COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 710

STABILITY COMPARISONS OF ESTIMATORS

Donald W.K. Andrews

July, 1984

STABILITY COMPARISONS OF ESTIMATORS

Donald W. K. Andrews

Cowles Foundation for Research in Economics
Yale University

July, 1984

ABSTRACT

A property of estimators called stability is investigated in this paper. The stability of an estimator is a measure of the magnitude of the affect of any single observation in the sample on the realized value of the estimator. High stability often is desirable for robustness against misspecification and against highly variable observations.

Stabilities are determined and compared for a wide variety of estimators and econometric models. Estimators considered include: least squares, maximum likelihood (including both LIML and FIML), instrumental variables, M-, and multi-stage estimators such as two and three stage least squares, Zellner's feasible Aikten estimator of the multivariate regression model, and Heckman's estimator of censored regression and self-selection models. The general results of the paper apply to numerous additional estimators of various and sundry models.

The stability of an estimator is found to depend on the number of finite moments of its influence curve (evaluated at a random observation in the sample). An estimator's stability increases strictly and continuously from zero to one as the number of finite moments of its influence curve increases from one to infinity. The more moments, the higher the stability. Since it often is possible to construct estimators with a specified influence function, estimators with different stabilities can be constructed. For example, one can attain the maximum stability possible by formulating a bounded influence estimator, since they have an infinite number of finite moments.

1. Introduction and Conclusion

This paper investigates a property of estimators called stability. The stability of an estimator is a measure of the magnitude of the affect of any single observation in the sample on the realized value of the estimator. A number of reasons related to robustness suggest that often it is desirable for an estimator to be relatively insensitive to any particular observation in the sample, i.e., to have high stability. But, whether or not high stability is desirable for a given situation, it is useful for diagnostic purposes to have knowledge of the stabilities of different estimators in order to know which estimators are likely to rely more heavily on some single observation.

In words, the stability of an estimator is the greatest normalization factor such that the normalized deviation of the estimator, due to the deletion of a single observation, converges to zero with probability one as the sample size goes to infinity, for any sequence of deletions. That is, the stability of an estimator $\hat{\theta}_n$ of an R^J-valued parameter θ_0 is defined as

$$(1.1) \quad \Lambda(\hat{\theta}_n, P_{\theta_0}) = \sup\{\xi \in R : n^{\xi}(\hat{\theta}_n - \hat{\theta}_{n,k_n}) \xrightarrow{n\to\infty} \underset{\sim}{0} \text{ a.s. } [P_{\theta_0}], \forall\{k_n\}\} ,$$

where $\hat{\theta}_{n,k_n}$ is the estimator applied to the sample of size n with the k_n^{th} observation deleted, $\underset{\sim}{0}$ is a J-vector of zeros, a.s. abbreviates almost surely, P_{θ_0} is the underlying probability distribution generating the data, and $\{k_n\} \equiv \{k_n : k_n \leq n, n = 1, 2, \dots\}$ is a sequence of indices of deleted observations, one for each sample size. Thus, the stability of an estimator is an asymptotic measure of the sensitivity of the estimator to observations actually in the sample (rather than to non-random hypothetical

observations, as is measured by the influence curve, see Hampel (1974)). Under fairly general conditions, stability values lie between zero and one, with the extremes being attained by certain estimators. The results of this paper concern the determination of the stability of estimators in a fairly broad class, and for an extensive array of different econometric models. Models for which the results apply include among others: location (univariate and multivariate), location and scale, linear and nonlinear regression (with fixed or random regressors), linear and nonlinear simultaneous equations, panel data, and limited dependent variable (such as logit, probit, truncated and censored regression, and self-selection).

The class of estimators considered in this paper is defined to include all estimators which can be written as solutions (for θ) to a system of equations:

$$\sum_{i=1}^{n} r_i(Z_i, \theta) = \underset{\sim}{0} , \tag{1.2}$$

where $r_i(\cdot,\cdot)$ is a specified function which defines the estimator, and Z_i is a random vector of observed variables comprising the i^{th} observation. Note, Z_i may include variables in Z_ℓ for $\ell < i$. For example, in time series regression and simultaneous equations models, Z_i may include lagged variables. The number of estimators which can be written in the form (1.2) is quite large. For example, the following estimators are included: least squares, maximum likelihood (including full-information (FIML) and limited information (LIML) estimators of simultaneous equations models), instrumental variables, M-, and various multi-stage estimators such as Zellner's (1962) seemingly unrelated regression estimator, Heckman's (1979) estimator of censored regression and self-selection models, two stage

least squares (2SLS), and three stage least squares (3SLS). These examples are discussed below in Section 3.

Under suitable regularity conditions (outlined below), it is possible to write estimators in the class defined above in a linearized form:

$$\hat{\theta}_n = \theta_0 - I_n \frac{1}{n} \sum_{i=1}^{n} A^{-1} r_i(Z_i, \theta_0) \,, \tag{1.3}$$

where θ_0 is the true parameter vector, I_n is a $J \times J$ random matrix equal to the identity matrix plus a matrix of small order one as $n \to \infty$ a.s., and A is a $J \times J$ non-random non-singular matrix. If $r_i(\cdot,\cdot)$ is independent of i for i sufficiently large, then $A^{-1}r(z, \theta_0)$ is the influence curve of $\hat{\theta}_n$ evaluated at z, as defined by Hampel (1974). It is shown that the stability of $\hat{\theta}_n$ is directly related to the number of moments of $r_i(Z_i, \theta_0)$, $i = 1, 2, \ldots$. In particular, if r_U and r_L are stochastically greater than or equal to, and less than or equal to $r_i(Z_i, \theta_0)$ for all $i = 1, 2, \ldots$, respectively, then the stability of $\hat{\theta}_n$ lies in the interval $[1 - 1/p, 1 - 1/q]$, where r_U and r_L have p and q finite moments, respectively. If p equals q, the stability of $\hat{\theta}_n$ is established. Otherwise, the stability of $\hat{\theta}_n$ is given by a more complicated expression involving the tail probabilities of the random vectors $r_i(Z_i, \theta_0)$, $i = 1, 2, \ldots$. Thus, the qualitative result is obtained that the stability of an estimator depends on the number of moments of its linearized form (or influence curve) --the greater the number of moments, the greater the stability. Further, there is no upper bound beyond which additional moments no longer increase the stability of the estimator. Since $r_i(\cdot,\cdot)$ is chosen by the investigator, it is often straightforward to obtain estimators with specified linearized form. Hence, estimators with different stabilities can be constructed.

It should be noted that stability results depend on the number of

finite moments of the linearized estimator, not on the number of finite moments of the estimator itself. The latter has received considerable attention in the econometric literature, e.g., see Kinal (1980), since common estimators of simultaneous equations models have fewer than all moments finite even with normal errors. These results have no clear implications for stability since they deal with moments of the estimator rather than of the linearized form.

The examples of Section 3 provide a variety of models, estimators, and stability characteristics of different estimators. We briefly summarize the results here: In the linear regression model with fixed regressors, the least squares (LS) estimator has stability which depends on the number of finite moments of the errors. On the other hand, Huber (1973) M-estimators have the maximum stability of one in this model, regardless of the distribution of the errors. In the linear regression model with random regressors, the LS estimator has stability which depends on the number of finite moments of the errors and the regressors, whichever is smaller. In contrast, Krasker and Welsch's (1982) bounded influence regression estimator has stability equal to one for all error and regressor distributions. Results for the LS estimator and M-estimators in the nonlinear regression model parallel those in the linear model, except the dependence on the number of finite moments of the regressors, when applicable, is replaced by that of the derivative of the regression function (with respect to the parameter vector) evaluated at the true parameter.

The instrumental variable (IV) estimator of a single equation from a system of linear equations has stability which depends on the number of finite moments of the errors and the instruments. In comparison, Krasker and Welsch's (1983) weighted instrumental variable (WIV) estimator for this

model has a bounded influence function, and hence, has stability equal to one--the maximum--regardless of the distribution of the errors and instruments.

The stability of maximum likelihood (ML) and pseudo-ML estimators depends on the number of finite moments of their score functions. In logit and probit models, this corresponds to the number of finite moments of the regressors. In the censored regression model, it corresponds to the number of finite moments of the errors and regressors. Heckman's (1979) two-stage estimator of this model has the same stability properties as the ML estimator. Similarly, the ML estimator and Zellner's (1962) feasible Aitken estimator for the seemingly unrelated (i.e., multivariate) nonlinear regression model have the same stability properties. Their stability depends on the number of finite moments of the errors and the derivatives of the regression functions (with respect to the parameter vector) evaluated at the true parameter. Following the examples of Section 3, the calculation of stabilities of other estimators in other models is straightforward.

Clearly, if $r_i(Z_i, \theta_0)$, $i = 1, 2, \ldots$ are uniformly bounded, then all of their moments exist and the maximum stability is attained. Bounded influence estimators, referred to above, are characterized by this property. In contrast, other estimators have stability which depends on the true underlying probability distribution, since the true distribution determines the number of finite moments of $r_i(Z_i, \theta_0)$, $i = 1, 2, \ldots$. This is illustrated by the examples of Section 3.

As mentioned above, several reasons related to robustness suggest that high stability is often a desirable property for estimators. We now discuss these reasons. First, economic data are rarely so "clean" that it is prudent to put great weight on a single observation. For example,

the imprecisions of economic data are manifested by the continuous revisions made to macroeconomic time series, and the subjective nature of some microeconomic survey data. Several factors contribute to this imprecision: There is pure measurement error at the data collection stage. The correspondence between observed or "constructed" variables and the variables which are relevant from the perspective of economic theory is usually imperfect, and sometimes considerably so. The precise definitions of variables may be problematic even from a theoretical perspective, as exemplified by the money supply and a market share (in a nebulous market). Finally, recording errors made in stages of data collection, transmission, and analysis are inevitable. Such errors are often beyond the control of the econometrician who may have no input into the collection and transmission stages. In fact, the econometrician may have only scant knowledge of the degree of imprecision of the data. In such cases, it is unwise to let any single observation have great weight in determining an estimator's value.

The imprecision of econometric models themselves also adds to the desirability of high stability. Economic theory cannot yield complete model specifications, so even in the presence of a simple true model, a specified model is likely to be just an approximation. Moreover, the existence of simple true models is questionable. To be tractable and useful, econometric models must be simple. However, in most cases such models are at best approximations of a much more complicated socio-economic phenomena. An observation which appears to be highly informative may be so only because of a spuriously precise specification of the model. For example, in a linear regression model an observation which is an outlier in the space of regressor variables can be highly informative. That is, it can greatly reduce estimator variances. However, if it is recognized that the extension of the

regression function to the outlying observation may be nonlinear with unknown functional form, then the informative content of the observation is drastically reduced. In such a case, the effect of the observation on the computed variance of an estimator with low stability is spurious and deceptive. Such an observation also can cause a significant bias for an estimator with low stability. An estimator with higher stability is more robust to such specification difficulties because no single observation is given excessive weight.

A third reason for interest in high stability is that, in some models, estimators which are highly sensitive to a single observation perform quite poorly even if the model is specified correctly and the variables are measured without error. This arises when the observations are highly variable. In this case, any single observation is potentially a randomly generated outlier with little informative content, and hence, should not be given disproportionate weight. For example, in a regression model or simultaneous equations model with fat-tailed errors, the least squares (LS) estimator has low stability because an outlying error realization can dramatically alter the value of the estimator. As expected, the relative efficiency of the LS estimator is quite poor in this situation. On the other hand, various robust procedures have high stability, and consequently, perform quite well even with highly variable observations. The statistical literature on robustness has analyzed problems of this sort in some detail, see Huber (1981).

The above arguments for high stability are not always applicable, of course, and so, estimators with high stability are not always preferable. However, for diagnostic purposes it still may be useful to know which estimation procedures are more likely to weight some single observation heavily. Hence, even in this case, estimator stability is of interest. Note, stability

comparisons can be made between different estimators for the same model or between estimators of different models. If an econometrician is more familair with one model than another, stability comparisons of the latter sort may yield useful qualitative information about the second estimator's sensitivity to single observations in the sample based on knowledge of the first estimator's sensitivity.

