Handout for Econ 624
Classical Nonlinear Econometric Models\textsuperscript{1}

Ingmar R. Prucha

5/12/2005

\textsuperscript{1}Personal notes. These notes are in draft form and provide brief summaries of important material presented in the lecture and/or discussion section. The notes are based in part on Pötscher and Prucha (1997). They are not written in the form of a textbook, and the material will be augmented in class with a more in depth discussion. The notes should not be distributed beyond the class room. Feedback, including feedback on typos, is welcome.
Contents

1 HISTORY AND SCOPE 1
2 MODELS, DATA GENERATING PROCESSES, AND ESTIMATORS 5
3 BASIC STRUCTURE OF THE CLASSICAL CONSISTENCY PROOF 10
4 CONSISTENCY: CATALOGUES OF ASSUMPTIONS 17
  4.1 Least Mean Distance Estimators ........................... 17
    4.1.1 General ........................................ 17
    4.1.2 Nonlinear Least Squares ................................ 18
    4.1.3 Maximum Likelihood ................................ 21
  4.2 Generalized Method of Moments Estimators .................. 23
5 BASIC STRUCTURE OF THE ASYMPTOTIC NORMALITY PROOF 25
  5.1 Outline of the Asymptotic Normality Proof .................. 25
  5.2 Basic Theorems for Asymptotic Normality ................... 27
  5.3 Proof of Basic Theorems for Asymptotic Normality .......... 30
6 ASYMPTOTIC NORMALITY: CATALOGUES OF ASSUMPTIONS 34
  6.1 Least Mean Distance Estimators ........................... 35
    6.1.1 General ........................................ 35
    6.1.2 Nonlinear Least Squares ................................ 38
    6.1.3 Maximum Likelihood ................................ 41
  6.2 Generalized Method of Moments Estimators .................. 44
7 NUMERICAL OPTIMIZATION 50
  7.1 Introduction ........................................... 50
  7.2 Examples of Gradient Methods ................................ 51
    7.2.1 Method of Steepest Descent ........................ 51
    7.2.2 Newton-Raphson Method .............................. 52
    7.2.3 Berndt, Hall, Hall and Hausman (BHHH) Method ...... 52
    7.2.4 Gauss-Newton Method ............................... 53
  7.3 Miscellaneous Remarks .................................... 54
A APPENDIX: FURTHER DISCUSSION OF EXAMPLES OF LMD AND GMM ESTIMATION 56
  A.1 LMD Estimation ........................................ 56
    A.1.1 ML Estimation of Linear and Nonlinear Regression Models 56
    A.1.2 ML Estimation of Discrete Response Models .......... 57
    A.1.3 ML Estimation of Censored Regression Models ...... 58
    A.1.4 ML Estimation of a Count Data Model ............... 59
A.1.5 NLS Estimation of a Linear Regression Model with Heteroskedastic Disturbances ............................................. 60

A.2 GMM Estimation .................................................................................. 61
   A.2.1 GMM Estimation of Linear and Nonlinear Regression Models ................................................................. 61
   A.2.2 GMM Estimation of Simultaneous Linear and Nonlinear Regression Models ........................................ 62
   A.2.3 Education Example ................................................................. 64
1 HISTORY AND SCOPE

Many estimators such as the least squares estimator, maximum likelihood estimators and generalized method of moments estimators are defined as the solution of a minimization (maximization) problem. They are often referred to as M-estimators or extremum estimators. A review of the literature shows that the proofs employed to demonstrate the consistency and the asymptotic normality of M-estimators have a quite similar structure. The basic methods used in these proofs have their origin in numerous contributions in the statistics literature. More specifically, these methods date back to articles by Doob (1934), Cramér (1946), Wald (1949), and LeCam (1953), who consider the maximum likelihood estimator in the case of independent and identically distributed (i.i.d.) data and to the analysis of the least squares estimator by Jennrich (1969) and Malinvaud (1970). In his seminal article Huber (1967) analyzes the asymptotic properties of M-estimators in the case of i.i.d. data processes and allows for certain types of misspecification. Hoadley (1971) considers the asymptotic properties of the maximum likelihood estimator for independent and not necessarily identically distributed data processes.

