Journal of Econometrics 37 (1988) 157–169. North-Holland

LARGE-SAMPLE PROPERTIES OF METHOD OF MOMENT ESTIMATORS UNDER DIFFERENT DATA-GENERATING PROCESSES

B.M.S. VAN PRAAG and T. KLOEK

Econometric Institute, Erasmus University, 3000 DR Rotterdam, The Netherlands

J. DE LEEUW

University of California, Los Angeles, CA 90024-1563, USA

In this paper we study the large-sample properties of method of moment estimators (MME) of population parameters β that result as explicit or implicit solitions of equations $f(\mu, \beta) = 0$, where μ is a vector of moments. The distribution of the corresponding large-sample estimator b resulting from solving $f(m, b) = 0$, where m is a sample moment estimator of μ, is shown to vary with the sample design. We study the case of a random sample, of a pooled sample, a sample from a known population distribution, and a so-called repeated sample. Also we provide a numerical example. Our main conclusion is that any seemingly harmless assumption on the sample design may have severe repercussions on the estimation of confidence bounds. Therefore we advocate to take a kind of minimal set of assumptions unless we have hard *a priori* knowledge.

1. Introduction

In econometrics, like in all sciences, scientific development may be characterized by the rise and decline of paradigms. The first statistical paradigm in econometrics had not much to do with statistics in the sense of econometrics today. Statistical methods were purely descriptive and involved taking averages, sample standard deviations, and at the top of sophistication regressing one variable Y on one or several variables $X_1, \dots, X_k$. Researchers did not care about the fact that the estimated measures would vary among data sets. No standard deviations of estimators, t-ratios, etc. were calculated. With the work of Haavelmo (1944) and the Cowles Commission it was realized that econometrics is not only data collection and measurement but also statistics. However, if we have a data set $\{y_t, x_t\}_{t=1}^{T}$ and regress y on x yielding a regression slope coefficient $b = \Sigma x_t y_t / \Sigma x_t^2$, then its stochastic properties can only be derived on the basis of some assumptions on the data-generating process. The standard procedure became to assume that the regression line was the *true* model, i.e., the data-generating process is described by an equation $y_t = \beta x_t + \alpha + \varepsilon_t$, where ε_t is a random error term with $\mathrm{E}(\varepsilon_t) = 0$, $\sigma^2(\varepsilon_t) = \sigma^2$. Clearly, this is a very restrictive assumption. In later literature [e.g., White (1980, 1982), Hansen (1982)] this assumption has been relaxed by assuming that the model to be estimated may be a misspecification of the real model.

In all those situations there is of course a natural inclination to test hypotheses about the true model by constructing test statistics the distribution of which under the null hypothesis can be derived. However, all these results are *conditional* in the sense that they only hold if the basic assumption of the true model holds. More recently it seems that we are returning to the modest standpoint that any model is really only an approximation of reality and that we can never claim that we have the true model analytically specified [see White (1980), Van Praag (1981), Chamberlain (1982), Hansen (1982)]. This implies that we cannot derive the probability distribution of our estimators on the basis of a model assumption. This situation is well-known in statistics, but it is not a problem for *large* samples as we may use the Central Limit Theorem. It states that the sum $T^{-1/2}(\Sigma X_t - \Sigma\mu_t)$ is asymptotically normal $N(0, V)$ under mild conditions, which may be loosely summarized by the following three assumptions:

(a) The X_t are i.i.d.
(b) The X_t have finite expectation.
(c) The X_t have a finite covariance matrix.

This assumption set (or some of its many variations, allowing for not too different distributions or slight interdependence) seems to be roughly speaking the *minimal* assumption set to make asymptotic statements about sample statistics without knowing the distribution of the X_t's itself. This idea may also be employed for many econometric problems.

First let us sketch the idea in its most simple setting. We have a random variable Y to be explained by a constant α and a random variable X. A best explanation is given by minimizing $E(Y - \beta X - \alpha)^2$ yielding the *population* regression coefficients $\hat{\beta}$ and $\hat{\alpha}$. It is well known that $\hat{\beta} = \sigma_{yx}/\sigma_{xx}$, where σ_{yx} and σ_{xx} are the covariance between Y and X and the variance of X, respectively. The natural sample estimator is $\hat{b} = s_{yx}/s_{xx}$ where s_{yx} and s_{xx} stand for the corresponding sample moments. Now it is well-known that the second-order sample moments $T^{1/2}(s - \sigma)$ are asymptotically normal due to the Central Limit Theorem, if their variances, i.e., fourth-order moments exist. As $\hat{b}$ is a differentiable function in $\hat{\beta} = \sigma_{yx}/\sigma_{yy}$, it can be derived after linearization that $T^{1/2}(\hat{b} - \hat{\beta})$ is also asymptotically $N(0, V(\hat{b}))$.