The stability measure is based on the deviations $\hat{\theta}_n - \hat{\theta}_{n,k}$, $k = 1, \ldots, n$. In the literature these deviations have been found useful for other related purposes. In analyzing the behavior of the least squares estimator in the linear regression model, Cook (1977, 1979) and Belesley, Kuh, and Welsch (1980) use these deviations to help detect influential observations. Also, these deviations are proportional to the deviations of an estimator from its jackknifed pseudo-values. Tukey (1958) has suggested a nonparametric estimator of the variance of the original estimator $\hat{\theta}_n$, based on the latter deviations (see also Miller (1974)). The relationship between stability and the influence curve, a very important tool of robust statistics, has been mentioned above. A finite sample analogue of the influence curve suggested by Tukey (1970), viz., the sensitivity curve, is also related to stability. If we denote the sensitivity curve of $\hat{\theta}_n$ formed using all n observations except the k^{th} by $SC_{n,k}(z)$, then $SC_{n,k}(z)$ evaluated at the deleted observation Z_k is proportional to the deviation $\hat{\theta}_n - \hat{\theta}_{n,k}$. That is, Tukey's finite sample sensitivity curve (constructed with an observation deleted) evaluated at points in the actual sample is the basis of the stability measure. Finally, a different approach to some issues related to stability is given by O'brien's (1975) analysis of the sensitivity of the least squares estimator in the linear regression model to random perturbations in the data.

This paper is organized as follows: Section 1 introduces the basic idea contained in the paper, attempts to motivate it, and summarizes the results in an informal manner. Section 2 presents definitions, assumptions, and the general results. For purposes of illustration, the linear regression model with the least squares estimator is used as a running example throughout this section. Section 3 discusses numerous additional applications of the general results. Section 4 contains proofs of the results given in Section 2.

2. General Results

2.1. Model and Estimator Assumptions

The general model considered in this section is described by an infinite sequence $\{Z_i\} \equiv \{Z_i : i = 1, 2, \ldots\}$ of random vectors of arbitrary dimensions. A sample of size n corresponds to the observation of the first n terms in this sequence. For increased generality, the i^{th} term Z_i is allowed to include elements of the random vectors Z_ℓ, for $\ell < i$. Thus, Z_i may include lagged variables. The distribution of the sequence $\{Z_i\}$, denoted P_{θ_0}, depends on an R^J-valued parameter θ_0. All probabilistic statements below are made for $\{Z_i\}$ distributed according to P_{θ_0}. Thus, "almost surely" means "almost surely under P_{θ_0}." The sequence $\{Z_i\}$ is assumed to be weakly dependent over time. That is, the dependence between random vectors dies out as the difference in subscripts of the variables becomes infinitely large. (For the case of cross-sectional data, the observations are often independent and this requirement is satisfied.) More precisely, $\{Z_i\}$ is assumed to be strong mixing. This is a realistic assumption for many economic time-series (and

cross-section) situations. It is considerably weaker than other assumptions, such as independence, m-dependence, or auto-regressive moving average (ARMA) structure (see Chanda (1974), but cf. Andrews (1984b)), which are often used in econometric models. Moreover, strong mixing does not imply stationarity or any assumption related to identical distributions.

Strong mixing is defined as follows: Let $\{Q_i : i = 1, 2, \ldots\}$ be a sequence of random vectors. Let $\mathcal{B}_{i\ell}$ denote the σ-field generated by $Q_i, Q_{i+1}, \ldots, Q_\ell$ for $1 \leq i < \ell \leq \infty$. That is, $\mathcal{B}_{i\ell}$ is the collection of all events determined by $Q_i, Q_{i+1}, \ldots, Q_\ell$. $\{Q_i\}$ is strong mixing if $\alpha(n) \to 0$ as $n \to \infty$, where $\alpha(n)$ are the strong mixing numbers of $\{Q_i\}$ defined by

$$(2.1) \qquad \alpha(n) \equiv \sup_{\ell \geq 1} \sup_{A \in \mathcal{B}_{1,\ell}, B \in \mathcal{B}_{\ell+n,\infty}} |P(A \cap B) - P(A)P(B)| .$$

Note, if $\{Q_i\}$ are independent, then $\alpha(n) = 0$, $\forall n \geq 1$; if $\{Q_i\}$ are m-dependent, then $\alpha(n) = 0$, $\forall n > m$; and if $\{Q_i\}$ have ARMA structure with absolutely continuous innovations, then $\alpha(n)$ declines to zero at an exponential rate as $n \to \infty$ (see Chanda (1974)). We assume:

A1) $\{Z_i\}$ are strong mixing with strong mixing numbers $\alpha(n)$ which satisfy $\alpha(n) = o(n^{-\alpha/(\alpha-1)})$ as $n \to \infty$, for some $\alpha \geq 1$ (where $\alpha = 1$ requires $\alpha(n) \equiv 0$ for n sufficiently large).

We now turn to two simple examples which we carry through this section to illustrate the more general model and results. Section 3 discusses other applications of the results of this section. The first example considered here is the classical linear regression (CLR) model,

$$(2.2) \qquad y_i = x_i'\theta_0 + u_i , \quad i = 1, 2, \ldots, n ,$$

where y_i is the observed dependent variable, x_i is the observed R^J-vector

of fixed regressors, u_i is an independent, identically distributed (i.i.d.), mean zero, unobserved error, and θ_0 is an R^J-valued unknown parameter vector. In this case, $Z_i = (y_i, x_i')'$. We suppose that the regressors are uniformly bounded, and that $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i x_i' = H$, for some non-singular $J \times J$ matrix H. The second example we consider is the random regressor linear regression (RRLR) model. This model is identical to the CLR model except the regressors are assumed to be random, not fixed. We assume the regressors are i.i.d. and independent of the errors, and $Ex_i x_i' = H$ is non-singular. Clearly, A1 is satisfied in both of these models with $\alpha = 1$. Note, the rather restrictive assumptions placed on these models are for purposes of exposition; the general results given below allow them to be relaxed considerably.

The class of estimators considered for the general model includes all estimators which can be written as (measurable) solutions for θ to a system of equations of the form

$$\sum_{i=1}^{n} r_i(Z_i, \theta) = \underset{\sim}{0}, \tag{2.3}$$

for some R^J-valued (measurable) functions $r_i(\cdot,\cdot)$, $i = 1, 2, \ldots$, which are defined on some neighborhood of the true parameter θ_0. For notational convenience we abbreviate $r_i(Z_i, \theta)$ by $r_i(\theta)$. The j^{th} element of $r_i(\theta)$ is denoted $r_{ij}(\theta)$. Section 3 shows that many well-known estimators of econometric models can be written as such. For the two examples of this section, we consider the least squares (LS) estimator. For this estimator,

$$r_i(Z_i, \theta) = (y_i - x_i'\theta)x_i \equiv r_i^{LS}(\theta). \tag{2.4}$$

Results concerning the stability of an estimator $\hat{\theta}_n$ are of interest only if the estimator satisfies certain minimal conditions regarding its performance. Hence, the following assumption is not particularly restrictive:

B1a) $\{r_i(\theta)\}$ is sufficiently well-defined that a (measurable) solution $\hat{\theta}_n$ to (2.3) exists (though is not necessarily unique) for n sufficiently large a.s., and $\hat{\theta}_n \xrightarrow{n\to\infty} \theta_0$ a.s.

b) Further, $\hat{\theta}_{n,k_n} \xrightarrow{n\to\infty} \theta_0$ a.s., for any sequence of positive integers $\{k_n\}$ with $k_n \leq n$, $\forall n$.

Conditions which imply strong consistency of the estimator $\hat{\theta}_n$ usually also imply strong consistency of $\hat{\theta}_{n,k_n}$, the estimator which ignores the $(k_n)^{th}$ observation. Most estimators considered in econometrics satisfy these conditions under certain assumptions on the underlying model.[1,2] Such assumptions can be found in the literature. In particular, the LS estimator for the CLR and RRLR models satisfy B1, see Lai, Robbins, and Wei (1978), Anderson and Taylor (1979), and White (1980).

One of the assumptions usually needed for consistency of an estimator defined as a solution to (2.3) is that the expectation of the defining equations is zero or approaches zero as the sample size increases. We shall make this assumption explicit:

$$\text{B2)} \quad n^{\nu-1} \sum_{i=1}^{n} E r_i(\theta_0) \xrightarrow{n\to\infty} \underset{\sim}{0}, \quad \forall \nu < 1 - 1/(2 \wedge (p/\alpha)),$$

where " $\wedge$ " is the minimum operator. In the CLR and RRLS models $E r_i^{LS}(\theta_0) \equiv \underset{\sim}{0}$, so B2 is satisfied.

We now state several definitions used below. A random variable (rv) X is said to have g finite moments if

$$(2.5) \qquad E|X|^{g+\delta} \begin{cases} < \infty & \text{for all } \delta < 0 \\ = \infty & \text{for all } \delta > 0 \, . \end{cases}$$

If $E|X|^{\delta} = \infty$ $(< \infty)$ for all $\delta > 0$, X is said to have 0 (∞) finite moments. Thus, every rv has a unique number g of finite moments and $g \in [0,\infty]$. For examples, a normal rv has ∞ finite moments, and a t rv with d degrees of freedom has d finite moments. The number of finite moments of a random vector or matrix is defined to be the smallest number of finite moments of any of its elements.

For a random vector or matrix X, let $|X|$ denote X with all of its elements replaced by their absolute values, and $\|X\|$ denote the Euclidean norm of X.

A rv X is said to be stochastically less (greater) than or equal to a rv Y, and we write $X \overset{ST}{\leq} Y$ $(X \overset{ST}{\geq} Y)$, if $F_X(x) \geq F_Y(x)$ $(F_X(x) \leq F_Y(x))$, $\forall x \in R$, where F_X and F_Y are the distribution functions (df's) of X and Y, respectively. The same term is applied to random vectors and matrices if the above condition is satisfied element by element.

Now we construct a random vector, r_U, which is stochastically greater than or equal to $r_i(\theta_0)$ for all i. Let $F_U(w)$ be a J-vector with j^{th} element given by $\inf_{i \geq 1} P(|r_{ij}(\theta_0)| \leq w)$, for $j = 1, \ldots, J$ and $w \in R$. Let r_U be a random J-vector whose elements have univariate df's given by the vector $F_U(w)$. Denote the number of finite moments of r_U by p. The following assumption requires that $r_i(\theta_0)$ for $i = 1, 2, \ldots$ are stochastically dominated by an $L^{2\alpha-1}$ random vector (where α is a measure of the dependence of the sequence $\{Z_i\}$, see A1):

B3) $E|r_U|^{2\alpha-1} < \underset{\sim}{\infty}$, where $\underset{\sim}{\infty}$ is a J-vector of infinities. Equivalently, $p > 2\alpha-1$.

(Note, B3 rules out the case where some element of r_U is point mass at infinity.) In general, if B3 does not hold then either $\hat{\theta}_n$ is strongly consistent but it is somewhat more difficult to prove (e.g., see Hannan and Kanter (1977)), or $\hat{\theta}_n$ is not strongly consistent (as exemplified by the LS estimator when the errors in the CLR or RRLR model have infinite mean). In consequence, B3, or conditions which imply B3, is a common assumption in the literature (e.g., see assumptions 3 and 5 of Burguete, Gallant, and Souza (1982, pp. 162 and 167)). The LS estimator in the CLR and RRLR models satisfies B3 since $r_U^{LS} \overset{ST}{\leq} |u_1| \cdot \sup_{i \geq 1} |x_i|$ where $E|u_1| < \infty$ in the CLR model, and $r_U^{LS} \overset{ST}{\leq} |u_1 \cdot x_1|$ where $E|u_1 \cdot x_1| < \infty$ in the RRLR model. Note, since p is not necessarily greater than or equal to 2, $\hat{\theta}_n$ is not necessarily asymptotically normal.