A review of the literature up to the beginning of the 1980s is given in Burguete, Gallant and Souza (1982) and Amemiya (1983). For an account of related contributions in the statistics literature see, e.g., Humak (1983). The theory reviewed in these references assumes that the model is essentially static in nature and that the data generating process exhibits a certain degree of temporal homogeneity, e.g., some form of stationarity. Developments in more recent years have focused on the extension of the theory to dynamic models, and in particular to situations where the data generating process can exhibit not only temporal dependence but also certain forms of temporal heterogeneity.

First progress towards a general theory for dynamic nonlinear econometric models was made by Bierens (1981, 1982a, 1984). His theory allows for temporal dependence in the data generating process and takes the dynamic structure of the model explicitly into account. Although Bierens does not take the data generating process to be stationary, his theory still assumes a certain degree of temporal homogeneity of the process. Bierens’ analysis focuses mainly on least squares and robust estimation of a nonlinear regression model. Hansen (1982) considers generalized method of moments estimators in the context of dynamic models under the stronger homogeneity assumption that the data generating process is stationary. Many economic data exhibit, besides temporal dependence, also temporal heterogeneity. Therefore the asymptotic properties of estimators under such conditions have been analyzed by Domowitz and White (1982), White and Domowitz (1984), Bates and White (1985) and Domowitz (1985). Although the results of the latter papers can in principle be applied to processes generated by certain dynamic models, the results are not genuinely geared towards such models. In particular, in specifying the dependence properties of the data generating process, these papers do not explicitly take into account the dynamic structure of the model. Furthermore, as pointed out by Andrews (1987) and Pötscher and Prucha (1986a,b), some of the maintained as-
sumptions in these papers are rather restrictive. Results by Wooldridge (1986), Gallant (1987a, Ch.7) and Gallant and White (1988) extend the theory of inference in dynamic nonlinear models to data generating processes that can exhibit both temporal dependence and heterogeneity.

Important ingredients in the typical proof of consistency and asymptotic normality of M-estimators are uniform laws of large numbers (ULLNs) and central limit theorems (CLTs). As a consequence, recent progress in the theory of inference in dynamic nonlinear models builds on progress in the derivation of ULLNs and CLTs. ULLNs for data generating processes, that are stationary or asymptotically stationary, are available in, e.g., LeCam (1953), Ranga Rao (1962), Jennrich (1969), Malinvaud (1970), Gallant (1977), Bierens (1981, 1982a, 1984, 1987), Amemiya (1985) and Pötscher and Prucha (1986a). \(^1\) Hoadley’s (1971) ULLN and its versions in Domowitz and White (1982) and White and Domowitz (1984) apply to temporally dependent and heterogeneous data generating processes. However, as pointed out by Andrews (1987) and Pötscher and Prucha (1986a,b) the maintained assumptions of this ULLN are restrictive (essentially requiring the random variables involved to be bounded). Andrews (1987) and Pötscher and Prucha (1986b, 1989, 1994b) introduce ULLNs for temporally dependent and heterogeneous processes under assumptions more appropriate for a theory of asymptotic inference in nonlinear econometric models. Furthermore, these papers specify the dependence structure in generic form in the sense that they assume the existence of laws of large numbers (LLNs) for certain “bracketing” functions of the data generating process, rather than to assume, e.g., a particular mixing property for the data generating process. As a consequence, these ULLNs are rather versatile tools that can be applied to processes with various dependence structures. Within the context of these ULLNs the dependence structure is relevant essentially only insofar as a LLN has to hold for the “bracketing” functions. The demonstration that a ULLN holds for a process with a particular dependence structure is therefore reduced to the demonstration that a LLN holds. For further results see also Andrews (1992), Newey (1991), and Pötscher and Prucha (1994a).