More generally, let us assume a population approximation yielding a population parameter $\hat{\beta}$ that is the unique solution of a set of equations $f(\mu, \beta) = 0$, where μ is a vector of moments and f is differentiable in $(\mu, \hat{\beta})$, then we may estimate $\hat{\beta}$ by $\hat{b}$, that is the unique solution of $f(m, b) = 0$, where m is the sample analogue of μ. Let μ stand for a kth-order moment vector and μ_2 stand for the corresponding $2k$th-order moment vector and let us assume $\mu_2 < \infty$, then it may be derived that $T^{1/2}(\hat{b} - \hat{\beta})$ is also asymptotically normal N(0, V). The specification of the covariance matrix V depends on the asymptotic distribution of m.

Clearly looking at economic models as approximations implies a fundamental change for hypothesis testing. If one assumes model A to be true, we may derive the distribution of the parameter estimators of model B, say $\hat{\theta}(B)$, under the condition that A holds, and inversely. In the present approach we may find the *joint* distribution of $\hat{\theta}(A)$ and $\hat{\theta}(B)$ given the same data set. For instance, let $\hat{\theta}(A)$ be $h_1(m_1)$ and $\hat{\theta}(B)$ be $h_2(m_2)$. If the joint distribution of (m_1, m_2) is known, the joint distribution of $h_1(m_1)$ *and* $h_2(m_2)$ may be derived as well. Notice that B is not necessarily nested in A. Secondly, we may study the joint distribution of $h(m^{(1)})$ *and* $h(m^{(2)})$, where $m^{(1)}$ and $m^{(2)}$ are referring to a vector of the same moments, calculated on two *different* samples.

The structure of this paper is the following. In section 2 we consider the traditional *random* sample. In section 3 we study the *heterogeneous* sample. Section 4 investigates the case of a sample from a parent distribution that is known to be *normal* or *elliptical*. In section 5 we consider *panels*. Section 6 gives a numerical example and section 7 concludes. The objective of this paper is not so much to strive for utmost generality as to provide some easily applicable formulae for the situations we encounter in practice. It should be emphasized, that we assume that our samples consist of i.i.d. observations, that is, we are not considering time series. We believe that the philosophy of this paper may be extended to that domain, but for reasons of space and exposition we do not make an attempt at this place.

2. Method of moment estimators in a random sample

Let us assume a random vector X on which we have T i.i.d. observations $\{X_t\}_{t=1}^{T}$. We are interested in a population parameter vector $\hat{\beta}$, that is a unique solution of an equation set $f(\mu, \beta) = 0$, where μ is a vector of moments and f a differentiable vector function in an open neighbourhood about $(\mu, \hat{\beta})$. Let m be the sample analogue of μ, where $T^{1/2}(m - \mu)$ is asymptotically N(0, V) distributed. The estimator $\hat{b}$ is the solution of $f(m, \hat{b}) = 0$. So we have $f(\mu, \hat{\beta}) - f(m, \hat{b}) = 0$ and it follows that up to a first-order approximation there holds

$$\nabla f_{\mu}(m - \mu) + \nabla f_{\beta}(\hat{b} - \hat{\beta}) = 0,$$

where ∇f_{μ} and ∇f_{β} are gradient matrices of f with respect to μ and β.

If the number of parameters β equals the number of equations $f(\cdot) = 0$ and if ∇f_{β} has full rank, we get asymptotically

$$T^{1/2}(\hat{b} - \hat{\beta}) = -\left[\nabla f_{\beta}\right]^{-1} \nabla f_{\mu} T^{1/2}(m - \mu),$$

and consequently, if $T^{1/2}(m-\mu)$ is N(0, V), we get

$$T^{1/2}(\hat{b}-\hat{\beta}) \xrightarrow{D} \mathrm{N}(0, V(\hat{b})),$$

where

$$V(\hat{b}) = T(\nabla f_\beta)^{-1}\nabla f_\mu V\left[(\nabla f_\beta)^{-1}\nabla f_\mu\right]'. \tag{2.1}$$