Next we construct a random matrix, Dr, which is stochastically greater than or equal to $\left|\frac{\partial}{\partial\theta} r_i(\theta_0)\right|$ for all i. Let $F_{Dr}(w)$, $w \in R$, be a $J \times J$ matrix with $(\ell,j)^{th}$ element $\inf_{i \geq 1} P\left(\left|\frac{\partial}{\partial\theta_\ell} r_{ij}(\theta_0)\right| \leq w\right)$, for $\ell, j = 1, \ldots, J$. Let Dr be a $J \times J$ random matrix whose elements have univariate df's given by the matrix $F_{Dr}(w)$. Dr is used to state a uniform smoothness condition on $r_i(\theta)$ at θ_0. We assume:

B4a) $A \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E\frac{\partial}{\partial\theta} r_i(\theta_0)$ exists and is non-singular.

b) $E\|Dr\|^\eta < \infty$, for some η satisfying $\eta \geq 2$ and $\eta > \alpha$.

(Note, the assumption $\eta \geq 2$ can be relaxed in the results that follow.) Assumption B4a is common in the literature (e.g., see assumption 6 of Burguete, Gallant, and Souza (1982, p. 169)) because it is usually necessary for asymptotic normality (with a non-singular covariance matrix). The

estimators considered here are not necessarily asymptotically normal, but this particular assumption is still used. For the LS estimator in the CLR and RRLR models, B4a corresponds to the assumptions above that $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$ and $Ex_i x_i'$ exist and are non-singular, respectively. B4b holds in the CLR model since the x_i are uniformly bounded, and in the RRLR model if $E(x_1'x_1)^2 < \infty$.

The result $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} r_i(\hat{\theta}_n) \xrightarrow{n\to\infty} A$ a.s. is commonly used in the literature when showing asymptotic normality of an estimator $\hat{\theta}_n$. We also use this result, and impose the following additional smoothness condition on $r_i(\theta)$ to ensure that it holds:[3]

B5a) $\sup_{i\geq 1} EW_{ij}^{\alpha+\delta} < \infty$, for some $\delta > 0$, for $j = 1, \ldots, J$, where $W_{ij} \equiv \sup_{\tilde{\theta}\in\Theta_0} \left\| \frac{\partial^2}{\partial\theta\partial\theta'}(r_{ij}(\tilde{\theta}) - r_{ij}(\theta_0)) \right\|$, Θ_0 is some neighbourhood of θ_0; and

b) $\frac{1}{n}\sum_{i=1}^{n} \left\| \frac{\partial^2}{\partial\theta\partial\theta'} r_{ij}(\theta_0) \right\| = O(1)$ as $n \to \infty$, a.s., $\forall j = 1, \ldots, J$.

For the LS estimator in the CLR and RRLR models B5 is automatically satisfied, since $\frac{\partial^2}{\partial\theta\partial\theta'} r_i(\theta) \equiv \underset{\approx}{0}$, where $\underset{\approx}{0}$ is a matrix of zeros.

2.2. Stability Results

First, we present a result which gives a linearized form of the estimator $\hat{\theta}_n$. It also shows that the smoothness conditions on $r_i(\theta)$ are sufficient to yield strong consistency of $\hat{\theta}_n$ and $\hat{\theta}_{n,k_n}$ at a faster rate of convergence than n^0.

Theorem 1. Under assumptions A1 and B1-B5,

(a) $\hat{\theta}_n = \theta_0 - I_n \frac{1}{n} \sum_{i=1}^{n} A^{-1} r_i(\theta_0)$, where I_n is a $J \times J$ random matrix equal to the identity matrix plus a matrix of small order one as $n \to \infty$ a.s., and

(b) for all sequences of positive integers $\{k_n\}$ with $k_n \leq n$, $\lim_{n\to\infty} n^{\nu}(\hat{\theta}_{n,k_n} - \theta_0) = \underset{\sim}{0}$ a.s., $\forall \nu < 1 - 1/(2 \wedge (p/\alpha))$.

Comments: 1. The linearized form of $\hat{\theta}_n$, viz., $\theta_0 - \frac{1}{n} \sum_{i=1}^{n} A^{-1} r_i(\theta_0)$, highlights the importance of the rv's $r_i(\theta_0)$, $i = 1, \ldots, n$, in determining the stochastic properties of the estimator $\hat{\theta}_n$. In particular, the linearized form suggests that the stability of $\hat{\theta}_n$ may be related to the tail behavior of $r_i(\theta_0)$, $i = 1, \ldots, n$. It is shown below that this is the case.

2. If $r_i(\cdot,\cdot)$ is independent of i for i sufficiently large, as is often the case, then the influence curve of $\hat{\theta}_n$ is $A^{-1} r(z, \theta_0)$. Thus the linearized form of $\hat{\theta}_n$ is determined by its influence curve.

3. Part b of the Theorem shows that the rate of convergence of $\hat{\theta}_n$ to θ_0 depends on the number of moments of $r_i(\theta_0)$, $i = 1, \ldots, n$ (as measured by the number p of finite moments of the stochastically dominating random vector r_U). In addition, there is a tradeoff between the number of moments of r_U and the degree of dependence over time (as indexed by α , see A1). Note, the dependence of the rate of convergence, ν , on the number of moments of r_U , p , and the degree of dependence, α , only exists below a cut off point. If $p \geq 2\alpha$, then the maximal rate of convergence is obtained, and additional moments are of no consequence. This contrasts with the results obtained below for the stability of $\hat{\theta}_n$. In the latter case no such cut off point exists.

4. In the CLR and RRLR models, the linearized form of the LS estimator is $\theta_0 - \frac{1}{n}\sum_{i=1}^{n} H^{-1}u_i x_i$, α equals one, and p equals the number of finite moments of u_i and $u_i x_i$, respectively. In both models, if u_i has two or more moments, the maximal rate of convergence is obtained, i.e., the upper bound on ν is one-half.

5. The proof (see Section 4) makes use of McLeish's (1975) three series theorem for strong mixing rv's, and a result of Loeve (1955).

We now establish two lower bounds on the stability of an estimator $\hat{\theta}_n$:

Theorem 2. Let A1 and B1-B5 hold. Then,

(a) $\Lambda(\hat{\theta}_n, P_{\theta_0}) \geq 1 - 1/p$, and

(b) $\Lambda(\hat{\theta}_n, P_{\theta_0}) \geq \sup\{\xi \in R : \sum_{n=1}^{\infty} [1 - F^*_{nj}(n^{1-\xi})] < \infty, \forall j = 1, \ldots, J\}$ $(\geq 1 - 1/p)$,

where $F^*_{nj}(x) \equiv \min_{i \leq n} F_{ij}(x)$, and $F_{ij}(\cdot)$ is the df of $r_{ij}(\theta_0)$.

Comments: 1. The lower bound of part a is more readily interpretable than that of part b, but part b is a stronger result. That is, the lower bound of part b is greater than or equal to that of part a.

2. The lower bound of part a is a linear function of the reciprocal of the number p of finite moments of r_U . The lower bound increases strictly and continuously from 0 to 1 as p increases from 1 to ∞ . This result differs from rate of convergence results for strong consistency (see Theorem 1). The latter exhibit a cut off point beyond which additional moments do not increase the rate of convergence.

3. For the LS estimator in the CLR model, p equals the number of finite moments of the error u_1 . For example, if u_1 has t-distribution with d degrees of freedom, then the lower bound given by part a is

$1 - 1/d$, and it ranges continuously from 0 for the Cauchy $(d = 1)$ to 1 for the normal $(d = \infty)$. With regard to part b of Theorem 2, $F^*_{nj}(n^{1-\xi}) = F_{u_1}(n^{1-\xi}/(\max_{i\leq n}|x_{ij}|))$ in this case, where $F_{u_1}(\cdot)$ is the df of u_1 . Note, $\sum_{n=1}^{\infty}[1 - F_{u_1}(n^{1-\xi}/(\max_{i\leq n}|x_{ij}|))] < \infty$, $\forall j$, if and only if $\sum_{i=1}^{n}[1 - F_{u_1}(n^{1-\xi})] < \infty$. And,

$$(2.6)\quad \sum_{n=1}^{\infty}[1 - F_{u_1}(n^{1-\xi})] = \sum_{n=1}^{\infty} P(|u_1|^{1/(1-\xi)} > n) \in [E|u_1|^{1/(1-\xi)},\ E|u_1|^{1/(1-\xi)} + 1],$$

using Loeve's (1955, p. 242) moments inequality. Thus, in this case, the lower bound of part b reduces to $1 - 1/p$, as in part a.

4. In the RRLR model, p equals the number of finite moments of $u_1 \cdot x_1$. If x_1 has as many or more moments than u_1 , then the situation is exactly as above in the CLR model. However, if x_i has fewer moments than u_1 , then the variability of the regressors determines the value of p and the lower bound $1 - 1/p$ is less than in the CLR model (with the same error distribution). For the RRLR model, $F^*_{nj}(n^{1-\xi}) = F_{nj}(n^{1-\xi})$, $\forall j$, and an argument similar to that of comment 3 shows that the lower bound of part b reduces to $1 - 1/p$.

5. The condition $\eta \geq 2$ in assumption B4b can be relaxed in this Theorem. Specifically, (a) under the assumptions of Theorem 2 except that of $\eta \geq 2$, for any $\tilde{p} \in (2\alpha-1, p]$, if $\eta \geq 2 \wedge (\tilde{p}/\alpha)$, then $\Lambda(\hat{\theta}_n, P_{\theta_0}) \geq 1 - 1/\tilde{p}$, and (b) under the assumptions of Theorem 2 except those of $\eta \geq 2$ and $p > 2\alpha-1$ (see B3), for any $\tilde{p} > 0$, if $\eta \geq 2 \wedge (\tilde{p}/\alpha)$ and $p > 1$, then $\Lambda(\hat{\theta}_n, P_{\theta_0}) \geq \sup\{\xi \in R : \sum_{n=1}^{\infty}[1 - F^*_{nj}(n^{1-\xi})] < \infty$ and $\xi < 2(1 - 1/(2 \wedge (\tilde{p}/\alpha)))\}$.

6. The proof makes use of a Taylor expansion of $\sum_{i=1}^{n} r_i(\hat{\theta}_n)$, the first Borel-Cantelli Lemma, a moment inequality of Loeve (1955), and Theorem 1b (to show various terms are $o(1)$ as $n \to \infty$ a.s.).

The next result provides an upper bound on the stability of an estimator $\hat{\theta}_n$ in terms related to the number of moments of $r_i(\theta_0)$, $i = 1, \ldots, n$. Further, it shows that the stability of $\hat{\theta}_n$ actually equals the lower bound of Theorem 2 part b. This result requires a stronger condition on the asymptotic weak dependence of the process $\{Z_i\}$ than strong mixing, because the second Borel-Cantelli Lemma is used in its proof. This Lemma is usually stated for independent sequences, but it also holds for φ-mixing processes (defined below), see Iosefescu and Theodorescu (1969, Lemma 1.1.2'). However, it has not been shown to hold, and may not hold, for strong mixing processes. (On the other hand, strong mixing processes do satisfy a related result, viz., Kolmogorov's zero-one law, see Andrews (1984a).)

