University College London

Department of Economics

M.Sc. in Economics

MC3: Econometric Theory and Methods

Course Notes 3

Notes on approximate inference

Andrew Chesher

31/10/2005

1. Introduction

These notes start with an introduction to large sample approximate inference and the concepts of convergence in probability to a constant and to a random variable. These ideas are used to develop approximate methods for making inferences using the OLS and GLS estimators in non-normal models. A general method for developing the limiting distribution of M-estimators is sketched and applied to the non-linear least squares estimator.

2. Approximate inference

The results set out in the previous notes let us make inferences about coefficients of regression functions, β, when the distribution of y given X is Gaussian (normal) and the variance of the unobservable disturbances is known. But in practice the normal distribution at best holds approximately and of course we never know the value of the nuisance parameter σ. So how can we proceed? The most common approach and the one outlined here involves employing *approximations* to exact distributions[1]. They have the disadvantage that they can be inaccurate and the magnitude of the inaccuracy can vary substantially from case to case. They have the following advantages.

1. They are usually very much easier to derive than exact distributions,
2. They are often valid for a wide family of distributions for y while exact distributional results are only valid for a specific distribution of y, and we rarely know which distribution to use to produce an exact distributional result.

The most common sort of approximation employed in econometrics is what is known as a *large sample approximation.*

Suppose we have a statistic, S_n, computed using n realisations, for example the OLS estimator, $\hat{\beta}$, or the variance estimator $\hat{\sigma}^2$, or one of the test statistics developed earlier.

To produce a large sample approximation to the distribution of the statistic, S_n, we regard this statistic as a member of a sequence of statistics, $S_1, \ldots, S_n, \ldots$, indexed by n, the number of realisations. We write this sequence as $\{S_n\}_{n=1}^{\infty}$. Denote the distribution function of S_n by $P[S_n \leq s] = F_{S_n}(s)$. We then consider how the distribution function $F_{S_n}(s)$ behaves as we pass through the sequence, that is as n takes larger and larger values. In particular we ask what properties the distribution function has as n tends to infinity[2]. The distribution associated with the limit of the sequence of statistics is sometimes referred to as a *limiting distribution.* Sometimes this distribution can be used to produce an approximation to $F_{S_n}(s)$ which can be used to conduct approximate inference using S_n. In many cases the limiting behaviour of $F_{S_n}(s)$ is the same under a variety of conditions. Then the approximation may be useful under all, or many of, these conditions.

[1] The main, and increasingly popular, alternative to the use of approximations is to use the bootstrap, *The Jacknife, the Bootstrap and Other Resampling Plans*, B. Efron, Philadelphia: Society for Industrial and Applied Mathematics, 1982. The relationship between bootstrap methods and large sample approximations is drawn out in *The Bootstrap and the Edgeworth Expansion*, P. Hall, New York: Springer-Verlag, 1992, a book written at a rather advanced level.

[2] It is very important to realise that this is merely a construction to enable an approximation to be found. Whether a sample ever could be large in a particular application of a large sample approximation is irrelevant.

2.1. Convergence in probability to a constant and consistency

In many cases of interest the distributions of a sequence of statistics becomes *concentrated on a single point*, say c, as we pass through the sequence, increasing n. That is, $F_{S_n}(s)$ becomes closer and closer to a step function as n is increased, a step function which is zero up to c, and at c, jumps to 1. In this case we say that S_n *converges in probability* to the constant c.

Formally, we say that a sequence of (possibly vector valued) statistics converges in probability to a constant (possibly vector), c, if, for all[3] $\varepsilon > 0$,

$$\lim_{n\to\infty} P[\|S_n - c\| > \varepsilon] = 0,$$

that is, if for every $\varepsilon, \delta > 0$, there exists N (which typically depends upon ε and δ), such that for all $n > N$

$$P[\|S_n - c\| > \varepsilon] < \delta.$$

We then write $\text{plim}_{n\to\infty} S_n = c$, or, $S_n \xrightarrow{p} c$, and c is referred to as the *probability limit* of S_n.

When $S_n = \hat{\theta}_n$ is an *estimator* of a parameter, θ, which takes the value θ_0 and $\hat{\theta}_n \xrightarrow{p} \theta_0$, we say that $\hat{\theta}_n$ is a *consistent estimator.* Determining whether or not an estimator is consistent can be quite straightforward. If every member of the sequence $\{E\left[\hat{\theta}_n\right]\}_{i=1}^{\infty}$ and $\{Var\left[\hat{\theta}_n\right]\}_{i=1}^{\infty}$ exists, and

$$\begin{aligned} \lim_{n\to\infty} E\left[\hat{\theta}_n\right] &= \theta \\ \lim_{n\to\infty} Var\left[\hat{\theta}_n\right] &= 0 \end{aligned}$$

then we say that $\hat{\theta}_n$ *converges in mean square* to θ. It is quite easily shown that convergence in mean square implies convergence in probability[4]. It is often easy to derive expected values and variances of statistics. So a quick route to proving consistency is to prove convergence in mean square. Note, though, that an estimator can be consistent but *not* converge in mean square. There are commonly occurring cases in econometrics[5] where estimators are consistent but the sequences of moments required for consideration of convergence in mean square do not exist.