Here we applied the so-called delta-method [see, e.g., Rao (1973)]. Let us now consider how V is consistently estimated. Let $M_t = g(X_t)$, $\mu = E(M)$, $m = (1/T)\Sigma M_t$. (We think of g as any measurable function of X with finite expectation.) Then we have plim(m) = μ, and

$$V = T\cdot \mathrm{E}(m-\mu)(m-\mu)',$$

is estimated consistently by

$$\hat{V} = \frac{1}{T}\sum_{t=1}^{T}(M_t-\mu)(M_t-\mu)' \approx \frac{1}{T}\sum_{t=1}^{T}(M_t-m)(M_t-m)'. \tag{2.2}$$

Actually (2.1) assessed by evaluating $\hat{\beta}$ by $\hat{b}$ and V by (2.2) yields a general formula for the covariance matrix of the asymptotic distribution of $T^{1/2}(\hat{b}-\hat{\beta})$. If V does not exist, it follows that (2.1) is not finite as well. This formula (2.1) requires the calculation of sometimes formidable derivatives but it is a very general formula. In practice most estimators of linear models are functions of the sample covariance matrix S of X, while the covariance matrix of vec(S) depends on fourth-order moments [see also Wesselman (1987)].

Consider for instance two competing models A and B, where the estimators $\hat{b}_A$ and $\hat{b}_B$ are solutions of $f_A(S, \hat{b})$ and $f_B(S, \hat{b})$, depending on the same covariance matrix S. Then we may derive the joint distribution of $(\hat{b}_A, \hat{b}_B)$ by applying (2.1) on the function

$$f(S,b) = \begin{bmatrix} f_A(S,b) \\ f_B(S,b) \end{bmatrix}.$$

This may be an alternative approach to compare various non-nested approximating models. In the literature Hansen (1982) considered the *generalized* method of moments estimators (GMME). The population parameter $\hat{\beta}$ is the solution of an equation set $\mathrm{E}_x f(X,\beta) = 0$ which is estimated by $\hat{b}$, the solution of

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t, b) = 0.$$

Here the distribution of the separate random elements depends on the unknown $\hat{\beta}$. It follows that we arrive at a formula similar to (2.1) where the matrix V however depends on $\hat{\beta}$ as well. This is caused by the difference between $\mathrm{E}f(X,\beta)$ and $f(\mu,\beta)=f(\mathrm{E}M,\beta)$. There are many cases where the operations $f(\cdot)$ and E( ) may be interchanged by suitable definition of M. For instance, in the regression case we switch from the natural variable X_t to the cross-products $X_{it}X_{jt}=M_{ij,t}$. There are other cases where such a change is impossible. An example is the Probit model. It follows that Hansen's model of GMME includes the class of MME considered here. However, as the MME are so easy to handle we believe it worthwhile to consider the MME in its own right.

3. Heterogeneous samples pooled together

Let us assume we have q mutually independent samples $\{X_t^{(1)}\}_{t=1}^{T_1}, \ldots, \{X_t^{(q)}\}_{t=1}^{T_q}$, drawn from q different distributions each of which is a random sample itself, and let us define $T=\sum_{i=1}^{q}T_i$, with $T_i/T=p_i$. The mixture of those samples will be called a *heterogeneous sample*. This situation may be the case if several samples from different sources are pooled or if we only have aggregate figures of subgroups available. It also may occur if we do not have enough computer memory to store the whole data set.

Let us assume the subsample moment vectors are $m^{(1)},\ldots,m^{(q)}$ and their covariance matrices are $V^{(1)},\ldots,V^{(q)}$, respectively, where those matrices may be estimated by (2.2). In that case we may calculate the overall moment vector

$$m=p_1m^{(1)}+\cdots+p_qm^{(q)}, \tag{3.1}$$

the expectation of which is μ. This yields the consistent estimator $\hat{b}$, the unique solution of $f(m,b)=0$. As m is a sum of q mutually independent vectors, the pooled covariance matrix is found to be

$$TV=T\sum_{i=1}^{q}p_i\{V^{(i)}+(m^{(i)}-\mu)(m^{(i)}-\mu)'\}. \tag{3.2}$$

In the special case that the samples have the same $m^{(i)}=m$ the latter term vanishes. In practice the population parameter μ is estimated by m. If T is large, $(m-\mu)T^{1/2}$ is approximately normal with expectation zero and covariance matrix (3.2). The asymptotic results on $b(m)$, derived in the previous sections, may be derived where V in (2.1) is replaced by (3.2).

The situation becomes different if the subsamples become so small that we cannot reasonably assume the vectors $m^{(i)}$ to be approximately normal. The limiting situation is that where $T=q$ is large and $T_i=1$ $(i=1,\ldots,q)$. In that case (3.1) is just a sum of many independent but not necessarily identically distributed vectors; it is well-known that also in that situation $(m-\mu)T^{1/2}$ may tend to normality, if the conditions of the Central Limit Theorem are fulfilled. We shall not elaborate that case.