A sequence of random vectors $\{Q_i\}$ is <u>φ-mixing</u> if $\varphi(n) \to 0$ as $n \to \infty$, where $\varphi(n)$ are the φ-mixing numbers of $\{Q_i\}$ defined by

$$(2.7)\quad \varphi(n) \equiv \sup_{\ell \geq 1} \sup_{A \in \mathcal{B}_{1,\ell}: P(A)>0; B \in \mathcal{B}_{\ell+n,\infty}} |P(A \cap B) - P(A)P(B)|/P(A)$$

$$= \sup_{\ell \geq 1} \sup_{A \in \mathcal{B}_{1,\ell}: P(A)>0; B \in \mathcal{B}_{\ell+n,\infty}} |P(B|A) - P(B)| ,$$

where $\mathcal{B}_{i,\ell}$ is the σ-field generated by $\{Q_i, Q_{i+1}, \ldots, Q_\ell\}$. Sequences of independent and m-dependent rv's are clearly φ-mixing. Billingsley (1968) provides additional examples. However, the φ-mixing condition is considerably stronger than the strong mixing condition. For example, stationary Gaussian sequences of rv's are φ-mixing if and only if they are m-dependent, see Ibragimov and Linnik (1971), whereas they are strong mixing under the

weak condition that they possess a continuous, positive spectral density, see Kolmogorov and Rozonov (1960). Thus, the φ-mixing assumption may be stronger than is reasonable for some economic applications, but it is difficult to avoid at present.

For the next result we assume:

A1') $\{Z_i\}$ are φ-mixing with strong mixing numbers as in A1.

For the upper bound on the stability of $\hat{\theta}_n$ given below, we need to construct a random vector, r_L, which is stochastically less than or equal to $|r_i(\theta_0)|$ for all i. Let $F_L(w)$ be the J-vector whose j^{th} element is $\sup_{i\geq 1} P(|r_{ij}(\theta_0)| \leq w)$ for $j = 1, \ldots, J$ and $w \in R$, and let r_L be a random J-vector whose elements have univariate df's given by the vector $F_L(w)$. Let q denote the number of finite moments of r_L. Note, the number of finite moments of r_U, p, is necessarily less than or equal to q.

Theorem 3. Let A1' and B1-B5 hold. Then,

(a) $\Lambda(\hat{\theta}_n, P_{\theta_0}) \leq 1 - 1/q$, provided $p > 2\alpha q/(q+1)$, and

(b) $\Lambda(\hat{\theta}_n, P_{\theta_0}) = \inf\{\xi \in R : \sum_{n=1}^{\infty} [1 - F^*_{nj}(n^{1-\xi})] = \infty$, for some integer j in $\{1,\ldots,J\}\}$

$= \sup\{\xi \in R : \sum_{n=1}^{\infty} [1 - F^*_{nj}(n^{1-\xi})] < \infty, \forall j = 1, \ldots, J\}$,

where F^*_{nj} is as in Theorem 2.

Comments: 1. Part a holds for all $q < \infty$. If $q = \infty$, part a is shown to hold (see Section 4) provided r_L is not identically $\underset{\sim}{0}$. In consequence, the right-hand-side in part b is less than or equal to one provided r_L is not identically $\underset{\sim}{0}$.

2. In some cases (e.g., when the observations are identically distributed), p equals q, and the stability of an estimator is given by the number of finite moments of the linearized form of the estimator — the more moments, the greater the stability. In particular, there is no cut off beyond which the existence of more moments is of no consequence. If p is less than q, then the stability of $\hat{\theta}_n$ lies in an interval determined by p and q, and its exact value is given by the somewhat complicated expression given in part b.

3. For the LS estimator in the RRLR model, $r_L \overset{ST}{=} |u_1 \cdot x_1| \overset{ST}{=} r_U$, and so, $q = p$ and the stability of $\hat{\theta}_n$ is $1 - 1/p$. In the CLR model, $r_L \overset{ST}{\geq} |u_1| \cdot \min_{i \geq 1} |x_1|$. If the regression function has a constant term, for example, then q is less than or equal to the number of finite moments of $|u_1|$, which is p. Hence, $q = p$ and the stability of $\hat{\theta}_n$ is $1 - 1/p$. For example, if the errors have t-distribution with d degrees of freedom, then the stability of $\hat{\theta}_n$ is $1 - 1/d$ in the CLR model. The stability of the usual estimator $\frac{1}{n-J} \sum_{i=1}^{n} (y_i - x_i' \hat{\theta}_{LS})^2$ of the error variance, σ^2, is $1 - 1/(2p)$ in the CLR model. Thus, the variance estimator is less stable than the LS estimator of the regression parameters. This corroborates results found in the literature comparing the robustness of these two estimators.

4. The condition $\eta \geq 2$ of assumption B4b can be relaxed in this Theorem. Specifically, (a) the assumptions $\eta \geq 2$ and $p > 2\alpha q/(q+1)$ can be replaced in Theorem 3 part a by $\eta \geq 2\Lambda(\tilde{p}/\alpha)$ for some $\tilde{p} \in (2\alpha q/(q+1), p]$, and (b) under the assumptions of Theorem 3 except those of $\eta \geq 2$ and $p > 2\alpha - 1$, if $\eta \geq 2\Lambda(\tilde{p}/\alpha)$ for some $\tilde{p} > 0$, then $\Lambda(\hat{\theta}_n, P_{\theta_0}) \leq \inf C$, where $C \equiv \{\xi \in R: \sum_{n=1}^{\infty} [1 - F^*_{nj}(n^{1-\xi})] = \infty$ for some j, and $\xi < 2(1 - 1/(2\Lambda(\tilde{p}/\alpha)))\}$. (Note, the infimum of a null set is defined to be infinity.)

3. Examples

This section contains a number of examples where the general results of Section 2 apply. The models and estimators are described as briefly as possible. In consequence, sufficient conditions for strong consistency (assumption B1) are not always given in their entirety. Such conditions can be found in the references cited. In all cases, the defining functions of the estimators, viz., $r_i(\theta)$, $i = 1, \ldots, n$, are assumed to be chosen to satisfy the conditions B2-B5.

It is possible to include some two and three stage estimators in the class considered in Section 2, e.g., Heckman's (1979) two stage estimator of the Tobit model--example 7, Zellner's (1962) feasible Aitken estimator for the seemingly unrelated nonlinear regression model--example 8, 2SLS, and 3SLS. Proceed as follows: Suppose part of the parameter vector θ_0 , call the part λ_0 , is estimated in a first stage via the solution to $\sum_{i=1}^{n} r_{1i}(\lambda) = \underset{\sim}{0}$, and (a not necessarily disjoint) part, call it β_0 , is estimated in a second stage via the solution to $\sum_{i=1}^{n} r_{2i}(\hat{\lambda}_n, \beta) = \underset{\sim}{0}$, where $\hat{\lambda}_n$ is the first stage estimator. In place of θ_0 consider an alternative parameter vector $\tilde{\theta}_0 = (\lambda_0', \beta_0')'$. Now, a single stage estimator, $\hat{\tilde{\theta}}_n$, of the desired form can be defined by taking

$$(3.1) \qquad r_i(\tilde{\theta}) = \begin{pmatrix} r_{1i}(\lambda) \\ r_{2i}(\lambda,\beta) \end{pmatrix}, \quad \text{for} \quad \tilde{\theta} = \begin{pmatrix} \lambda \\ \beta \end{pmatrix}.$$

This estimator satisfies A1 or A1', and B1-B5, if the separate stage estimators do. (The matrix $E\frac{\partial}{\partial\theta} r_i(\tilde{\theta}_0)$ in B4a is triangular, and hence, is non-singular if the diagonal blocks are non-singular.) Thus, the results of Section 2 apply. The extension for three stage estimators is straightforward

In the examples that follow we assume independence of the observations because this is the usual assumption made in the references cited. In most cases, this assumption can be relaxed by replacing it with an assumption of strong mixing. Strong consistency is proved, then, using the strong law of large numbers for strong mixing rv's (see McLeish (1975)).

1. Classical linear regression (CLR) model--Classical M-estimators (references: Huber (1973), Yohai and Maronna (1979)). The model is the CLR model described in Section 2. We adopt slightly different notation:

$$(3.2) \qquad y_i = x_i'\beta_0 + u_i \ , \ i = 1, \ldots, n \ , \ Z_i \equiv (y_i, x_i')' \ , \ \theta_0 = (\beta_0', \sigma_0)' \ .$$

The estimator $\hat{\theta}_n$ is defined by

$$(3.3) \qquad r_i(\theta) = \begin{pmatrix} \psi((y_i - x_i'\beta)/\sigma)x_i \\ \psi^2(|y_i - x_i'\beta|/\sigma) - c \end{pmatrix} , \quad \text{for} \quad \theta = \begin{pmatrix} \beta \\ \sigma \end{pmatrix} ,$$

where c is a given constant, ψ is a bounded, smooth, odd function, and the true parameter σ_0 solves $E\psi^2(|u_i|/\sigma_0) = c$. The estimator $\hat{\theta}_n$ has the maximum stability, one, whether or not the errors u_i have any moments. This contrasts sharply with the LS estimator, see Section 2.

2. Random regressor linear regression (RRLR) model--General M-estimators (references: Maronna and Yohai (1981), Krasker and Welsch (1982)). The model is the RRLR model described in Section 2 with the notation of example 1. The estimator $\hat{\theta}_n$ is defined by

$$(3.4) \qquad r_i(\theta) = \begin{pmatrix} \tilde{\psi}(x_i, (y_i - x_i'\beta)/\sigma)x_i \\ \chi(|y_i - x_i'\beta|/\sigma) \end{pmatrix} ,$$

where, for each x_i, $\tilde{\psi}(x_i, \cdot)$ is bounded, odd, and non-negative on R^+, $\chi(\cdot)$ is nondecreasing and bounded, $E|x_i|\sup_u|\tilde{\psi}(x_i, u)| < \infty$, and the true parameter σ_0 solves $E\chi(|u_i|/\sigma_0) = 0$. The stability of $\hat{\theta}_n$ depends on the number of finite moments of $\tilde{\psi}(x_i, u_i/\sigma_0)x_i$. If $\tilde{\psi}$ is taken such that this is bounded uniformly for x_i and u_i (as in Krasker and Welsch (1982), for example), then the general M-estimator is a bounded influence estimator, and has stability equal to one--the maximum.

3. Linear, limited information simultaneous equations model--Instrumental variable (IV) estimator (references: Sargan (1959), Heiler (1981)). The model is the same as the RRLR model but the regressors and errors are not necessarily independent:

$$(3.5) \qquad y_i = x_i'\theta_0 + u_i, \quad i = 1, \ldots, n, \quad Z_i = (y_i, x_i', w_i')',$$

where w_i is a random vector of instrumental variables which is independent of the error u_i but not of the regressor x_i. The estimator $\hat{\theta}_n$ is defined by

$$(3.6) \qquad r_i(\theta) = (y_i - x_i'\theta)w_i.$$

The stability results for the IV estimator $\hat{\theta}_n$ are the same as for the LS estimator in the RRLR model with the number of finite moments of the instruments replacing those of the regressors. In particular, if the instruments or the errors have fewer than all moments finite, the IV estimator has stability less than one.

4. Linear, limited information simultaneous equations model--Weighted instrumental variables (WIV) estimator (reference: Krasker and Welsch (1983)). The model is as in example 3 with a slight change in notation:

$$(3.7)\quad y_i = x_i'\beta_0 + u_i\ ,\quad i = 1, \ldots, n\ ,\quad Z_i \equiv (y_i, x_i', w_i')\ ,\quad \theta_0 = (\beta_0', \alpha_0')'\ .$$

The estimator $\hat{\theta}_n$ is defined by

$$(3.8)\quad r_i(\theta) = \begin{pmatrix} \min\{1,\ c/[|(y_i - x_i'\beta)/\sigma| \cdot (w_i'B^{-1}w_i)^{1/2}]\} \cdot (y_i - x_i'\beta)w_i \\ S\ \mathrm{vec}[\gamma(c^2/w_i'B^{-1}w_i) \cdot w_i w_i' - B] \end{pmatrix}\ ,\ \text{for } \theta = \begin{pmatrix}\beta\\ \alpha\end{pmatrix}\ ,$$

where c and σ given constants, the parameter vector $\alpha \equiv S\ \mathrm{vec}\ B$, S is a known $[J(J+1)/2] \times J^2$ selection matrix such that $S\ \mathrm{vec}\ B$ is the vector obtained by vectorizing the lower triangle of the symmetric $J \times J$ matrix B, $\gamma(t) \equiv E \min(\eta^2, t)$ for $\eta \sim N(0,1)$, and the true parameter vector $\alpha_0 \equiv S\ \mathrm{vec}\ B_0$ solves $B_0 = E\gamma(c^2/w_i'B_0^{-1}w_i)w_i w_i'$. (Note, σ can be estimated by adding it to the parameter vector θ and adding an element to $r_i(\theta)$.) As defined, $r_i(\theta)$ does not satisfy our conditions for smoothness in θ. However, a version of $r_i(\theta)$ which is smoothed at the corners satisfies our conditions, yet differs very little from $r_i(\theta)$. It can be seen that $r_i(\theta_0)$ is a bounded random vector. Hence, $\hat{\theta}_n$ and the WIV estimator of β_0, given by the sub-vector $\hat{\beta}_n$, have stability one.