Consistency is generally regarded as a desirable property for an estimator to possess. Note though that in all practical applications of econometric methods we have a finite sized sample at our disposal. The consistency property on its own does not tell us about the quality of the estimate that we calculate using such a sample. It might be better sometimes to use an inconsistent estimator that generally takes values close to the unknown θ than a consistent estimator that is very inaccurate except at a much larger sample size than we have available. The consistency property does tell us that with a large enough sample our estimate would likely be close to the unknown truth, but not how close, nor even how large a sample is required to get an estimate close to the unknown truth. To get some idea of this we consider the concept of convergence in distribution.

[3] Here the notation $||\cdot||$ is used to denote the Euclidean length of a vector, that is: $\|z\| = (z'z)^{1/2}$. This is the absolute value of z when z is a scalar.

[4] A proof using Chebyshef's inequality is available in most of the intermediate textbooks.

[5] For example, the two stage least squares estimator in just identified linear models, i.e. the indirect least squares estimator.

2.2. Convergence in distribution

A sequence of statistics $\{S_n\}_{n=1}^{\infty}$ that converges in probability to a constant has a variance (if one exists) which becomes small as we pass to larger values of n. If we multiply S_n by a function of n, chosen so that the variance of the transformed statistic remains approximately constant as we pass to larger values of n, then we may obtain a sequence of statistics which converge not to a constant but to a *random variable.* If we can work out what the distribution of this random variable is, then we can use this distribution to approximate the distributions of the transformed statistics in the sequence, and in particular of the statistic we calculate using a particular finite sized sample. This approach can be formalised as follows. First we define convergence in distribution.

Consider a sequence of random variables $\{T_n\}_{n=1}^{\infty}$. Denote the distribution function of T_n by

$$P[T_n \leq t] = F_{T_n}(t).$$

Let T be a random variable with distribution function

$$P[T \leq t] = F_T(t).$$

We say that $\{T_n\}_{n=1}^{\infty}$ *converges in distribution* to T if for all $\varepsilon > 0$ there exists N (which will generally depend upon ε) such that for all $n > N$,

$$|F_{T_n}(t) - F_T(t)| < \varepsilon$$

at all points t at which $F_T(t)$ is continuous. Then we write $T_n \overset{d}{\to} T$. The definition applies for vector and scalar random variables. In this situation we will also talk in terms of T_n *converging in probability* to (the random variable) T.

Now return to the sequence $\{S_n\}_{n=1}^{\infty}$ that converges in probability to a constant. Let $T_n = h(n)(S_n)$ with $h(\cdot) > 0$ chosen so that $\{T_n\}_{n=1}^{\infty}$ *converges in distribution* to a random variable T that has a non-degenerate distribution. Usually the statistics S_n are centered so that their expectations converge to a constant value, and they are usually scaled so that the distribution of T is parameter free, if this can be done. A common case that will arise is that in which $h(n) = n^{\alpha}$. In this course we will only encounter the special case in which $\alpha = 1/2$, that is $h(n) = n^{1/2}$.

We can use the limiting random variable T to make approximate probability statements as follows. Since $S_n = T_n/h(n)$,

$$\begin{aligned} P[S_n \leq s] &= P[T_n/h(n) < s] \\ &= P[T_n < s \times h(n)] \\ &\simeq P[T < s \times h(n)] \\ &= F_T(s \times h(n)) \end{aligned}$$

which allows approximate probability statements concerning the random variable S_n.

Example 2.1. *Consider the mean, $\bar{X}_n$ of n independently and identically distributed random variables with common mean and variance respectively μ and σ^2. One of the simplest Central Limit Theorems (see below) says that, if $T_n = n^{1/2}(\bar{X}_n - \mu)/\sigma$ then $T_n \overset{d}{\to} T \sim N(0,1)$. We can* *use this result to say that* $T_n \simeq N(0,1)$ *where "*$\simeq$*" here means "is approximately distributed as". This sort of result can be used to make approximate probability statements. Since* T *has a standard normal distribution*

$$P[-1.96 \leq T \leq 1.96] = 0.95$$

and so, approximately,

$$P[-1.96 \leq \frac{n^{1/2}(\bar{X}_n - \mu)}{\sigma} \leq 1.96] \doteqdot 0.95$$

leading, if σ^2 *were known, to the approximate* 95% confidence interval *for* μ,

$$\{\bar{X}_n - 1.96\sigma/n^{1/2}, \bar{X}_n + 1.96\sigma/n^{1/2}\},$$

approximate in the sense that

$$P[\bar{X}_n - 1.96\sigma/n^{1/2} \leq \mu \leq \bar{X}_n + 1.96\sigma/n^{1/2}] \doteqdot 0.95$$

Approximate hypothesis tests, *that is tests whose size is approximately equal to some nominated size are produced in a similar way.*

It is *very important* to realise that in making this approximation there is *no* sense in which we ever think of the sample size actually becoming large. Sometimes the use of large sample approximations is criticised by saying that "the sample isn't large" or a particular form of approximation is criticised by saying "the sample size couldn't become large". These are both ignorant comments. These large sample approximations are just that, approximations, and they are as good or bad as the size of the error incurred in using the approximation. Of course that depends to some extent on the sample size but it also depends on other factors.

For example we know that when y given X is normally distributed the OLS estimator is exactly normally distributed conditional on X. For non-normal y, under some conditions, as we will see, the limiting distribution of an appropriately scaled OLS estimator is normal. The quality of that normal approximation depends upon the sample size, but also upon the extent of the departure of the distribution of y given X from normality and upon the disposition of the values of the covariates. For y close to normality the normal approximation to the distribution of the OLS estimator is good even at very small sample sizes.