4. The normal and elliptical parent distributions

A popular assumption in practical work is that the sample is drawn from a specific distribution that is known. This has two advantages. First the MME can be interpreted not only as a geometric approximation but mostly also as an ML-estimator. Second, the asymptotic distribution of $b(m)$ can be analytically specified in some cases.

It is common to assume that the family of k-vectors $\{X_t\}_{t=1}^{T}$ has been drawn from a normal distribution $N(0, \Sigma)$.[1] It is well-known that in case of normality all higher-order moments are functions of Σ, hence any moment m and the derived parameter $b(m)$ is known as soon as Σ is known and we may replace all higher-order moments by suitable expressions in terms of second-order moments. Then $b(\mu) \equiv \tilde{b}(\Sigma)$. Similarly, estimation of higher than second-order moments μ by m boils down to estimation of Σ by S. As soon as the covariance matrix Π of the sample covariance matrix S is known, we know the large sample distribution of any $\tilde{b}(S)$. It is nearly costless to generalize this result referring to a parent normal distribution for the case of covariance functions. Let us assume that the parent distribution of X is *elliptical*, where an elliptical distribution is described by a density function $Cg(x'\Sigma^{-1}x)$ with C a normalizing constant. The normal distribution is a special case with $g(x'\Sigma^{-1}x) = \exp(-1/2x'\Sigma^{-1}x)$. The Student distribution is another member. This distribution family, extensively studied by Muirhead (1982) and others, is very rich and nevertheless only slightly less tractable than the multivariate normal subfamily. In the elliptical case it can be shown that the covariance matrix Π of the sample covariance matrix S, defined by its elements

$$\pi_{hi,jl} = T \cdot \operatorname{cov}(s_{hi}, s_{jl}),$$

equals

$$\pi_{hi,jl} = \kappa(\sigma_{hi}\sigma_{jl} + \sigma_{hj}\sigma_{il} + \sigma_{hl}\sigma_{ij}) + \sigma_{hj}\sigma_{il} + \sigma_{hl}\sigma_{ij}, \tag{4.1}$$

where κ is the *common kurtosis parameter* [see Van Praag, Dijkstra and Van

[1] This applies for the structural model, but also for the so-called functional model specification, where the stochastic element is introduced by means of a random error term. Prucha and Kelejian (1984) generalized the latter approach to assuming the error term to obey a multivariate t-distribution.

Velzen (1985)]. In case of normality, $\kappa = 0$. It can be shown that the marginal distributions of $X_1, \ldots, X_k$ have the same kurtosis defined by $3\kappa = \kappa_4^j/(\kappa_2^j)^2$ ($j = 1, \ldots, k$), where κ_2^j and κ_4^j are the second- and fourth-order cumulants of X_j. The estimation of κ and a test whether X is elliptically distributed can be based on the identities [see, e.g., Cramèr (1946)]

$$\kappa_2^j = \mu_2^j - (\mu_1^j)^2 (= \sigma_{jj}),$$

$$\kappa_4^j = \mu_4^j - 4\mu_1^i \mu_3^j - 3(\mu_2^j)^2 + 12\mu_2^j(\mu_1^j)^2 - 6(\mu_1^j)^4, \tag{4.2}$$

where

$$\mu_i^j = \mathrm{E}(X_j^i).$$

As in our example $\mu_1 = 0$ we find as sample estimate of 3κ

$$m_4^j/(m_2^j)^2 - 3 = 3\hat{\kappa}, \qquad j = 1, \ldots, k. \tag{4.3}$$

[See for a less naive estimator Mardia (1970).]

We see that, if X is elliptically distributed, then

$$T^{1/2}(\hat{b} - \hat{\beta}) \xrightarrow{\mathrm{D}} \mathrm{N}(0, V(\hat{b})) \quad \text{with} \quad V(\hat{b}) = [\nabla \hat{b}_\Sigma]' \cdot \Pi \cdot \nabla \hat{b}_\Sigma,$$

and Π defined by (4.1). It can be shown that $\kappa > -\frac{2}{3}$ [Rao (1973, p. 143)] in order that the integral of the density converges. We ignore the case $\kappa = -\frac{2}{3}$ which yields a two-point distribution [see Kendall and Stuart (1977, p. 88)]. The advantage of the elliptical assumption is clear. If κ is known, we do not need to estimate any fourth-order moments of X, as (4.1) may be evaluated by second-order moments only. If κ is unknown, we need one fourth-order moment to be inserted in (4.3). However, for a sample not positively known to be drawn from an elliptical population, it will not give the correct estimates of the standard deviations of $\hat{b}$ (see section 7 for empirical comparisons). Let us denote (4.1) as