5. Nonlinear regression model--Least squares estimator (references: Jennrich (1969), Malinvaud (1970), Bierens (1981), Wu (1981), Domowitz and White (1982)). The model is

$$(3.9)\qquad y_i = f(x_i, \beta_0) + u_i\ ,\quad i = 1, \ldots, n\ ,\quad Z_i = (y_i, x_i')'\ ,\quad \theta_0 \equiv \beta_0\ ,$$

where the errors u_i are strong mixing, mean zero rv's, the regressors x_i may be fixed or random but must satisfy conditions for "proper" behavior as $n \to \infty$ (see references), and the regression function $f(\cdot,\cdot)$ is smooth. The LS estimator $\hat{\theta}_n$ is defined by the function

$$(3.10) \qquad r_i(\theta) = (y_i - f(x_i, \theta))\frac{\partial}{\partial\theta}f(x_i, \theta) .$$

The stability of $\hat{\theta}_n$ depends on the random vectors $\left|u_i \frac{\partial}{\partial\theta}f(x_i, \theta_0)\right|$ in the manner described in Theorem 3. For examples, if the regressors are i.i.d. random vectors or are fixed and uniformly bounded, then its stability is $1 - 1/p$, where p is the number of finite moments of $\left|u_1 \frac{\partial}{\partial\theta}f(x_1, \theta_0)\right|$.

6. Nonlinear regression model--Classical M-estimators (references: Bierens (1981), Burguete, Gallant, and Souza (1982), Andrews (1983)). The model is as in example 5, except $\theta_0 \equiv (\beta_0', \sigma_0)'$ and the assumption of mean zero errors is replaced by the assumption that $E\psi(u_i/\sigma_0) = 0$, for ψ and σ_0 given below.[4] The estimator $\hat{\theta}_n = (\hat{\beta}_n', \hat{\sigma}_n)'$ is defined by the function

$$(3.11) \qquad r_i(\theta) = \begin{pmatrix} \psi((y_i - f(x_i, \beta))/\sigma)\cdot\frac{\partial}{\partial\theta}f(x_i, \beta) \\ \psi^2((y_i - f(x_i, \beta))/\sigma) - \gamma \end{pmatrix}, \quad \text{for } \theta \equiv (\beta', \sigma)' ,$$

where γ is a (known) constant given by $\gamma = \int\psi^2(s)d\Phi(s)$ for $\Phi(\cdot)$ the standard normal df, σ_0 is an unknown scale parameter defined by $E\psi^2(u_i/\sigma_0) = \gamma$, and ψ is a bounded function as in example 1. Since ψ is bounded, the stabilities of $\hat{\theta}_n$ and $\hat{\beta}_n$ depend on the vectors $\left|\frac{\partial}{\partial\theta}f(x_i, \theta_0)\right|$. If the regressors are i.i.d. random vectors, their stability is $1 - 1/p$, where p is the number of finite moments of $\left|\frac{\partial}{\partial\theta}f(x_1, \beta_0)\right|$. If the regressors are fixed and uniformly bounded, their stability is one.

7. Censored regression (or Tobit) model--Heckman's two stage estimator (reference: Heckman (1979)). The model

$$(3.12) \qquad y_i = (x_i'\beta_0 + u_i) \vee 0 \ , \quad i = 1, \ldots, n \ , \quad Z_i = (y_i, x_i')' \ ,$$

where " $\vee$ " is the maximum operator, the regressors x_i are i.i.d. random vectors, the errors u_i are independent, normal$(0, \sigma_0^2)$ rv's, and $\theta_0 \equiv (\beta_0', \sigma_0)'$. Heckman's two stage procedure uses an estimator of the form (2.3) at each stage:

1st stage: The estimator $\hat{\lambda}_n$ is a maximum likelihood (ML) probit estimator of $\lambda_0 \equiv \beta_0/\sigma_0$. Its defining function is

$$(3.13) \qquad r_{1i}(\lambda) = \frac{1_{[y_i>0]} - \Phi(x_i'\lambda)}{\Phi(x_i'\lambda)(1 - \Phi(x_i'\lambda))}\phi(x_i'\lambda)x_i \ ,$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution function, respectively, and $1_{[\cdot]}$ denotes the indicator function.

2nd stage: The estimator $(\hat{\beta}_n', \hat{\sigma}_n)'$ is the LS estimator of $(\beta_0', \sigma_0)'$ given $\hat{\lambda}_n$, using only the uncensored observations. Its defining function is

$$(3.14) \qquad r_{2i}(\hat{\lambda}_n, \beta, \sigma) = (y_i - x_i'\beta - (\hat{\phi}_i/\hat{\Phi}_i)\sigma)\begin{pmatrix} x_i \\ \hat{\phi}_i/\hat{\Phi}_i \end{pmatrix} \cdot 1_{[y_i>0]} \ ,$$

where $\hat{\phi}_i = \phi_i(x_i'\hat{\lambda}_n)$ and $\hat{\Phi}_i = \Phi_i(x_i'\hat{\lambda}_n)$.

This two stage estimator can be written as a single stage estimator of form (2.3) by considering the estimator $\hat{\tilde{\theta}}_n = (\hat{\lambda}_n', \hat{\beta}_n', \hat{\sigma}_n)'$ of $\tilde{\theta}_0 \equiv (\beta_0'/\sigma_0, \beta_0', \sigma_0)'$ defined by the function

$$(3.15) \qquad r_i(\tilde{\theta}) = \begin{pmatrix} r_{1i}(\lambda) \\ r_{2i}(\lambda, \beta, \sigma) \end{pmatrix} .$$

Since the errors have all moments finite in this example, the stability of $\hat{\tilde{\theta}}_n$ under P_{θ_0} depends on the number p of finite monents of the regressors x_i. In particular, the stability of $\hat{\tilde{\theta}}_n$ and of $(\hat{\beta}_n', \hat{\sigma}_n)'$ is $1 - 1/p$.

8. Seemingly unrelated nonlinear regression--Zellner's feasible Aitken estimator (references: Zellner (1962), Gallant (1975)). The model consists of M equations:

$$(3.16)\quad y_{im} = f_m(x_{im}, \beta_{0m}) + u_{im}\ ,\quad m = 1, \ldots, M\ ,\quad i = 1, \ldots, n\ ;$$

$$Z_i = (y_{i1}, \ldots, y_{iM}, x_{i1}', \ldots, x_{iM}')'\ .$$

Under Gallant's (1975) assumptions, the error vectors $u_i \equiv (u_{i1}, \ldots, u_{iM})'$ satisfy $Eu_i = \underset{\sim}{0}$, $Eu_iu_i' = \Sigma_0$, and $Eu_iu_\ell' = \underset{\approx}{0}$ for $i \neq \ell$; the variables x_{im} are fixed; and the regression functions $f_m(x_{im}, \beta_{0m})$ are smooth, have bounded first derivatives, and behave like i.i.d. rv's for n large. Let $\theta_0 \equiv (\beta_0', \alpha_0')'$, where $\beta_0 \equiv (\beta_{01}', \ldots, \beta_{0M}')'$, $\alpha_0 \equiv S \operatorname{vec} \Sigma_0$, and S is known $M(M+1)/2 \times M^2$ selection matrix such that α_0 is the vector obtained by vectorizing the lower triangle of the symmetric matrix Σ_0.
The feasible Aitken estimator has three stages, each of which yields an estimator which is the solution to a system of equations.
1st stage: The estimator $\hat{\lambda}_n$ is an equation by equation LS estimator of β_0. Its defining function is

$$(3.17)\qquad r_{1i}(\beta) = \begin{pmatrix} (y_{i1} - f_1(x_{i1}, \beta_1))\frac{\partial}{\partial \beta_1} f_1(x_{i1}, \beta_1) \\ \vdots \\ (y_{iM} - f_M(x_{iM}, \beta_M))\frac{\partial}{\partial \beta_M} f_M(x_{iM}, \beta_M) \end{pmatrix}\ ,\ \text{for}\ \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}\ .$$

2nd stage: The estimator $\hat{\alpha}_n$ of α_0 is based on the first-stage residuals. Its defining function is

(3.18) $\quad r_{2i}(\hat{\lambda}_n, \alpha) = S\ \mathrm{vec}(\Sigma - \hat{u}_i\hat{u}_i') \ ,$

where $\hat{u}_i \equiv (y_{i1} - f_1(x_{i1}, \hat{\lambda}_{1n}), \ldots, y_{iM} - f_M(x_{iM}, \hat{\lambda}_{Mn}))'$, Σ is an $M \times M$ matrix defined by $\mathrm{vec}\ \Sigma = D\alpha$, and D is the known $M^2 \times M(M+1)/2$ duplication matrix defined such that Σ is symmetric and α is the vectorization of the lower triangle of Σ .

3rd stage: The estimator $\hat{\beta}_n$ is a multi-equation weighted LS estimator of β_0 . Its defining function is

(3.19) $\quad r_{3i}(\hat{\alpha}_n, \beta) = (y_i - f_i(\beta))'\hat{\Sigma}_n^{-1} \frac{\partial}{\partial\beta} f_i(\beta) \ ,$

where $\mathrm{vec}\ \hat{\Sigma}_n \equiv D\hat{\alpha}_n$, $y_i \equiv (y_{i1}, \ldots, y_{iM})'$, and $f_i(\beta) \equiv (f_i(x_{i1}, \beta_1), \ldots, f_M(x_{iM}, \beta_M))'$.

We write this multi-stage estimator in the single stage form of (2.3) by taking

(3.20) $\quad r_i(\tilde{\theta}) = (r_{1i}(\lambda)', r_{2i}(\lambda,\alpha)', r_{3i}(\alpha,\beta)')' \ ,$

to yield an estimator $\hat{\tilde{\theta}}_n \equiv (\hat{\lambda}_n', \hat{\alpha}_n', \hat{\beta}_n')'$ of the parameter vector $\tilde{\theta}_0 \equiv (\beta_0', \alpha_0', \beta_n')'$. In this example, the stability of the estimators $\hat{\tilde{\theta}}_n$ and $(\hat{\beta}_n, \hat{\alpha}_n)$ is $1 - 2/p$, where p is the number of finite moments of the errors u_i .