So, the sequence $\{S_n\}_{n=1}^{\infty}$ indexed by the sample size is just a hypothetical construct in the context of which we can develop an approximation to the distribution of a statistic. Of primary interest is the extent to which, at the value of n that we have, the deviations $|F_{T_n}(t) - F_T(t)|$ are large or small. This is commonly studied by Monte Carlo simulation or by considering higher order approximations.

2.3. Functions of statistics - Slutsky's Theorem

The result embodied in *Slutsky's Theorem* is of great help in developing the limiting distributions of statistics. Slutsky's Theorem states that if T_n is a sequence of random variables that converges in probability to a constant c, and $g(\cdot)$ is a continuous function, then $g(T_n)$ converges in probability to $g(c)$. T_n can be a vector or matrix of random variables in which case c is a vector or matrix of constants. Sometimes c is called the *probability limit* of T_n.

A similar result holds for convergence to a random variable, namely that if T_n is a sequence of random variables that converges in probability to a random variable T, and $g(\cdot)$ is a continuous function, then $g(T_n)$ converges in probability to $g(T)$. For example, if

$$T_n' = \left[\begin{array}{ccc} T_n^{1\prime} & \vdots & T_n^{2\prime} \end{array} \right]$$

and

$$T_n \xrightarrow{d} T = \left[\begin{array}{ccc} T^{1\prime} & \vdots & T^{2\prime} \end{array} \right]'$$

then

$$T_n^1 + T_n^2 \xrightarrow{d} T^1 + T^2$$

with similar results for matrix and scalar products, inverses and so forth. Note that convergence in probability to a constant is essentially a special case of convergence in distribution in which the limiting distribution has all its probability mass concentrated on a single point. So above, if $T_n^1 \xrightarrow{d} T^1$ and $T_n^2 \xrightarrow{p} c$, then $T_n^1 + T_n^2 \xrightarrow{d} T^1 + c$ where we interpret the result as a random variable T^1 shifted in location by a constant c.

Often the limiting distribution of a statistic is obtained by writing the statistic as a function of simple components, obtaining the limiting distributions and/or probability limits of the components and then combining these using the preceding results. We will see this operation done with the OLS estimator shortly. First we introduce some theorems which enable us to state the limiting distributions of relatively simple expressions which appear as components of some econometric statistics.

2.4. Limit theorems

The theorems we need are called central limit (or sometimes just limit) theorems. We give two of these here.

The *Lindberg-Levy Central Limit Theorem* gives the limiting distribution of a mean of identically distributed random variables. The Theorem states that if $\{Y_i\}_{i=1}^{\infty}$ are mutually independent random (vector) variables each with expected value μ and positive definite covariance matrix Ω then if $\bar{Y}_n = n^{-1}\sum_{i=1}^{n} Y_i$,

$$n^{1/2}(\bar{Y}_n - \mu) \xrightarrow{d} Z, \qquad Z \sim N(0, \Omega).$$

Many of the statistics we encounter in econometrics can be expressed as means of *non-identically distributed* random vectors, whose limiting distribution is the subject of the *Lindberg-Feller Central Limit Theorem.* The Theorem states that if $\{Y_i\}_{i=1}^{\infty}$ are independently distributed random variables with $E[Y_i] = \mu_i$, $Var[Y_i] = \Omega_i$ with finite third moments and

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \frac{1}{n}\sum_{i=1}^{n} \mu_i = \bar{\mu}_n$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mu_i = \mu \qquad \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \Omega_i = \Omega,$$

where Ω is finite and positive definite, and for each j

$$\lim_{n\to\infty}\left(\sum_{i=1}^{n}\Omega_i\right)^{-1}\Omega_j = 0, \tag{2.1}$$

then

$$n^{1/2}(\bar{Y}_n - \bar{\mu}_n) \xrightarrow{d} Z, \qquad Z \sim N(0,\Omega).$$

2.5. The approximate distribution of the OLS estimator

Consider the OLS estimator $S_n = \hat{\beta}_n = (X_n'X_n)^{-1}X_n'y_n$ where we index $\hat{\beta}$ etc., by n to indicate that a sample of size n is involved. We know that when

$$y_n = X_n\beta + \varepsilon \qquad E[\varepsilon_n|X_n] = 0 \qquad Var[\varepsilon_n|X_n] = \sigma^2 I_n$$

then

$$E[\hat{\beta}_n|X_n] = \beta$$

and

$$Var[\hat{\beta}_n|X_n] = \sigma^2(X_n'X_n)^{-1} = n^{-1}\sigma^2(n^{-1}X_n'X_n)^{-1} = n^{-1}\sigma^2(n^{-1}\sum_{i=1}^{n}x_ix_i')^{-1}.$$

2.5.1. Consistency

First consider when the consistency property applies. To do this we have to consider the manner in which X_n varies as we pass down the sequence $\{S_n\}_{n=1}^{\infty}$. If the x_i's were independently sampled from some distribution such that $n^{-1}\sum_{i=1}^{n}x_ix_i' = n^{-1}X'X \xrightarrow{p} \Sigma_{xx} = E[xx']$, and if this matrix of expected squares and cross-products is non-singular then

$$\lim_{n\to\infty} Var[\hat{\beta}_n|X_n] = 0.$$

In this case $\hat{\beta}_n$ converges in mean square to β (recall that $E[\hat{\beta}|X] = \beta$), so $\hat{\beta}_n \xrightarrow{p} \beta$ and the OLS estimator is consistent.