$$\pi_{hi,jl} = \kappa \pi_{hi,jl}^{(1)} + \pi_{hi,jl}^{(2)},$$

where $\Pi^{(1)}$ and $\Pi^{(2)}$ are defined by the first and latter two terms of (4.1), respectively. Accordingly we may write

$$\Pi = \kappa \Pi^{(1)} + \Pi^{(2)}. \tag{4.4}$$

The matrix Π is a symmetric positive semi-definite (s.p.s.d.) matrix for all

$\kappa > -\frac{2}{3}$. Then it is easy to see that Π is a s.p.s.d. matrix, increasing in κ. This leads to an important corollary. Substituting (4.4) into $[\nabla b_{\Sigma}]'\Pi[\nabla b_{\Sigma}]$ it follows that for given S the covariance matrix $V(\hat{b})$ is increasing in κ. Notice that the second term at the right-hand side of (4.4) is the covariance matrix when $\kappa = 0$, that is, under the assumption that X has been drawn from a normal population. [See also Wesselman and Van Praag (1987).] We notice two *caveats* with respect to this extension to the elliptical distribution. First, higher-order moments than second-order ones may involve other parameters than Σ and κ, depending on the shape of the elliptical density $g(x)$. Hence, if higher than second-order moments are involved in m, knowledge of π is not sufficient to establish the distribution of the MME. Second, κ is the common kurtosis of the κ marginal distributions of $X_1, \ldots, X_k$. If they are significantly unequal, then our data are certainly not drawn from an elliptical population. Testing this equality would involve eighth-order moments.

5. Panels

Finally let us consider a panel. There is a crucial difference betwen the first and the latter waves of a panel. The first wave is a random sample in the sense of section 2. The second wave is correlated with the first. For instance, consider two waves of a household panel, which we denote by $\{X_t^{(1)}\}_{t=1}^T$ and $\{X_t^{(2)}\}_{t=1}^T$, respectively. In general we may assume independence between households, but most certainly $X_t^{(2)}$ will depend in a stochastic sense on the earlier observation $X_t^{(1)}$. Formally, we may assume that the first wave is a random sample of observations $\{X_t^{(1)}\}_{t=1}^T$ with a corresponding vector of relevant sample moments $m^{(1)}$. Consider now the second wave $\{X_t^{(2)}\}_{t=1}^T$ with corresponding $m^{(2)}$. And let us assume we are interested in an estimate of $\hat{\beta}$, solving the population equation $f(\mu, \beta) = 0$. It is estimated on the second wave by $\hat{b}^{(2)}$ which solves $f(m^{(2)}, b) = 0$. We may call $\hat{b}^{(2)}$ the second-wave estimate.

It is obvious that there is a dependency between $\hat{b}^{(2)}$ and $\hat{b}^{(1)}$, the corresponding estimator on the first wave. Their relation depends on that between $m^{(1)}$ and $m^{(2)}$. Consider the joint distribution of $(m^{(1)}, m^{(2)})$. If T is large, that distribution tends to the normal and this holds also for the conditional distribution of $m^{(2)}|m^{(1)}$ [see also Steck (1957)]. Let us denote the conditional covariance matrix by $\Sigma_{m^{(2)}|m^{(1)}}$, then the conditional asymptotic distribution of $(\hat{b}^{(2)} - \hat{\beta}^{(2)})T^{1/2}$ is normal with expectation zero and covariance matrix as in (1.1), with V replaced by $\Sigma_{m^{(2)}|m^{(1)}}$. The precise specification of $\Sigma_{m^{(2)}|m^{(1)}}$ depends on the intertemporal relationship between $X_t^{(1)}$ and $X_t^{(2)}$. In this paper we shall consider the simplest specification. Let the vector m be partitioned as $m = (m_1, m_2)$, where m_1 has length p_1 and m_2 has length p_2, and let us assume $m_1^{(1)} = m_1^{(2)}$ while $m_2^{(1)}$ and $m_2^{(2)}$ are conditionally independent, that is, the conditional density $h(m_2^{(2)}|m_1^{(1)}, m_2^{(1)}) = h(m_2^{(2)}|m_1^{(1)})$.