9. Maximum likelihood (ML) and pseudo-maximum likelihood estimators (references: Wald (1949), Huber (1967), Hoadley (1971), Akaike (1973), Crowder (1976), White (1982)). ML and pseudo-ML estimators (both defined as solutions to likelihood equations) can be written in the form (2.3) for all econometric models, provided the log-likelihood (or pseudo-log-likelihood) function is based on the assumption of independent observations, and is

differentiable in its parameter θ . ML estimators are defined by the score function

$$(3.21) \qquad r_i(\theta) = \frac{\partial}{\partial\theta}\log p(z_i, \theta) ,$$

where $p(z_i, \theta)$ is the density of Z_i with respect to some measure μ . Pseudo-ML estimators are defined identically, except $p(z_i, \theta)$ is some specified density which is not necessarily assumed to be the true density of Z_i . In addition, an estimator defined by (3.21) is called a pseudo-ML estimator if the observations are not independent, since in this case $\sum_{i=1}^{n} \log p(z_i, \theta)$ is not the log-likelihood of the sample. For the results of Section 2 to hold, all that is needed is that the observations are strong mixing or φ-mixing, and the score function satisfies the conditions B1-B5 on $r_i(\theta)$. Under quite general conditions, ML and pseudo-ML estimators have been shown to be strongly consistent, so B1 is not a problem. Further, assumptions B2-B5 are easy to verify and are satisfied in most econometric models. The stability of ML and pseudo-ML estimators depends on the number of moments (and perhaps tail behavior) of their score functions as established in Section 2. Examples include:

(i) Binary logit model: The rv y_i takes values 0 or 1. The probability that y_i equals 1 is $P_i(\theta) \equiv \exp(x_i'\theta)/(1 + \exp(x_i'\theta))$, where x_i is a fixed or random explanatory variable. The ML estimator is defined by

$$(3.22) \qquad r_i(\theta) = [y_i - \exp(x_i'\theta)/(1 + \exp(x_i'\theta))]x_i .$$

The first multiplicand of $r_i(\theta)$ lies in $(-1,1)$, so the stability of $\hat{\theta}_n$ depends on the explanatory variables x_i , $i = 1, \ldots, n$. If the

x_i are fixed and uniformly bounded, the stability of $\hat{\theta}_n$ is one. If the x_i are i.i.d. with p finite moments, the stability is $1 - 1/p$. (Note, the extension to the multinomial logit model is straightforward.

(ii) Binary probit model: The model is the same as the logit model, except $P_i(\theta) \equiv \Phi(x_i'\theta)$, where $\Phi(\cdot)$ is the standard normal df. The ML probit estimator $\hat{\theta}_n$ is defined by

$$(3.23) \qquad r_i(\theta) = \frac{y_i - \Phi(x_i'\theta)}{\Phi(x_i'\theta)[1 - \Phi(x_i'\theta)]}\phi(x_i'\theta)x_i .$$

It is easy to see that the stability properties of the ML probit estimator are the same as those of the ML logit estimator.

(iii) Censored regression (Tobit) model (see Amemiya (1973)): The model is the same as in example 7. The ML estimator $\hat{\theta}_n = (\hat{\beta}_n', \hat{\sigma}_n)'$ is defined by

$$(3.24) \quad r_i(\theta) = \begin{pmatrix} \dfrac{\phi_i(\theta)x_i}{1 - \Phi_i(\theta)}1[y_i=0] + \dfrac{1}{\sigma^2}(y_i - x_i'\beta)x_i 1[y_i>0] \\ \dfrac{\phi_i(\theta)x_i'\beta}{\sigma^2(1 - \Phi_i(\theta))}1[y_i=0] + \dfrac{1}{\sigma}\left[\left(\dfrac{y_i - x_i'\beta}{\sigma}\right)^2 - 1\right]1[y_i>0] \end{pmatrix},$$

where $\phi_i(\theta) \equiv \phi(x_i'\beta/\sigma)$ and $\Phi_i(\theta) \equiv \Phi(x_i'\beta/\sigma)$. The form of $r_i(\theta)$ shows that the ML estimator has the same stability properties as Heckman's two stage estimator (see example 7).

(iv) Seemingly unrelated nonlinear regression model: The model is the same as in example 8 where $\theta_0 \equiv (\beta_0', \alpha_0')'$. The pseudo-ML estimator of θ_0 formed using the multivariate normal$(\underset{\sim}{0}, \Sigma)$ distribution for the errors $u_i \equiv (u_{i1}, \ldots, u_{iM})'$ is defined by

$$(3.25)\quad r_i(\theta) = \begin{pmatrix} (y_i - f_i(\beta))'\Sigma^{-1}\frac{\partial}{\partial\beta}f_i(\beta) \\ S\ \mathrm{vec}[\Sigma - (y_i - f_i(\beta))(y_i - f_i(\beta))'] \end{pmatrix}, \quad \text{for} \quad \theta = (\beta', \alpha')' ,$$

where $\alpha = S\ \mathrm{vec}\ \Sigma$ and S is defined in example 8. The pseudo-ML estimator is very similar to the feasible Aitken estimator of example 8. They both have the same stability properties.

For brevity we have not included the 2SLS, 3SLS, LIML, and FIML estimators of linear simultaneous equations models in the examples given above. 2SLS and 3SLS can be written in the form (2.3) via the method of examples 7 and 8 (using Theil's (1953) interpretation of 2SLS). LIML can be so written using its interpretation as the FIML estimator of an incomplete system of equations (see Godrey and Wickens (1977), and Phillips and Wickens (1978, pp. 276, 351)). Finally, FIML is trivially of the form (2.3) under the assumption of independent errors.

4. Proofs

The proofs of Theorem 1 and other results below use the following lemma:

Lemma 1. Let $\{Y_i\}$ be a sequence of mean zero, strong mixing rv's with strong mixing numbers which satisfy A1. Assume $\sup_{i\geq 1} E|Y_i|^{c-\delta} < \infty$, for some $c > 1$ and all δ arbitrarily small and positive. Then, for any sequence of positive integers $\{k_n\}$ with $k_n \leq n$,

$$n^{\zeta}\bar{Y}_{n,k_n} \xrightarrow{n\to\infty} 0 \quad \text{a.s.}, \quad \forall \zeta < 1 - 1/(2 \wedge (c/\alpha)) ,$$

where $\bar{Y}_{n,k_n} \equiv \frac{1}{n-1}\sum_{i=1, i\neq k_n}^{n} Y_i$.

Proof of Theorem 1. Let $\tilde{\sum_i}$ denote a summation over i from 1 to n with $i \neq k_n$. Using $\tilde{\sum_i} r_{ij}(\hat{\theta}_{n,k_n}) = 0$, a Taylor expansion of $n^{\nu-1}\tilde{\sum_i} r_{ij}(\hat{\theta}_{n,k_n})$ about θ_0 yields

$$(4.1)\quad 0 = n^{\nu-1}\tilde{\sum_i} r_{ij}(\theta_0) + \frac{1}{n}\tilde{\sum_i}\left[\frac{\partial}{\partial\theta'} r_{ij}(\theta_0) + (\hat{\theta}_{n,k_n} - \theta_0)'\frac{\partial^2}{\partial\theta\partial\theta'} r_{ij}(\theta^*_{n,k_n})\right] n^{\nu}(\hat{\theta}_{n,k_n} - \theta_0),$$

and so,

$$(4.2)\quad 0 = o(1) + (a_j + o(1))n^{\nu}(\hat{\theta}_{n,k_n} - \theta_0) \quad \text{a.s.}, \quad \forall \nu < 1 - 1/(2 \wedge (p/\alpha)),$$

for $j = 1, \ldots, J$, where θ^*_{n,k_n} is a random vector on the line segment joining $\hat{\theta}_{n,k_n}$ and θ_0, a_j is the j^{th} row of A, and $o(1)$ is a random vector of appropriate dimension which is of small order one as $n \to \infty$ a.s. (4.2) follows from (4.1) using (i) Lemma 1 and B3 to show $n^{\nu-1}\tilde{\sum_i}[r_{ij}(\theta_0) - Er_{ij}(\theta_0)] = o(1)$ as $n \to \infty$ a.s., (ii) the assumption B2 that $n^{\nu-1}\tilde{\sum_i} Er_{ij}(\theta_0) = o(1)$ as $n \to \infty$, (iii) Lemma 1 and B4 to show $\frac{1}{n}\tilde{\sum_i}\left[\frac{\partial}{\partial\theta} r_{ij}(\theta_0) - E\frac{\partial}{\partial\theta} r_{ij}(\theta_0)\right] = o(1)$ as $n \to \infty$ a.s., and (iv) equation (4.5) below and the strong consistency of $\hat{\theta}_{n,k_n}$ to give $\frac{1}{n}\tilde{\sum_i}(\hat{\theta}_{n,k_n} - \theta_0)'\frac{\partial^2}{\partial\theta\partial\theta'} r_{ij}(\theta^*_{n,k_n}) = o(1)$ as $n \to \infty$ a.s. Stacking equations (4.2) for $j = 1, \ldots, J$ to form a system of equations yields part b of the Theorem, since A is non-singular. Part a follows in a similar fashion from (4.1) and (4.2) taking $\nu = 0$ and $k_n = n-1$.

It remains to show (4.5). By Lemma 1 and B5a,

$$(4.3)\quad \frac{1}{n}\tilde{\sum_i}(W_{ij} - EW_{ij}) \xrightarrow{n\to\infty} 0 \quad \text{a.s., and} \quad \frac{1}{n}\tilde{\sum_i} EW_{ij} = O(1) \quad \text{as} \quad n \to \infty, \quad \forall j = 1, \ldots, J.$$

Now, for any sequence of rv's $\{\tilde{\theta}_n\}$ such that $\tilde{\theta}_n \xrightarrow{n\to\infty} \theta_0$ a.s., $\tilde{\theta}_n$ is in Θ_0 for n sufficiently large a.s. (where Θ_0 is some neighborhood of θ_0, see B5), and so, for n sufficiently large,

$$(4.4)\quad \frac{1}{n}\sum_i \left\| \frac{\partial^2}{\partial\theta\partial\theta'} r_{ij}(\tilde{\theta}_n) - \frac{\partial^2}{\partial\theta\partial\theta'} r_{ij}(\theta_0) \right\| \le \frac{1}{n}\sum_i W_{ij} = O(1) \quad \text{as} \quad n \to \infty \quad \text{a.s.},$$

by (4.3). Hence, using B5b,

$$(4.5)\qquad \frac{1}{n}\sum_i \left\| \frac{\partial^2}{\partial\theta\partial\theta'} r_{ij}(\tilde{\theta}_n) \right\| = O(1) \quad \text{as} \quad n \to \infty\ , \quad \text{a.s.}, \quad \forall j = 1, \ldots, J\ . \ \square$$

Proof of Lemma 1. First we show that under the assumptions of the Lemma

$$(4.6)\qquad n^{\zeta}\bar{Y}_n \xrightarrow{n\to\infty} 0 \quad \text{a.s.}, \quad \forall \zeta < 1 - 1/(2 \wedge (c/\alpha))\ .$$

We apply McLeish's (1975) Lemma 2.9 to the rv's $X_n \equiv Y_n/n^{1-\zeta}$, where using his notation we set $d_n = 1$, $\forall n$, $g_n(x) = |x|^{s(\delta)}$ for $s(\delta) \equiv (c-\delta) \wedge 2\alpha$, and $\bar{X}_n \equiv X_n 1[|X_n| \le d_n]$. Since $\sum_{n=1}^{\infty} E^{1/\alpha} g_n(|X_n|) < \infty$, provided $\zeta < 1 - \alpha/s(\delta)$, we have $\sum_{n=1}^{\infty} (X_n - E\bar{X}_n)$ converges a.s. by his Lemma 2.9. Now, by the proof of Loeve's (1955) Theorem 16.4.A (p. 241),

$$(4.7)\qquad \sum_{n=1}^{\infty} (EX_n - E\bar{X}_n) < \infty\ , \quad \text{provided} \quad |s(\delta)| \ge 1\ .$$

Thus, provided $(c-\delta) \wedge 2\alpha \ge 1$ (which requires $c > 1$ and δ arbitrarily small and positive), we have $\sum_{n=1}^{\infty} X_n$ converges a.s. Applying Kronecker's Lemma gives (4.6) for $\zeta < 1 - \alpha/s(\delta)$. Since δ is arbitrarily small, (4.6) holds for all $\zeta < 1 - \alpha/s(0)$, as desired.