Another way to proceed is to imagine that as n increases, a fixed matrix X_n is replicated over and over again, in which case after m replications when the matrix of covariate values is X_{mn} containing X_n replicated m times,

$$Var[\hat{\beta}_{mn}|X_{mn}] = m^{-1}\sigma^2(X_n'X_n)^{-1}$$

which passes to zero as m tends to infinity as long as $X_n'X_n$ is non-singular.

2.5.2. Limiting distribution

To make large sample approximate inference using the OLS estimator, consider the centred statistics

$$S_n = \hat{\beta}_n - \beta$$

and the associated scaled statistics

$$\begin{aligned} T_n &= n^{1/2} S_n \\ &= n^{1/2}(\hat{\beta}_n - \beta) \\ &= (n^{-1} X_n' X_n)^{-1} n^{-1/2} X_n' \varepsilon_n. \end{aligned}$$

Note that, in the last line, the functions of n multiply together to produce $n^{1/2}$ as in the second line. Thinking in terms of obtaining the x values by random sampling, under suitable conditions[6]

$$(n^{-1} X_n' X_n)^{-1} \xrightarrow{p} \Sigma_{xx}^{-1}.$$

Consider the term

$$n^{-1/2} X_n' \varepsilon_n = n^{-1/2} \sum_{i=1}^{n} x_i \varepsilon_i.$$

Let $R_i = x_i \varepsilon_i$ and note that

$$E[R_i] = 0, \qquad Var[R_i] = \sigma^2 x_i x_i'.$$

Under suitable conditions on the vectors x_i, the R_i's satisfy the conditions of the Lindberg-Feller Central Limit Theorem and we have

$$n^{-1/2} \sum_{i=1}^{n} R_i = n^{-1/2} X_n' \varepsilon_n \xrightarrow{d} N(0, \sigma^2 \Sigma_{xx}).$$

Finally, by Slutsky's Theorem

$$T_n = n^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \sigma^2 \Sigma_{xx}^{-1}).$$

We use this approximation to say that

$$n^{1/2}(\hat{\beta}_n - \beta) \simeq N(0, \sigma^2 \Sigma_{xx}^{-1}).$$

In practice of course σ^2 and Σ_{xx} are unknown and we replace them by estimates, e.g. $\hat{\sigma}_n^2$ and $n^{-1} X_n' X_n$. If these are consistent estimates then we can use Slutsky's Theorem to obtain the limiting distributions of the resulting statistics.

For example in testing the hypothesis $H_0 : R\beta = r$, we have already considered the statistic

$$S_n = (R\hat{\beta}_n - r)' \left(R \left(X_n' X_n \right)^{-1} R' \right)^{-1} (R\hat{\beta}_n - r)/\sigma^2$$

[6] In the previously circulated version Σ_{XX} was (incorrectly) not inverted on the right hand side of the next expression. Thanks to Albrecht Glitz for pointing this out.

where the subscript "n" is now appended to indicate the sample size under consideration. In the normal linear model, $S_n \sim \chi^2_{(j)}$. When y given X is non-normally distributed the limiting distribution result given above can be used, as follows.

Rewrite S_n as

$$S_n = \left(n^{1/2}(R\hat{\beta}_n - r)\right)' \left(R\left(n^{-1}X_n'X_n\right)^{-1}R'\right)^{-1}\left(n^{1/2}(R\hat{\beta}_n - r)\right)/\sigma^2.$$

Let P_n be such that

$$P_n\left(R\left(n^{-1}X_n'X_n\right)^{-1}R'\right)P_n' = I_j$$

and consider the sequence of random variables

$$T_n = \frac{n^{1/2}}{\sigma}P_n(R\hat{\beta}_n - r).$$

$T_n \xrightarrow{d} N(0, I_j)$ as long as $P_n \xrightarrow{p} P$ where $P'(R\Sigma_{xx}^{-1}R')P = I_j$. Application of the results on limiting distributions of functions of random variables given in Section 2.3 gives

$$T_n'T_n \xrightarrow{d} \chi^2_{(j)}.$$

Now

$$T_n'T_n = \frac{n}{\sigma^2}(R\hat{\beta}_n - r)'\left(R\left(n^{-1}X_n'X_n\right)^{-1}R'\right)^{-1}(R\hat{\beta}_n - r)$$

where we have used

$$P_n'P_n = \left(R\left(n^{-1}X_n'X_n\right)^{-1}R'\right)^{-1}.$$

Cancelling the terms involving n:

$$T_n'T_n = S_n \xrightarrow{d} \chi^2_{(j)}.$$

Finally, if $\hat{\sigma}^2_n$ is a consistent estimator of σ^2 then it can replace σ^2 in the formula for S_n and the approximate $\chi^2_{(j)}$ still applies, that is:

$$\left(n^{1/2}(R\hat{\beta}_n - r)\right)' \left(R\left(n^{-1}X_n'X_n\right)^{-1}R'\right)^{-1}\left(n^{1/2}(R\hat{\beta}_n - r)\right)/\hat{\sigma}^2_n \xrightarrow{d} \chi^2_{(j)}.$$

The other results we developed earlier for the normal linear model with "known" σ^2 also works as approximations when a normality restrictions does not hold and when σ^2 is replaced by a consistent estimator.