The general situation is the following. Let the unconditioned distribution of (m_1, m_2) be asymptotically normal with covariance matrix

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}.$$

Then the conditional asymptotic distribution of $m_2^{(2)}$ given $m_1^{(2)} = m_1^{(1)}$ is the normal distribution where

$$\mathrm{E}\left(m_2^{(2)} | m_1^{(2)} = m_1^{(1)}\right) = \mu_2 + V_{21} V_{11}^{-1}\left(m_1^{(1)} - \mu_1\right), \tag{5.1}$$

$$TV\left(m_2^{(2)} | m_1^{(2)} = m_1^{(1)}\right) = V_{22} - V_{21} V_{11}^{-1} V_{12}. \tag{5.2}$$

If $m_1^{(1)}$ is a consistent estimator of the population parameter μ_1, (5.1) will tend to μ_2. However, it is obvious from (5.2) that $TV(m_2^{(2)} | m_1^{(2)} = m_1^{(1)}) < V_{22}$ if $V_{21} V_{11}^{-1} V_{12} > 0$. For functions $\hat{b}^{(2)} = b(m_1^{(2)}, m_2^{(2)})$ it implies that $\hat{b}^{(2)}$ is only random by virtue of $m_2^{(2)}$ but not by $m_1^{(2)}$, which is known beforehand to equal $m_1^{(1)}$.

Eq. (2.1) is still applicable, but the covariance matrix in (2.1) is reduced for two reasons. First no derivative is taken with respect to m_1 as it is non-random. Hence the dimension $(p \times p)$ of V in (2.1) reduces to $(p_2 \times p_2)$, where p_2 is the length of m_2. Second V is replaced by (5.2). The matrix (5.2) itself is estimated by its sample analogue. Hence the covariance matrix of $(\hat{b}^{(1)} - \hat{\beta}^{(1)})T^{1/2}$ may be written as

$$V(\hat{b}^{(1)}) = \nabla b V [\nabla b]',$$

where $[\nabla b_{m_1}] = [\nabla f_b]^{-1} \nabla f_{m_1}$ [see (2.1)].

The covariance matrix of $(\hat{b}^{(2)} - \hat{\beta}^{(2)})T^{1/2}$, given $m_1^{(2)} = m_1^{(1)}$, equals

$$V\left(\hat{b}^{(2)} | m_1^{(2)} = m_1^{(1)}\right) = \left[\nabla b_{m_2}\right]\left(V_{22} - V_{12} V_{11}^{-1} V_{21}\right)\left[\nabla b_{m_2}\right]'.$$

This implies for standard deviations that any standard deviation of a component of $\hat{b}^{(2)}$ is *smaller* than that of the corresponding component of $\hat{b}^{(1)}$. This outcome may be quite important for the construction of tests and confidence bounds.

A rather familiar example of this situation is also hidden in the classical regression model

$$Y = \beta' X + \varepsilon,$$

where X is assumed to be non-random (i.e., determined in a previous wave) and hence Σ_{XX} fixed. For the OLS regression estimator $\hat{b} = S_{XX}^{-1} S_{XY}$ this implies that $S_{XX}(= m_1)$ is constant over subsequent samples and that we are in the second wave of a panel as just described. This implies then for the usual OLS model, that assuming the OLS classical model with S_{XX} constant, while the sample is completely random (i.e., S_{XX} random as well) yields too optimistic conclusions for reliability, too small standard deviations, and too narrow confidence intervals [see also White (1980), Van Praag (1981), and Van Praag, De Leeuw and Kloek (1986)].

6. A numerical example

In this section we illustrate the previous findings for the case of OLS regression, where we assess the standard deviations of the regression coefficients under various assumptions on the data-generating process. Our data base consists of a cross-section of 2206 observations of households, which are assumed to be i.i.d. For each household we know family size fs, its after-tax household income y_c, and an estimate by the head of the household of the minimum income $y_{\min}$ 'he would need to make ends meet for his household'. Obviously this amount $y_{\min}$ depends on fs, while it is 'anchored' to current income y_c. So a relation

$$\ln y_{\min} \approx \beta_0 + \beta_1 \ln fs + \beta_2 \ln y_c, \tag{6.1}$$

lies at hand. For the theoretical background, which is evidently psychologically flavoured, we refer to Helson (1964), Van Praag (1971), Goedhart et al. (1977) and Van Praag (1985).[2]

Our regression estimate is

$$\ln y_{\min} = 4.563 + 0.157 \ln fs + 0.508 \ln y_c, \qquad R^2 = 0.527, \quad N = 2206.$$

The standard deviations of the regression coefficients are evaluated under various assumptions. We compare in table 1 the following assumptions:

(1) $y_{\min}$, fs, y_c are i.i.d. random observations (section 2).

