisfies

$$\text{plim } D_T = J^{-1} \tag{6}$$

and the step size λ_T is obtained from solving

$$\max_{\underline{0}\leq\lambda\leq\underline{A}} L(\tilde{\theta} + \lambda D_T L_\theta(\tilde{\theta})) \tag{7}$$

where $A > 1$.

Equation (6) will be satisfied if $D_T = -L_{\theta\theta}(\tilde{\theta})^{-1}$, as for Newton-Raphson, or if D_T is the outer product estimator presented in Berndt, Hall, Hall, and Hausman (1974).

Theorem 1: If Assumptions 1-5 are satisfied then plim $\lambda_T = 1$ and plim $T^a(\hat{\theta} - \overline{\theta}) = 0$.

Proof: Measurability of λ_T follows from Jennrich (1969), Lemma 2 Let $\dot{\theta} = \tilde{\theta} + D_T L_\theta(\tilde{\theta})$. We will work with sequences which are tail equivalent to $\hat{\theta}_T$, $\tilde{\theta}_T$, $\overline{\theta}$, and $\dot{\theta}$ without new notation. A mean value expansion of $L_\theta(\tilde{\theta})$ around $\hat{\theta}$ yields

$$\begin{aligned} T^a(\dot{\theta}-\hat{\theta}) &= T^a(\tilde{\theta}-\hat{\theta}) + D_T L_{\theta\theta}(\ddot{\theta}) T^a(\tilde{\theta}-\hat{\theta}) \\ &= [I + D_T L_{\theta\theta}(\ddot{\theta})] T^a(\tilde{\theta}-\hat{\theta}) = o_p(1). \end{aligned} \tag{8}$$

where $\ddot{\theta}$ lies between $\hat{\theta}$ and $\tilde{\theta}$, so that by Assumptions 3 and 4 plim $\ddot{\theta} = \theta_o$ and the third equality follows by Assumptions 2, 4, and 5.

Expanding $L(\dot{\theta})$ we find that

$$\begin{aligned} T^{2a}[L(\dot{\theta})-L(\hat{\theta})] &= \frac{1}{2} T^a(\dot{\theta}-\hat{\theta})' L_{\theta\theta}(\ddot{\theta}) T^a(\dot{\theta}-\hat{\theta}) \\ &= o_p(1) o_p(1) o_p(1) = o_p(1) . \end{aligned} \tag{9}$$

where $\ddot{\theta}$ lies between $\dot{\theta}$ and $\hat{\theta}$ and the second equality follows from equation (8) and Assumption 2.

Using equation (9) and expanding $L(\bar{\theta})$ around $L(\dot{\theta})$ in λ, and using the fact that $\hat{\theta}$ is the MLE,

$$0 \leq T^{2a}[L(\hat{\theta}) - L(\bar{\theta})] = T^{2a}[L(\dot{\theta}) - L(\bar{\theta})] + o_p(1) \tag{10}$$
$$= T^{2\alpha} L_\theta(\bar{\theta})' D_T L_\theta(\tilde{\theta})(1 - \lambda_T)$$
$$+ \frac{1}{2}(1 - \lambda_T)^2 T^{2a} L_\theta(\tilde{\theta})' D_T L_{\theta\theta}(\ddot{\theta}) D_T L_\theta(\tilde{\theta}) + o_p(1).$$

where $\ddot{\theta} = \tilde{\theta} + \dot{\lambda}_T D_T L_\theta(\tilde{\theta})$ for $\dot{\lambda}_T$ and 1. By Assumption 5 and J^{-1} positive definite $\lambda_T > 0$ with probability approaching one as T grows (see Berndt, Hall, Hall, and Hausman (1974) Gradient Theorem). By the first order condition for λ_T and $A>1$ it follows that

$$L_\theta(\bar{\theta})' D_T L_\theta(\tilde{\theta})(1 - \lambda_T) \leq 0 . \tag{11}$$

From equations (10) and (11) we obtain

$$0 \leq (1 - \lambda_T)^2 T^a L_\theta(\tilde{\theta})' D_T L_{\theta\theta}(\ddot{\theta}) D_T L_\theta(\tilde{\theta}) + o_p(1) \tag{12}$$
$$= (1-\lambda_T)^2 T^a(\tilde{\theta} - \hat{\theta})'(-J^{-1}) T^a(\tilde{\theta} - \hat{\theta}) + o_p(1)$$

where the last equality follows by $\dot{\lambda}_T$ bounded, so that by Assumptions 3, 5 and equation 8, plim $\ddot{\theta} = \theta_o + o_p(1)\text{plim}(\dot{\theta} - \tilde{\theta}) = \theta_o$, and by

$$T^a(\tilde{\theta} - \hat{\theta}) = T^a(\tilde{\theta} - \dot{\theta}) + o_p(1) = O_p(1).$$

Equation (5) of Assumption 4 and Equation (12) imply that if λ_T does not converge in probability to one equation (12) will be violated. Let $\varepsilon, \delta > 0$ and (T_n) be a subsequence such that for each T_n,

$$P((\lambda_{T_n} - 1)^2 > \varepsilon) > \delta \tag{13}$$

Equation (5) and J positive definite imply that there exists T^1 and $\eta > 0$ such that $T > T^1$ implies

$$P(T^{2a}(\tilde{\theta} - \hat{\theta})'(-J^{-1})(\tilde{\theta} - \hat{\theta}) < -\eta) > 1 - \delta/2 \tag{14}$$

so that for all $T_n > T^1$

$$P(T^{2a}(\lambda_{T_n} - 1)^2(\tilde{\theta} - \hat{\theta})'(-J^{-1})(\tilde{\theta} - \hat{\theta}) < -\eta\varepsilon) > \delta/2 \tag{15}$$

violating equation (12).