2.6. The approximate distribution of the GLS estimator

In Course Notes 2 we saw that in the model

$$\begin{aligned} y &= X\beta + \varepsilon \\ E[\varepsilon|X] &= 0 \\ Var[\varepsilon|X] &= \Omega \end{aligned}$$

the GLS estimator, $\tilde{\beta} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}y$ (note using Ω, not an estimate $\hat{\Omega}$) is BLU, and when y given X is normally distributed

$$\tilde{\beta} \sim N(\beta, \left(X'\Omega^{-1}X\right)^{-1}).$$

When y given X is non-normally distributed we can proceed as above, working in the context of a transformed model in which transformed y given X has an identity covariance matrix giving, under suitable conditions $\tilde{\beta} \xrightarrow{p} \beta$ and the limiting distribution

$$n^{1/2}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, \left(n^{-1}X'\Omega^{-1}X\right)^{-1}).$$

We noted that in practice Ω is unknown and suggested using a feasible GLS estimator, $\tilde{\beta} = \left(X'\hat{\Omega}^{-1}X\right)^{-1} X'\hat{\Omega}^{-1}y$ in which $\hat{\Omega}$ was some estimate of the conditional variance of y given X. Suppose $\hat{\Omega}$ is a *consistent* estimator of Ω. Then it can be shown that $\tilde{\beta}$ is a consistent estimator of β and under suitable conditions

$$n^{1/2}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, \left(n^{-1}X'\Omega^{-1}X\right)^{-1}).$$

When $\hat{\Omega}$ is a consistent estimator the limiting distribution of the feasible GLS estimator is the same as the limiting distribution of the estimator that employs Ω. Of course the exact distributions differ in a finite sized sample to an extent that depends upon the accuracy of the estimator of Ω in that finite sized sample.

When the elements of Ω are functions of a finite number of parameters it may be possible to produce a consistent estimator, $\hat{\Omega}$.

For example consider a heteroscedastic model in which Ω is diagonal with diagonal elements

$$\omega_{ii} = f(x_i, \gamma).$$

A first step OLS estimation produces residuals, $\hat{\varepsilon}_i$ and

$$E[\hat{\varepsilon}_i^2|X] = (M\Omega M)_{ii} = M_i'\Omega M_i = \omega_{ii}M_{ii}$$

where M_i' is the ith row of M and M_{ii} is the (i, i) element of M. This simplification follows from the diagonality of Ω and the idempotency of M. We can therefore write[7]

$$\frac{\hat{\varepsilon}_i^2}{M_{ii}} = f(x_i, \gamma) + u_i$$

where $E[u_i|X] = 0$, and under suitable conditions a nonlinear least squares estimation will produce a consistent estimator of γ, leading to a consistent estimator of Ω.

[7] Under suitable conditions on the evolution of the x_i's, $M_{ii} \to 1$, and the division by M_{ii} on the left hand side below can be (and routinely is) ignored.

2.7. The approximate distribution of M-estimators

We have already met two M-estimators - the OLS and the NLS estimator. Below we will study another M-estimator, the maximum likelihood estimator. It is difficult to develop exact distributions for these estimators, except under very special circumstances (e.g. for the OLS estimator with normally distributed y given X) and, as pointed out earlier, in practice an exact distributional result rarely leads to practical exact inference. So, we need to develop approximations to the distributions of M-estimators. As before we will study large sample approximations. This Section gives an overview of the strategy used to develop large sample approximations to the distributions of M-estimators. The next Section considers the particular case of the NLS estimator.

Consider an M-estimator defined as

$$\hat{\theta}_n = \arg\max_{\theta} U(Z_n, \theta)$$

where θ is a vector of parameters and Z_n is a vector random variable. In the applications we will consider Z_n contains n random variables representing outcomes observed in a sample of size n. We wish to obtain the limiting distribution of $\hat{\theta}_n$.

The first step is to show that $\hat{\theta}_n \xrightarrow{p} \theta_0$, the true value of θ. This is done by placing conditions on U and on the distribution of Z_n which ensure that (a) for θ in a neighbourhood of θ_0, $U(Z_n, \theta) \xrightarrow{p} U^*(\theta)$, (b) the sequence of values (indexed by n) of θ that maximise $U(Z_n, \theta)$ converges in probability to the value of θ that maximises $U^*(\theta)$, (c) the value of θ that uniquely maximises $U^*(\theta)$ is θ_0, the unknown parameter value. Condition (c) is essentially an identification condition.