Now, simple algebra gives $\overline{Y}_n - \overline{Y}_{n,k_n} = \frac{1}{n}Y_{k_n} - \frac{1}{n}\overline{Y}_{n,k_n}$, and so,

$$(4.8) \qquad n^{\zeta}\overline{Y}_{n,k_n}\left(1 - \frac{1}{n}\right) = n^{\zeta}\overline{Y}_n - n^{\zeta-1}Y_{k_n} .$$

Using (4.6) we have

$$(4.9) \qquad n^{\zeta-1}Y_n = n^{\zeta}\overline{Y}_n - n^{\zeta}\left(\frac{n-1}{n}\right)\overline{Y}_{n-1} \xrightarrow{n\to\infty} 0 \quad \text{a.s.}$$

Thus, for any subsequence $\{k_n\}$ of $\{n\}$ with $k_n \le n$, $\forall n$,

$$(4.10) \qquad n^{\zeta-1}|Y_{k_n}| \le k_n^{\zeta-1}|Y_{k_n}| \xrightarrow{n\to\infty} 0 \quad \text{a.s.}, \quad \forall \zeta < 1 - 1/(2 \wedge (c/\alpha)) .$$

Combining (4.6), (4.8), and (4.10) gives the desired result. □

The proof of Theorem 2 uses the following Lemma:

Lemma 2. Let $\{Y_i\}$ be as in Lemma 1 and assume $|Y_i| \overset{ST}{\le} Y$, $\forall i$, for some rv Y which satisfies $E|Y|^{c-\delta} < \infty$ for some $c \ge \alpha$ and all δ arbitrarily small and positive. If

$$(4.11) \qquad \sum_{n=1}^{\infty} [1 - G_n^*(n^{1-\tau})] < \infty , \quad \text{for some } \tau < 1 ,$$

where $G_n^*(x) \equiv \min_{1\le i\le n} G_i(x)$ and $G_i(x)$ is the df of Y_i, then for all sequences of positive integers $\{k_n\}$ with $k_n \le n$,

$$\lim_{n\to\infty} n^{\zeta}(\overline{Y}_n - \overline{Y}_{n,k_n}) = 0 \quad \text{a.s.}, \quad \forall \zeta < \tau ,$$

where $\overline{Y}_{n,k_n}$ is as in Lemma 1.

Proof of Theorem 2. We prove the results of comment 5 following Theorem 2. These results imply those of the Theorem. We prove comment 5 part b first. It suffices to show: if

$$(4.12)\quad \sum_{n=1}^{\infty}[1 - F^{*}_{nj}(n^{1-\xi})] < \infty\ ,\quad \forall j = 1, \ldots, J\ ,\quad \text{and}\quad \xi < 2(1 - 1/(2 \wedge (\tilde{p}/\alpha)))\ ,$$

then for all sequences of positive integers $\{k_n\}$ with $k_n \leq n$, we have

$$(4.13)\quad \lim_{n\to\infty} n^{\zeta}|\hat{\theta}_n - \hat{\theta}_{n,k_n}| = 0 \quad \text{a.s.},\quad \forall \zeta < \xi\ .$$

Let $\bar{r}_n = \frac{1}{n}\sum_{i=1}^{n} r_i(\theta_0)$, $\bar{r}_{n,k_n} = \frac{1}{n-1}\sum_{i}^{\sim} r_i(\theta_0)$, $A_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}r_i(\theta_0)$, and $A_{n,k_n} = \frac{1}{n-1}\sum_{i}^{\sim}\frac{\partial}{\partial\theta}r_i(\theta_0)$. Using (4.1), (4.2), and Theorem 1 part b, we have

$$(4.14)\quad 0 = \bar{r}_n - \bar{r}_{n,k_n} + A_n(\hat{\theta}_n - \theta_0) - A_{n,k_n}(\hat{\theta}_{n,k_n} - \theta_0) + o(n^{-2\nu}) \quad \text{a.s.},$$

for $\nu < 1 - 1/(2 \wedge (\tilde{p}/\alpha))$. By definition of ν and ζ we can take ν such that $2\nu \geq \zeta$. This, plus manipulation of (4.14), gives

$$(4.15)\quad -n^{\zeta}(\hat{\theta}_n - \hat{\theta}_{n,k_n}) = n^{\zeta}A_n^{-1}(\bar{r}_n - \bar{r}_{n,k_n}) + n^{\zeta}A_n^{-1}(A_n - A_{n,k_n})(\hat{\theta}_{n,k_n} - \theta_0) + o(1) \quad \text{a.s.}$$

where A_n^{-1} exists for n sufficiently large a.s., since A is non-singular and $A_n \xrightarrow{n\to\infty} A$ a.s. by Lemma 1 and B4.

For all ℓ, $j = 1, 2, \ldots, J$, and $\tau \equiv 1 - 1/\eta$,

$$(4.16)\quad \sum_{n=1}^{\infty} P\left(\left|\frac{\partial}{\partial\theta_\ell} r_{nj}(\theta_0)\right| > n^{1-\tau}\right) \leq \sum_{n=1}^{\infty} P([Dr]_{\ell,j}^{1/(1-\tau)} > n) \leq E[Dr]_{\ell,j}^{1/(1-\tau)} + 1 < \infty\ ,$$

where the first inequality uses the definition of Dr , the second inequality follows by Loeve (1955, Moments Inequality, p. 242), and the third inequality follows by B4b for all η satisfying $\eta \geq 2 \wedge (\tilde{p}/\alpha)$ and $\eta > \alpha$. Lemma 2 applied element by element now gives

$$(4.17) \qquad n^{\delta}(A_n - A_{n,k_n}) \xrightarrow{n\to\infty} \underset{\sim}{0} \text{ a.s.}, \quad \forall \delta < \tau \ (\equiv 1 - 1/\eta) ,$$

where $\underset{\sim}{0}$ is a $J \times J$ matrix of zeros. Thus, using Theorem 1 part b,

$$(4.18) \qquad n^{\zeta}A_n^{-1}(A_n - A_{n,k_n})(\hat{\theta}_{n,k_n} - \theta_0) = o(1) \text{ as } n \to \infty \text{ a.s.},$$

provided $\zeta - \delta - \nu \leq 0$. Algebraic manipulation verifies this inequality.

For $\zeta < \xi$ where $\sum_{n=1}^{\infty} [1 - F_{nj}^{*}(n^{1-\xi})] < \infty$, $\forall j = 1, \ldots, J$, Lemma 2 gives

$$(4.19) \qquad n^{\zeta}(\bar{r}_n - \bar{r}_{n,k_n}) \xrightarrow{n\to\infty} \underset{\sim}{0} \text{ a.s.}$$

Equations (4.15), (4.18), and (4.19), and the result $A_n^{-1} \xrightarrow{n\to\infty} A^{-1}$ a.s. yield (4.13), as desired.

Now we show that comment 5 part b implies comment 5 part a. For all $j = 1, \ldots, J$, and all $\xi < 1 - 1/p$,

$$(4.20) \qquad \sum_{n=1}^{\infty} [1 - F_{nj}^{*}(n^{1-\xi})] \leq \sum_{n=1}^{\infty} P((r_{Uj})^{1/(1-\xi)} > n) \leq E(r_{Uj})^{1/(1-\xi)} + 1 < \infty ,$$

where the third inequality holds for all $\xi < 1 - 1/p$ since r_U has p finite moments, the second inequality follows by Loeve (1955, Moments Inequality, p. 242), and the first inequality holds by definition of r_U . For $\tilde{p} \in (2\alpha-1, p]$ (which requires $p > 2\alpha-1$) and $\xi < 1 - 1/\tilde{p}$, we can show $\xi < 2(1 - 1/(2 \wedge (\tilde{p}/\alpha)))$. This and (4.20) give part a of comment 5. □

Proof of Lemma 2. Simple algebra gives,

$$(4.21) \qquad n^{\zeta}(\overline{Y}_n - \overline{Y}_{n,k_n}) = n^{\zeta-1}Y_{k_n} - n^{\zeta-1}\overline{Y}_{n,k_n} .$$

By assumption,

$$(4.22) \qquad \sum_{n=1}^{\infty} P(n^{\tau-1}|Y_{k_n}| \geq 1) \leq \sum_{n=1}^{\infty} [1 - G_n^*(n^{1-\tau})] < \infty ,$$

so the first Borel-Cantelli Lemma gives $P(n^{\tau-1}|Y_{k_n}| \geq 1 \text{ i.o.}) = 0$, where i.o. abbreviates "infinitely often." Thus, $\forall \zeta < \tau$,

$$(4.23) \qquad n^{\zeta-1}|Y_{k_n}| \xrightarrow{n\to\infty} 0 \quad \text{a.s.}$$

Lemma 1 and the assumption $c \geq \alpha$ give

$$(4.24) \qquad n^{\zeta-1}|\overline{Y}_{n,k_n}| \xrightarrow{n\to\infty} 0 \quad \text{a.s.},$$

since $\zeta-1 < 0$. Equations (4.21), (4.23), and (4.24) combine to give the desired result. □

The proof of Theorem 3 uses the following Lemma:

Lemma 3. Let $\{Y_i\}$ be as in Lemma 2 with the further assumption that $\{Y_i\}$ is φ-mixing. If

$$\sum_{n=1}^{\infty} [1 - G_n^*(n^{1-\tau})] = \infty , \quad \text{for some } \tau < 1 ,$$

where $G_n^*(\cdot)$ is as in Lemma 2, then for some sequence of positive integers $\{k_n\}$ with $k_n \leq n$,

$$\limsup_{n\to\infty} n^{\zeta}|\overline{Y}_n - \overline{Y}_{n,k_n}| = \infty \quad \text{a.s.}, \quad \forall \zeta > \tau ,$$

where $\overline{Y}_{n,k_n}$ is as in Lemma 1.

Proof of Theorem 3. We prove the results of comment 4 following Theorem 3. These results (and Theorem 2) imply those of Theorem 3. Consider comment 4 part b first. The result is trivial if C is null, so assume C is non-empty. It suffices to show that for $\xi \in C$, any ζ larger than but arbitrarily close to ξ, and some sequence $\{k_n\}$ with $k_n \leq n$,

$$\lim_{n\to\infty} n^{\zeta}|\hat{\theta}_n - \hat{\theta}_{n,k_n}| = \underset{\sim}{0} \quad \text{a.s.} \tag{4.25}$$

does not hold. For ξ, ζ, and $\{k_n\}$ as above, (4.15) and (4.18) yield

$$n^{\zeta}|\hat{\theta}_n - \hat{\theta}_{n,k_n}| = n^{\zeta}|A_n^{-1}(\overline{r}_n - \overline{r}_{n,k_n})| + o(1) \quad \text{a.s.}, \tag{4.26}$$

since $\zeta < 2(1 - 1/(2 \wedge (\tilde{p}/\alpha)))$, and provided $\zeta - \delta - \nu \leq 0$ for some $\delta \leq 1 - 1/\eta$, where A_n^{-1} exists for n sufficiently large a.s. Given the former condition on ζ, the latter condition holds if $\eta \geq 2 \wedge (\tilde{p}/\alpha)$, as is assumed.

Using a proof by contradiction we show that for some sequence $\{k_n\}$

$$\limsup_{n\to\infty} n^{\zeta}|A_n^{-1}(\overline{r}_n - \overline{r}_{n,k_n})| = \underset{\sim}{\infty}^{1} \quad \text{a.s.}, \tag{4.27}$$

where $\underset{\sim}{\infty}^{1}$ denotes a J-vector with at least one element equal to ∞. Let ω denote a realization of the process $\{Z_i\}$. If (4.27) does not hold, then for all ω in a set with positive probability we have

$$n^{\zeta}|(A_n^{\omega})^{-1}(\overline{r}_n^{\omega} - \overline{r}_{n,k_n}^{\omega})| \leq M^{\omega} \cdot \underset{\sim}{e} , \quad \forall n = 1, 2, \ldots , \tag{4.28}$$

for some scalar $M^{\omega} < \infty$, where $\underset{\sim}{e}$ is a vector of ones and the superscript ω indicates the particular realization ω. For such ω and n sufficiently large,

$$(4.29)\quad n^{\zeta}|\bar{r}_n - \bar{r}_{n,k_n}| = n^{\zeta}|A_n A_n^{-1}(\bar{r}_n - \bar{r}_{n,k_n})| \leq M\cdot|A_n|\cdot\underset{\sim}{e} \leq M\cdot|A + \varepsilon\underset{\sim}{e}\underset{\sim}{e}'|\cdot\underset{\sim}{e} < \underset{\sim}{\infty}^2$$

where the superscript ω has been omitted in (4.29) for notational convenience, $\underset{\sim}{\infty}^2$ denotes a J-vector of infinities, the first inequality holds by simple algebra, and the second inequality holds for n sufficiently large given $\varepsilon > 0$ since $A_n \xrightarrow{n\to\infty} A$ a.s. But, Lemma 3 implies $\limsup_{n\to\infty} n^{\zeta}|\bar{r}_n - \bar{r}_{n,k_n}| = \underset{\sim}{\infty}^1$ a.s. for some sequence $\{k_n\}$. This contradicts (4.29) and implies (4.27) is true. (4.25) and (4.27) combine to give the result of comment 4 part b.