(2) $y_{\min}$, fs, y_c are drawn from an elliptical parent distribution with $3\kappa = -1, 0, 2, 4, 6$ [this includes the case of a normal parent distribution ($\kappa = 0$), section 4].

(3) $y_{\min}$, fs, y_c are the second wave of a panel, where fs and y_c are fixed (section 5).

(4) The sample is drawn from a population for which the approximating linear equation is the data-generating process, that is, there holds

$$\ln y_{\min} = \beta_0 + \beta_1 \ln fs + \beta_2 \ln y_c + \varepsilon,$$

with ε a normal error, not depending on the random regressors fs and y_c. The regressors fs and y_c are random and the joint distribution of fs, y_c is unknown.

[2] For a description of the data set, see Van Praag, Goedhart and Kapteyn (1980).

Table 1

Standard deviations under various assumptions.

	β_0	β_1	β_2
1. Random sample	0.330	0.017	0.034
2. Elliptical sample			
$\kappa = -1$	0.075	0.007	0.008
$\kappa = 0$	0.129	0.012	0.013
$\kappa = 2$	0.167	0.015	0.017
$\kappa = 4$	0.197	0.018	0.020
$\kappa = 6$	0.223	0.020	0.023
3. Second wave	0.227	0.013	0.023
4. Linear model with normal error[a]	0.262	0.017	0.026
5. Classical linear model	0.129	0.012	0.013

[a] This case has not been explicitly considered in the text but it can be straightforwardly derived by the reader by similar methods.

(5) The sample satisfies the traditional OLS assumptions, i.e., a controllable experiment (*fs* and y_c non-random) and the linear model holds with

$$\varepsilon \sim N(0, \sigma^2).$$

The first rather striking result emanating from table 1 is that the reliability assessment of estimates, as reflected by their standard deviations, varies a great deal with the prior assumptions made. For instance, when we consider the elliptical assumption with a varying value of κ, we find that normality is just one parameter choice among many and that the reliability of the same estimator on the same sample decreases rapidly if we opt for greater values of κ. Not by coincidence the classical linear model assumptions yield standard deviations identical to that under the normal assumption ($\kappa = 0$) [cf. Goldberger (1964)]. It follows that the reliability estimates of the classical linear model (or of the normal parent population) cannot have any special claim as being more valid than other estimates presented in the table, except those in line 1, corresponding to the random sample assumption. This is especially relevant as most standard-deviation estimates presented in table 1 are much larger than classical OLS estimates, which implies that the OLS standard deviations may give a rather optimistic look on the reliability. The consequence is that t-statistics used for significance tests may be too high; similar considerations hold for other frequently used test-statistics.

The main conclusion is that all assumptions 2 to 5, also that of OLS, are arbitrary choices and consequently that all confidence estimates have an arbitrary basis as well. Obviously, there is only one confidence estimate which does not suffer from arbitrary assumptions. That is the estimator investigated in section 2, of which the corresponding standard deviations are given in line

1. It is seen that for this example the standard deviations are larger than for most other assumptions. However, this is not always true (cf. $\kappa = 6$). The only fact that always holds is that the assumption of keeping some variables fixed reduces standard deviations in an unrealistic way. If one of the specific assumptions, on which lines 2 to 5 are based, hold, the general formula will yield automatically the special case. So it seems wise and prudent to start with the general formula, developed in section 2.

7. Concluding remarks

In this paper we consider moment estimators under weaker assumptions than usually made. The theory outlined in this paper holds for all method of moments estimators. For the special case of the OLS model our result in section 2 coincides with that of White (1980). For the special case of a linear equation system, Chamberlain (1982) finds the same result.

In this approach we may also define residual vectors e_t as the difference between the observation and its best prediction by the approximating function. Then we find for linear approximating functions that so-defined residuals and explanatory variables are uncorrelated when we use a Euclidean distance approximation. It does *not* follow, however, that the residuals and the explanatory variables are independent. This additional assumption is usually made, but it is not made here. Under these circumstances the ordinary least-squares estimator is still consistent, but the standard formula for its covariance matrix does not hold. So, if we are not certain that the independence assumption mentioned above is appropriate, it may be advisable not to use the traditional formula derived under the independence assumption, but the more general result discussed in this paper.