From equation (1) and the definition of $\dot{\theta}$ it follows that

$$\begin{aligned} T^a(\hat{\theta} - \bar{\theta}) &= T^a(\hat{\theta} - \dot{\theta}) + (1 - \lambda_T)D_T T^a L_\theta(\tilde{\theta}) \\ &= o_p(1) + (1 - \lambda_T)D_T o_p(1) = o_p(1) + o_p(1)o_p(1)o_p(1) \\ &= o_p(1) , \end{aligned} \tag{16}$$

where the second equality follows from equation (8) and the third equality from plim $\lambda_T = 1$ and Assumption 5.

3. Discussion

The fact that the maximum likelihood stepsize converges in probability to one can be used as a rough diagnostic check of the validity of a one-step

estimator. If the likelihood is correctly specified and the sample size is large then λ_T should be close to one, so that if the maximum likelihood step-size is far from one the sample may be undersized or the likelihood is misspecified.[3]

One of the applications of Theorem 1 is to situations where it is useful to guarantee that the iteration of equation (1) results in an increase in the likelihood function, such as when testing hypotheses. Of course, it seems reasonable to expect that a stepsize of one should result in an increase in the likelihood in large samples. We verify that a stepsize of one will result in an increase in the likelihood function in the following result.

Theorem 2: If Assumptions 1-5 are satisfied then, for $\dot{\theta} = \tilde{\theta} + D_T L_\theta(\tilde{\theta})$, $L(\dot{\theta}) > L(\tilde{\theta})$ with probability approaching one as the sample size grows.

Proof: Expanding $L(\tilde{\theta})$ in a mean value expansion around $\dot{\theta}$

$$T^{2a}[L(\dot{\theta}) - L(\tilde{\theta})] = -T^a L_\theta(\dot{\theta})'[T^a(\tilde{\theta} - \dot{\theta})] - T^{2a}\frac{1}{2}(\tilde{\theta} - \dot{\theta})'L_{\theta\theta}(\ddot{\theta})(\tilde{\theta}-\dot{\theta}) \quad (17)$$

where $\ddot{\theta}$ lies between $\tilde{\theta}$ and $\dot{\theta}$, so that plim $\ddot{\theta} = \theta_o$. Expanding $L_\theta(\dot{\theta})$ around $\hat{\theta}$

$$T^a L_\theta(\dot{\theta}) = L_{\theta\theta}(\ddot{\theta})T^a(\dot{\theta} - \hat{\theta}) = o_p(1)o_p(1) = o_p(1) \quad (18)$$

where the second equality follows from Assumption 2. Then by equations (17) and (18) and $T^a(\dot{\theta} - \hat{\theta}) = o_p(1)$,

$$T^{2a}(L(\dot{\theta}) - L(\tilde{\theta})) = -T^{a}L_{\theta}(\dot{\theta})'[T^{a}(\tilde{\theta} - \hat{\theta})] \tag{19}$$

$$- \frac{1}{2} T^{a}(\tilde{\theta} - \hat{\theta})'L_{\theta\theta}(\ddot{\theta})(\tilde{\theta} - \hat{\theta})T^{a} + o_{p}(1)$$

$$= o_{p}(1)o_{p}(1) + \frac{1}{2}T^{a}(\tilde{\theta} - \hat{\theta}) \quad J \quad T^{a}(\tilde{\theta} - \hat{\theta}) + o_{p}(1)$$

$$= \frac{1}{2} T^{a}(\tilde{\theta} - \hat{\theta})' \; J \; T^{a}(\tilde{\theta} - \hat{\theta}) + o_{p}(1).$$

The conclusion follows by positive definiteness of J and Assumption 4, as in the proof of Theorem 1. Thus, both λ_{T} and whether or not the likelihood function increases can be used as rough diagnostic checks for one-step procedures.

Finally, it is straightforward to show that our results apply to situations where criterion functions other than the likelihood function are used to obtain estimators. For example stepsize obtained minimizing a quadratic form in moment functions would yield on estimator which is asymptotically equivalent to a one-step generalized method of moments estimator (see Newey (1983) Theorem 3.2).

Footnotes

1. An important example where there is often a large number of parameters is the nested logit model of McFadden (1981).

1a. See, for example, Zacks (1971), pp. 250-251.

2. See, for example, Berkovec, Hausman, and Rust (1983).

3. A format test based on the asymptotic distribution $\sqrt{T}\,(\lambda_T - 1)$ would be would be difficult to obtain. Some preliminary calculations, which are not reported here, indicate that this asymptotic distribution would involve third derivatives of the likelihood and the reciprocal of a quadratic form of asymptotically normal random variables.

References

Berndt, E., B. Hall, R. Hall, and J. Hausman, 1974, Estimation and Inference in Nonlinear Structural Models, Annals of Economic and Social Measurement 3, pp. 653-666.

Jennrich, R. I., 1969, Asymptotic Properties of Non-linear Least Squares Estimators, Annals of Mathematical Statistics 40, pp. 633-643.

McFadden, D., 1981, Econometric Models of Probabilistic Choice, in: C. Manski and D. McFadden, eds., Structural Analysis of Discrete Data (MIT Press, Cambridge).

Newey, W., 1983, Generalized Method of Moments Specification Testing, Princeton University, Econometric Research Program Memorandum No. 306.

Zacks, S., 1971, The Theory of Statistical Inference (Wiley, New York).