Next we show comment 4 part b implies comment 4 part a. For $q < \infty$, it suffices to show $\xi \equiv 1 - 1/(q+\varepsilon)$ is in C for ε arbitrarily small and positive. For this ξ, $E|r_{Lj}|^{q+\varepsilon} = \infty$ for some integer j in $\{1, \ldots, J\}$. Thus,

$$(4.30)\quad \sum_{n=1}^{\infty}[1 - F^*_{nj}(n^{1-\xi})] \geq \sum_{n=1}^{\infty}[1 - F_{Lj}(n^{1/(q+\varepsilon)})] \geq E|r_{Lj}|^{q+\varepsilon} - 1 = \infty,$$

where the second inequality follows by Loeve (1955, Moments Inequality, p. 242). In addition, algebraic manipulation shows that $\tilde{p} \in (2\alpha q/(q+1), p]$ and $n \geq 2 \wedge (\tilde{p}/\alpha)$ implies $\xi < 2(1 - 1/(2 \wedge (\tilde{p}/\alpha)))$, for ε sufficiently small. Hence, $\xi \in C$.

For the case $q = \infty$, part a says $\Lambda(\hat{\theta}_n, P_{\theta_0}) \leq 1$. The latter is true whether or not $q = \infty$, if $r_L \not\equiv 0$. To see this, consider $\xi = 1+\varepsilon$ for ε arbitrarily small and positive. For this ξ,

$$(4.31) \qquad \sum_{n=1}^{\infty} [1 - F^*_{nj}(n^{1-\xi})] \geq \sum_{n=1}^{\infty} [1 - F_{Lj}(n^{-\varepsilon})] = \infty ,$$

unless $F_{Lj}(0) = 1$. Since $r_{Lj} \geq 0$, $\forall j$, (4.31) holds for all $\varepsilon > 0$ and all j unless $r_L \equiv \underset{\sim}{0}$. Thus, part a holds for $q = \infty$, and more generally, $\Lambda(\hat{\theta}_n, P_{\theta_0}) \leq 1$, provided $r_L \not\equiv \underset{\sim}{0}$. □

Proof of Lemma 3. It suffices to show the result for $\zeta \in (\tau,1]$. Using (4.21), we have

$$(4.32) \qquad n^{\zeta}|\overline{Y}_n - \overline{Y}_{n,k_n}| \geq n^{\zeta-1}|Y_{k_n}| - n^{\zeta-1}|\overline{Y}_{n,k_n}| .$$

Let $\{k_n\}$ be a sequence such that $G_{k_n}(n^{1-\tau}) = G^*_n(n^{1-\tau})$, $\forall n$. Then,

$$(4.33) \qquad \sum_{n=1}^{\infty} P(n^{\tau-1}|Y_{k_n}| > 1) = \sum_{n=1}^{\infty} [1 - G^*_n(n^{1-\tau})] = \infty .$$

The second Borel-Cantelli Lemma, which holds not only for independent events but for events determined by φ-mixing rv's as well, see Iosifescu and Theodorescu (1969, Lemma 1.1.2'), now gives $P(n^{\tau-1}|Y_{k_n}| > 1 \text{ i.o.}) = 1$. Thus,

$$(4.34) \qquad \limsup_{n\to\infty} n^{\zeta-1}|Y_{k_n}| = \infty \quad \text{a.s.}, \quad \forall \zeta \in (\tau,1] .$$

Also, since $\zeta \leq 1$, $n^{\zeta-1}|\overline{Y}_{n,k_n}|$ converges to zero as $n \to \infty$ a.s. by Lemma 1 and the assumption $c \geq \alpha$. Thus, (4.34) and (4.32) combine to give the desired result. □

FOOTNOTES

[1]Assumption B1 is actually weaker than strong consistency because B1 requires $\hat{\theta}_n \xrightarrow{n\to\infty} \theta_0$ a.s.$[P_\theta]$ only for $\theta = \theta_0$ rather than for all θ in some parameter space Θ.

[2]In models which are misspecified (i.e., a parametric family $\{P_\theta : \theta \in \Theta\}$ is specified, but the true distribution P of $\{Z_i\}$ is not in the family), the definition or identification of the estimand θ_0 is sometimes problematical. One solution, which is more or less satisfactory depending upon the situation, is to take the estimand θ_0 to be the a.s. limit of the estimator under P (e.g., see Bickel and Lehmann (1975), Maronna and Yohai (1981), Huber (1973), and White (1980, 1982).) For example, θ_0 may be defined as the unique solution to $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E_P r_i(\theta) = \underset{\sim}{0}$. Using this approach, B1a is satisfied under any form of misspecification (i.e., any P), provided the estimator $\hat{\theta}_n$ has an a.s. (non-random) limit under P.

[3]Assumption B5 requires that $r_i(\theta)$ is twice differentiable in some neighborhood of θ_0. This is not needed for asymptotic normality in general, but is needed for the stability results (see equation (4.14) of the proof of Theorem 2).

[4]This altering of the assumption of mean zero errors only affects the definition of the constant term, and hence, is relatively innocuous.

REFERENCES

Akaike, H. (1973). "Information theory and an extension of the likelihood principle," Proc. Second Inter. Symp. Information Thy. B. N. Petrov and F. Csáki, eds. Akadémai Kiado, Budapest.

Amemiya, T. (1973). "Regression analysis when the dependent variable is truncated normal," Econometrica 41 997-1016.

Anderson, T. W. and J. B. Taylor (1979). "Strong consistency of least squares estimates in dynamic models," Ann. Statist. 7 484-489.

Andrews, D. W. K. (1983). "Robust and efficient estimation of nonlinear regression models with dependent errors," manuscript, Cowles Foundation, Yale University.

Andrews, D. W. K. (1984a). "A zero-one result for the least squares estimator," Cowles Foundation Discussion Paper No. 698, Yale University.

Andrews, D. W. K. (1984b). "Non-strong mixing autoregressive processes," forthcoming J. Appl. Prob.

Belsley, D. A., E. Kuh, and R. E. Welsch (1980). Regression Diagnostics J. Wiley and Sons, New York.

Bickel, P. J. and E. L. Lehmann (1975). "Descriptive statistics for nonparametric models. I. Introduction," Ann. Statist. 3 1038-1044.

Bierens, H. J. (1981). Robust Methods and Asymptotic Theory in Nonlinear Econometrics. Lecture Notes in Economics and Mathematical Systems, No. 192. Springer-Verlag, New York.

Billingsley, P. (1968). Convergence of Probability Measures. J. Wiley and Sons, New York.

Burguete, J. F., A. R. Gallant, and G. Souza (1982). "On the unification of the asymptotic theory of nonlinear econometric models," Econometric Reviews 1 151-190.

Chanda, K. C. (1974). "Strong mixing properties of linear stochastic processes," J. Appl. Prob. 11 401-408.

Cook, R. D. (1977). "Detection of influential observations in linear regression," Technometrics 19 15-18.

Cook, R. D. (1979). "Influential observations in linear regression," J. Am. Statist. Assoc. 74 169-174.

Crowder, M. J. (1976). "Maximum likelihood estimation for dependent observations," J. Roy. Statist. Soc. B 38 45-53.

Domowitz, I. and H. White (1982). "Misspecified models with dependent observations," J. Econometrics 20 35-58.

Gallant, A. R. (1975). "Seemingly unrelated nonlinear regressions," J. Econometrics 3 35-50.

Godfrey, L. G. and M. R. Wickens (1978). "The estimation of incomplete models using subsystem LIML," Essex University Discussion Paper No. 99.

Hampel, F. R. (1974). "The influence curve and its role in robust estimation," J. Amer. Statist. Assoc. 69 383-393.

Hannan, E. J. and M. Kanter (1977). "Autoregressive processes with infinite variance," J. Appl. Prob. 14 411-415.

Heckman, J. J. (1979). "Sample selection bias as a specification error," Econometrica 47 153-161.

Heiler, S. (1981). "Strong and weak consistency of instrumental variables estimates and application to dynamic models," manuscript, Dortmund University.

Hoadley, B. (1971). "Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case," Ann. Math. Statist. 42 1977-1991.

Huber, P. J. (1967). "The behaviour of maximum likelihood estimates under nonstandard conditions," Proc. Fifth Berkeley Symp. Math. Statist. Prob. 221-233. University of California Press, Berkeley.

Huber, P. J. (1973). "Robust regression: Asymptotics, conjectures, and Monte Carlo," Ann. Statist. 1 799-821.

Huber, P. J. (1981). Robust Statistics. J. Wiley and Sons, New York.

Ibragimov, I. A. and Yu. V. Linnik (1971). Independent and Stationary Sequences of Random Variables. Wolters-Noordhoff Pub. Co., The Netherlands.

Iosifescu, M. and R. Theodorescu (1969). Random Processes and Learning. Die Grundlehren der mathematischen Wissenschaften Band 150. Springer-Verlag, New York.

Jennrich, R. I. (1969). "Asymptotic properties of non-linear least squares estimators," Ann. Math. Statist. 40 633-643.

Kinal, T. W. (1980). "The existence of moments of k-class estimators," Econometrica 42 517-527.

Kolmogorov, A. N. and Y. A. Rozonov (1960). "On strong mixing conditions for stationary Gaussian processes," Theor. Probability Appl. 5 204-208.

Krasker, W. S. and R. E. Welsch (1982). "Efficient bounded-influence regression estimation," J. Am. Stat. Assoc. 77 595-604.

Krasker, W. S. and R. E. Welsch (1983). "Resistant estimation for simultaneous-equations models using weighted instrumental variables," manuscript, Sloane School of Management, M.I.T.

Lai, T. L., H. Robbins, and C. Z. Wei (1978). "Strong consistency of least squares estimates in multiple regression," Proc. Natl. Acad. Sci. USA 75 3034-3036.

Loeve, M. (1955). Probability Theory. D. Van Nostrand, New York.

Maronna, R. A. and V. J. Yohai (1981). "Asymptotic behaviour of general M-estimates for regression and scale with random carriers," Z. Wahrscheinlichkeitstheorie 58 7-20.

McLeish, D. L. (1975). "A maximal inequality and dependent strong laws," Ann. Probability 3 829-839.

Phillips, P. C. B. and M. R. Wickens (1978). Exercises in Econometrics, Volume II. Ballinger Pub. Co., Cambridge, MA.

Theil, H. (1953). "Repeated least-squares applied to complete equation systems," manuscript, Central Planning Bureau, The Netherlands.

Wald, A. (1949). "Note on the consistency of the maximum likelihood estimate," Ann. Math. Statist. 20 595-601.

White, H. (1980). "Using least squares to approximate unknown regression functions," Int. Econ. Rev. 21 149-170.

White, H. (1982). "Maximum likelihood estimation of misspecified models," Econometrica 50 1-25.

Wu, C. F. (1981). "Asymptotic theory of nonlinear least squares estimation," Ann. Statist. 9 501-513.

Yohai, V. J., and R. A. Maronna (1979). "Asymptotic behaviour of M-estimators for the linear model," Ann. Statist. 7 258-268.

Zellner, A. (1962). "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias," J. Am. Stat. Assoc. 57 348-368.