To obtain the limiting distribution of $n^{1/2}\left(\hat{\theta}_n - \theta_0\right)$, consider situations in which the M-estimator can be defined as the unique solution to first order conditions

$$U_\theta(Z_n, \hat{\theta}_n) = 0$$

where

$$U_\theta(Z_n, \hat{\theta}_n) = \frac{\partial}{\partial\theta} U(Z_n, \theta)|_{\theta=\hat{\theta}_n}$$

This is certainly the case when $U(Z_n, \theta)$ is concave. We first consider a Taylor series expansion of $U(Z_n, \theta)$ regarded as a function of θ around $\theta = \theta_0$, as follows.

$$U_\theta(Z_n, \theta) = U_\theta(Z_n, \theta_0) + U_{\theta\theta}(Z_n, \theta_0)(\theta - \theta_0) + R(\theta, \theta_0, Z_n) \qquad (**)$$

Evaluating this at $\theta = \hat{\theta}_n$ (at which point $U_\theta(Z_n, \hat{\theta}_n) = 0$) gives

$$0 = U_\theta(Z_n, \hat{\theta}_n) = U_\theta(Z_n, \theta_0) + U_{\theta\theta}(Z_n, \theta_0)\left(\hat{\theta}_n - \theta_0\right) + R(\hat{\theta}_n, \theta_0, Z_n) \qquad (**)$$

where

$$U_{\theta\theta}(Z_n, \theta) = \frac{\partial^2}{\partial\theta\partial\theta'} U(Z_n, \theta).$$

The remainder term, $R(\hat{\theta}_n, \theta_0, Z_n)$, involves the third derivatives of $U(Z_n, \theta)$ and in many situations converges in probability to zero as n becomes large, in part because of the consistency of $\hat{\theta}_n$.

This allows us to write

$$U_\theta(Z_n,\theta_0)+U_{\theta\theta}(Z_n,\theta_0)\left(\hat{\theta}_n-\theta_0\right)\doteq 0 \qquad (**)$$

and then

$$\left(\hat{\theta}_n-\theta_0\right)\doteq -U_{\theta\theta}(Z_n,\theta_0)^{-1}U_\theta(Z_n,\theta_0).$$

Equivalently:

$$n^{1/2}\left(\hat{\theta}_n-\theta_0\right)\doteq -\left(n^{-1}U_{\theta\theta}(Z_n,\theta_0)\right)^{-1}n^{-1/2}U_\theta(Z_n,\theta_0).$$

In the situations we will encounter it is possible to find conditions under which

$$n^{-1}U_{\theta\theta}(Z_n,\theta_0)\xrightarrow{p}A(\theta_0) \qquad n^{-1/2}U_\theta(Z_n,\theta_0)\xrightarrow{d}N(0,B(\theta_0)),$$

for some matrices $A(\theta_0)$ and $B(\theta_0)$, concluding that

$$n^{1/2}\left(\hat{\theta}_n-\theta_0\right)\xrightarrow{d}N(0,A(\theta_0)^{-1}B(\theta_0)A(\theta_0)^{-1\prime}).$$

The details vary from case to case and the conditions under which the claimed convergence (in probability and distribution) occur are often complicated and will not be considered here. Of primary interest here are the forms taken by $A(\cdot)$ and $B(\cdot)$.

Consider the case of the OLS estimator:

$$\hat{\theta}_n=\arg\max_\theta\left\{-\sum_{i=1}^n\left(Y_i-x_i'\theta\right)^2\right\}$$

when $Y_i=x_i'\theta_0+\varepsilon_i$ and the ε_i's are independently distributed with expected value zero and common variance σ_0^2.

Note that OLS *maximises* the *negative* of the sum of squared residuals,

$$\begin{aligned}
U(Z_n,\theta) &= -\sum_{i=1}^n\left(Y_i-x_i'\theta\right)^2\\
n^{-1/2}U_\theta(Z_n,\theta) &= 2n^{-1/2}\sum_{i=1}^n(Y_i-x_i'\theta)x_i\\
n^{-1}U_{\theta\theta}(Z_n,\theta) &= -2n^{-1}\sum_{i=1}^n x_ix_i'
\end{aligned}$$

and, defining $\Sigma_{XX}\equiv\text{plim}_{n\to\infty}n^{-1}\sum_{i=1}^n x_ix_i'$:

$$A(\theta_0)=-2\Sigma_{XX}$$

which does not depend upon θ_0 in this special case,

$$B(\theta_0)=4\sigma_0^2\Sigma_{XX}$$

$$A(\theta_0)^{-1}B(\theta_0)A(\theta_0)^{-1\prime}=\sigma_0^2\Sigma_{XX}^{-1}$$

and finally the OLS estimator has the following limiting normal distribution.

$$n^{1/2}\left(\hat{\theta}_n-\theta_0\right)\xrightarrow{d}N(0,\sigma_0^2\Sigma_{XX}^{-1}).$$

2.8. (*) The approximate distribution of the NLS estimator

This Section studies a slightly more complicated case: the limiting distribution of the NLS estimator. At each step compare with the results above for the OLS estimator in the linear model.

The NLS estimator is defined as follows.

$$\hat{\theta}_n = \arg\max_{\theta} \left\{ -\sum_{i=1}^{n} \left(Y_i - g(x_i, \theta)\right)^2 \right\}$$

In the previous Section $g(x_i, \theta) = x_i'\theta$. There is:

$$\begin{aligned}
U(Z_n, \theta) &= -\sum_{i=1}^{n} \left(Y_i - g(x_i, \theta)\right)^2 \\
n^{-1/2} U_\theta(Z_n, \theta) &= 2n^{-1/2} \sum_{i=1}^{n} \left(Y_i - g(x_i, \theta)\right) g_\theta(x_i, \theta) \\
n^{-1} U_{\theta\theta}(Z_n, \theta) &= 2n^{-1} \sum_{i=1}^{n} \left(Y_i - g(x_i, \theta)\right) g_{\theta\theta}(x_i, \theta) - 2n^{-1} \sum_{i=1}^{n} g_\theta(x_i, \theta) g_\theta(x_i, \theta)'
\end{aligned}$$

where

$$\begin{aligned}
g_\theta(x_i, \theta) &\equiv \frac{\partial}{\partial \theta} g(x_i, \theta) \\
g_{\theta\theta}(x_i, \theta) &\equiv \frac{\partial^2}{\partial \theta \partial \theta'} g(x_i, \theta).
\end{aligned}$$