From this paper it is clear that for all estimators b, that are either explicitly or implicitly defined as functions of sample moments, the large-sample stochastic properties are basically known if we know the covariance matrix of the sample moments involved, say $V(m)$. If specific additional assumptions are made with respect to the sample design, it affects $V(m)$ but not m or $b(m)$. It follows that stochastic properties may be analyzed for classes of estimators that involve the same sample moments just by studying $V(m)$. This decomposability is called *population-sample decomposability* [see also Van Praag, de Leeuw and Kloek (1986)].

Compared to the established theory of the estimation of linear models and structures there seem to be three major methodological advantages inherent in this approach. First, it is based on weaker postulates; second, it unifies the approach of moment estimators, instead that we have to consider the theory of regression, of factor analysis, or of linear models as separate theories. Third, it yields more realistic confidence estimates that coincide with those of the classical theory if the appropriate assumptions apply to the real situation.

There is also a potential disadvantage. This occurs in the case that some of the observations are outliers or, more generally, influential observations [see Krasker et al. (1983)]. The fact that higher than second-order moments are used may have the consequence that the effects of rare influential observations may be more serious than in the traditional case. This possibility requires further research.

References

Browne, M.W., 1982, Covariance structures, in: D.M. Hawkins, ed., Topics in applied multivariate analysis (Cambridge University Press, Cambridge).

Chamberlain, G., 1982, Multivariate regression models for panel data, Annals of Applied Econometrics 1, 5–46.

Cramér, H., 1946, Mathematical methods of statistics (Princeton University Press, Princeton, NJ).

Goedhart, T., V. Halberstadt, A. Kapteyn and B.M.S. Van Praag, 1977, The poverty line: Concept and measurement, Journal of Human Resources 12, 503–520.

Goldberger, A., 1964, Econometric theory (Wiley, New York).

Hansen, L.P., 1982, Large sample properties of generalized method of moment estimators, Econometrica 50, 1029–1054.

Helson, H., 1964, Adaptation-level theory: An experimental and systematic approach to behavior (Harper, New York).

Haavelmo, T., 1944, The probability approach in econometrics, Econometrica 12, Suppl.

Jöreskog, K.G., 1973, A general method for estimating a linear structural equation system, in: A.S. Goldberger and O.D. Duncan, eds., Structural equation models in the social sciences (Seminar Press, New York) 85–112.

Kendall, M.C. and A. Stuart, 1977, Advanced theory of statistics, Vol. 1 (Griffin, London).

Krasker, W.S., E. Kuh and R.E. Welsch, 1983, Estimation for dirty data and flawed models, in: Z. Griliches and M.D. Intriligator, eds., Handbook of econometrics, Vol. 1 (North-Holland, Amsterdam) 652–700.

Mardia, K.V., 1970, Measures of multivariate skewness and kurtosis with applications, Biometrika 57, 519–539.

Muirhead, R.J., 1982, Aspects of multivariate statistical theory (Wiley, New York).

Prucha, J.R. and H.H. Kelejian, 1984, The structure of simultaneous equation estimators: A generalization to non-normal disturbances, Econometrica 52, 721–736.

Rao, C.R., 1973, Linear statistical inference and its application (Wiley, New York).

Steck, G.P., 1957, Linear statistical inference and its applications (Wiley, New York).

Van Praag, B.M.S., T. Goedhart and A. Kapteyn, 1980, The poverty line: A pilot survey in Europe, Review of Economics and Statistics 62, 461–465.

Van Praag, B.M.S., 1981, Model-free regression, Economics Letters 7, 139–144.

Van Praag, B.M.S. and N.L. van der Sar, 1985, Household cost functions and equivalence scales, Journal of Human Resources, forthcoming.

Van Praag, B.M.S., T.K. Dijkstra and J. Van Velzen, 1985, Least-squares theory based on general distributional assumptions with an application to the incomplete observations problem, Psychometrika 50, 25–36.

Van Praag, B.M.S., J. de Leeuw and T. Kloek, 1986, The population-sample decomposition approach to multivariate estimation methods, Applied Stochastic Models and Data Analysis 2, 99–119.

Wesselman, A.M. and B.M.S. van Praag, 1987, Elliptical regression operationalized, Economics Letters 23, 269–274.

Wesselman, A.M., 1987, The population-sample decomposition method: A distribution-free estimation technique for minimum distance parameters, Ph.D. dissertation (Erasmus University, Rotterdam) and Kluwer Academic Publications 19 (Kluwer, Dordrecht).

White, H., 1980, Using least squares to approximate unknown regression functions, International Economic Review 21, 149–170.

White, H., 1982, Maximum likelihood estimation of misspecified models, Econometrica 50, 1–25.