We know that

$$E\left[\left(Y_i - g(x_i, \theta_0)\right) g_{\theta\theta}(x_i, \theta_0) | X\right] = 0$$

and, under suitable conditions:

$$2n^{-1} \sum_{i=1}^{n} \left(Y_i - g(x_i, \theta_0)\right) g_{\theta\theta}(x_i, \theta_0) \xrightarrow{p} 0$$

so that

$$n^{-1} U_{\theta\theta}(Z_n, \theta_0) \xrightarrow{p} -2 \operatorname*{plim}_{n \to \infty} n^{-1} \sum_{i=1}^{n} g_\theta(x_i, \theta_0) g_\theta(x_i, \theta_0)' = A(\theta_0).$$

Now consider

$$n^{-1/2} U_\theta(Z_n, \theta_0) = 2n^{-1/2} \sum_{i=1}^{n} (Y_i - g(x_i, \theta_0)) g_\theta(x_i, \theta_0).$$

We know that

$$E\left[\left(Y_i - g(x_i, \theta_0)\right) g_\theta(x_i, \theta_0) | X\right] = 0.$$

Suppose that $Var[Y_i - g(x_i,\theta_0)|X] = \sigma_0^2$. Then

$$Var[n^{-1/2}U_\theta(Z_n,\theta_0)] = 4n^{-1}\sigma_0^2 \sum_{i=1}^{n} g_\theta(x_i,\theta_0)g_\theta(x_i,\theta_0)'$$

and we can find conditions under which

$$n^{-1/2}U_\theta(Z_n,\theta_0) \xrightarrow{p} N(0, 4\sigma_0^2 \operatorname*{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n} g_\theta(x_i,\theta_0)g_\theta(x_i,\theta_0)'),$$

so that, in the notation of the previous Section

$$B(\theta) = 4\sigma_0^2 \operatorname*{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n} g_\theta(x_i,\theta_0)g_\theta(x_i,\theta_0)'.$$

Finally, noting that

$$A(\theta_0)^{-1}B(\theta_0)A(\theta_0)^{-1\prime} = \sigma_0^2 \left(\operatorname*{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n} g_\theta(x_i,\theta_0)g_\theta(x_i,\theta_0)' \right)^{-1}$$

we obtain the limiting distribution of the NLS estimator as follows.

$$n^{1/2}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N(0, \sigma_0^2 \left(\operatorname*{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n} g_\theta(x_i,\theta_0)g_\theta(x_i,\theta_0)' \right)^{-1})$$

To check on this result, consider the case in which $g(x_i,\theta) = x_i'\theta$, when $g_\theta(x_i,\theta) = x_i$. In this case

$$\operatorname*{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n} g_\theta(x_i,\theta)g_\theta(x_i,\theta)' = \operatorname*{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n} x_i x_i{}' = \operatorname*{plim}_{n\to\infty} n^{-1}X'X$$

and we obtain the now familiar limiting distribution of the OLS estimator.

As an example suppose that $g(x,\theta) = \theta_0 + \theta_1 x^{\theta_2}$. In this case

$$g_\theta(x,\theta) = \begin{bmatrix} 1 \\ x^{\theta_2} \\ \theta_1 x^{\theta_2}\log(x) \end{bmatrix}$$

and if $\hat{\theta}$ denotes the nonlinear least squares estimator then the approximate variance of $n^{1/2}\left(\hat{\theta}_n - \theta\right)$ is as follows[8].

$$\sigma_0^2 \begin{bmatrix} 1 & \operatorname{plim}_{n\to\infty} n^{-1}\sum_{i=1}^n x^{\theta_2} & \operatorname{plim}_{n\to\infty} n^{-1}\sum_{i=1}^n \theta_1 x^{\theta_2}\log(x) \\ * & \operatorname{plim}_{n\to\infty} n^{-1}\sum_{i=1}^n x^{2\theta_2} & \operatorname{plim}_{n\to\infty} n^{-1}\sum_{i=1}^n \theta_1 x^{2\theta_2}\log(x) \\ * & * & \operatorname{plim}_{n\to\infty} n^{-1}\sum_{i=1}^n \theta_1^2 x^{2\theta_2}\log(x)^2 \end{bmatrix}^{-1}$$

[8] The matrix is symmetric and only elements on or above the leading diagonal are shown.

In practice one would estimate this using

$$\hat{\sigma}_0^2 \begin{bmatrix} 1 & n^{-1}\sum_{i=1}^n x^{\hat{\theta}_2} & n^{-1}\sum_{i=1}^n \hat{\theta}_1 x^{\hat{\theta}_2}\log(x) \\ * & n^{-1}\sum_{i=1}^n x^{2\hat{\theta}_2} & n^{-1}\sum_{i=1}^n \hat{\theta}_1 x^{2\hat{\theta}_2}\log(x) \\ * & * & n^{-1}\sum_{i=1}^n \hat{\theta}_1^2 x^{2\hat{\theta}_2}\log(x)^2 \end{bmatrix}^{-1}$$

where the $\hat{\theta}_i$'s are elements of the NLS estimator, $\hat{\theta}_n$, and $\hat{\sigma}_0^2$ is an estimator of the variance around the regression function, for example the mean of the squared NLS residuals,

$$\hat{\sigma}_0^2 = n^{-1}\sum_{i=1}^n (y_i - g(x_i,\hat{\theta}))^2.$$

2.9. Approximate distributions of functions of estimators - the "delta method"

Often we want to make inferences about some function of parameters, for example the ratio of two parameters. Value of time estimation provides a good example of this. If the indirect utility a person derives from action i (e.g. choosing plane route i) can be written as $U(\theta_1 c_i + \theta_2 t_i)$ where c_i is the cost incurred when action i is chosen and t_i is the time incurred when action i is chosen, then the ratio $v = \theta_2/\theta_1$ measures the value of time. (Why?)

Discrete choice modelling and estimation provide a route to getting estimates of θ_2 and θ_1 along with estimates of their accuracy and so we can compute an estimate of the value of time using[9] $\hat{v} = \hat{\theta}_2/\hat{\theta}_1$. How accurate will this estimate be, and what is its approximate distribution?

We proceed in a more general context in which we are interested in a scalar function of a vector of parameters, $h(\theta)$, and suppose that we have a consistent estimator $\hat{\theta}$ of θ whose approximate distribution is given by

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,\Omega)$$

where θ_0 is the data generating value of θ. What is the approximate distribution of $h(\hat{\theta})$?

Consider a Taylor series expansion of $h(\theta)$ around $\theta = \theta_0$ as follows

$$h(\theta) = h(\theta_0) + (\theta - \theta_0)' h_\theta(\theta_0) + \frac{1}{2}(\theta - \theta_0)' h_{\theta\theta}(\theta^*)(\theta - \theta_0)$$

where $h_\theta(\theta_0)$ is the vector of derivatives of $h(\theta)$ evaluated at $\theta = \theta_0$, $h_{\theta\theta}(\theta^*)$ is the matrix of second derivatives of $h(\theta)$ evaluated at $\theta = \theta^*$, a value between[10] θ and θ_0. Evaluate this at $\theta = \hat{\theta}$ and rearrange to give

$$n^{1/2}\left(h(\hat{\theta}) - h(\theta_0)\right) = n^{1/2}(\hat{\theta} - \theta_0)' h_\theta(\theta_0) + \frac{1}{2} n^{1/2}(\hat{\theta} - \theta_0)' h_{\theta\theta}(\hat{\theta}^*)(\hat{\theta} - \theta_0)$$

where $\hat{\theta}^*$ lies between $\hat{\theta}$ and θ_0. Since $\hat{\theta}$ is consistent, $\hat{\theta}^*$ must converge to θ_0 and if $h_{\theta\theta}(\theta_0)$ is bounded then the second term above disappears[11] as $n \to \infty$. So, we have[12]

$$n^{1/2}\left(h(\hat{\theta}) - h(\theta_0)\right) \xrightarrow{d} h_\theta(\theta_0)' Z$$

[9] In this Section the subscript "n" indicating the sample size is suppressed.

[10] In the sense that $||\theta^* - \theta_0|| < ||\theta - \theta_0||$.

[11] $n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,\Omega)$ and $(\hat{\theta} - \theta_0) \xrightarrow{p} 0$.

[12] In the previously circulated version, in the next expression, there was incorrectly a factor $n^{1/2}$ on the right hand side. Thanks to Albrecht Glitz for pointing this out.

where

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} Z \sim N(0, \Omega).$$

Using our result on linear functions of normal random variables

$$n^{1/2}\left(h(\hat{\theta}) - h(\theta_0)\right) \xrightarrow{d} N(0, h_\theta(\theta_0)'\Omega h_\theta(\theta_0)).$$

In the value of time example

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

and $h(\theta) = \theta_2/\theta_1$ leading to

$$h_\theta(\theta) = \begin{bmatrix} -\theta_2/\theta_1^2 \\ 1/\theta_1 \end{bmatrix}.$$

Write the approximate variance of $n^{1/2}(\hat{\theta} - \theta_0)$ as

$$\Omega = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{bmatrix}.$$

Then the approximate variance of $n^{1/2}\left(h(\hat{\theta}) - h(\theta_0)\right)$ is

$$\begin{aligned} \begin{bmatrix} -\theta_2/\theta_1^2 \\ 1/\theta_1 \end{bmatrix}' \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{bmatrix} \begin{bmatrix} -\theta_2/\theta_1^2 \\ 1/\theta_1 \end{bmatrix} &= \left(\theta_2^2/\theta_1^4\right)\omega_{11} - 2\left(\theta_2/\theta_1^3\right)\omega_{12} + \left(1/\theta_1^2\right)\omega_{22} \\ &= \left(1/\theta_1^2\right)\left(\left(\theta_2^2/\theta_1^2\right)\omega_{11} - 2\left(\theta_2/\theta_1\right)\omega_{12} + \omega_{22}\right) \end{aligned}$$

in which θ_1 and θ_2 are here taken to indicate the data generating values. Clearly if θ_1 is very close to zero then this will be large. Note that if θ_1 were actually zero then the development above would not go through because the condition on $h_{\theta\theta}(\theta_0)$ being bounded would be violated.

The method we have used here is sometimes called the "delta method".