Since LLNs and CLTs are available for, e.g., \(\alpha\)-mixing and \(\phi\)-mixing processes it is tempting to simply postulate that the process of the endogenous and exogenous variables is \(\alpha\)-mixing or \(\phi\)-mixing. This approach is used, e.g., in Domowitz and White (1982), White and Domowitz (1984), Bates and White (1985) and Domowitz (1985). However, as already alluded to above, in case the data are generated by a dynamic nonlinear model this assumption is not satisfactory. This is so, since then the endogenous variables typically depend on the infinite history of the exogenous variables and the disturbances. Now, even if the exogenous variables and the disturbances are \(\alpha\)-mixing or \(\phi\)-mixing, the endogenous variables need not inherit the same property, since \(\alpha\)-mixing and \(\phi\)-mixing are not necessarily preserved by transformations which involve the infinite past, see, e.g., Ibragimov and Linnik (1971), Chernick (1981), Andrews (1984), Athreya

\(^1\)For independent random variables and in particular for i.i.d. random variables ULLNs have been established in the empirical process literature under much weaker conditions, see, e.g., Gänsler (1983) and Pollard (1984, 1990).
and Pantula (1986a,b), and Doukhan (1994). Hence, an assumption that the process of endogenous and exogenous variables is $\alpha$-mixing or $\phi$-mixing does not seem to be adequate for a general treatment of dynamic models. In fact, as discussed next, it is possible to build a theory of asymptotic inference without these mixing conditions.

Intuitively one can expect LLNs and CLTs to hold for (functions of) the data generating process, provided both the dynamic system (generating the endogenous variables) and the process of exogenous variables and disturbances have a sufficiently “fading memory”, even if the data generating process is not $\alpha$-mixing or $\phi$-mixing. This suggests that consistency and asymptotic normality results can also be obtained in such a context. The contributions of Bierens (1981, 1982a, 1984), Wooldridge (1986), Gallant (1987a, Ch.7), and Gallant and White (1988) can be viewed as a demonstration that this is indeed true under certain regularity conditions. The basic approach taken in all these references is to show that LLNs and CLTs hold for (functions of) the data generating process by demonstrating that the (functions of the) data generating process can be approximated by processes with a sufficiently fading memory. However, these references differ in the approximation concept employed: Bierens’ approximation concept leads to the definition of processes that are “stochastically stable w.r.t. an $\alpha$-mixing [or $\phi$-mixing] base”; using this approximation concept he proves LLNs and CLTs for such processes. Wooldridge (1986), Gallant (1987a, Ch.7) and Gallant and White (1988) employ the concept of “near epoch dependence w.r.t. an $\alpha$-mixing [or $\phi$-mixing] base”, and then make use of a result by McLeish (1975a) that processes with such a dependence structure fall into the class of mixingales, for which LLNs and CLTs are available in McLeish (1974, 1975a,b, 1977).

Pötscher and Prucha (1991a,b) recently provided a survey and a critical discussion of these developments towards an asymptotic theory for M-estimators in dynamic nonlinear models. The first of these papers, Pötscher and Prucha (1991a), also introduced an encompassing framework for the approaches taken by Bierens on the one hand and by Gallant, White and Wooldridge on the other hand. These approaches differ mainly in the employed approximation concepts, i.e., stochastic stability versus near epoch dependence. The relationship between these approaches had not been explored in the literature before and a clear understanding of their relative merits was lacking. Apart from providing an understanding of the differences and common grounds between these rival approaches, the encompassing framework of Pötscher and Prucha (1991a,b) also resulted in LLNs and CLTs under simpler and weaker sets of assumptions than in, e.g., Gallant (1987a) and Gallant and White (1988). In turn this lead to catalogues of assumptions for consistency and asymptotic normality of M-estimators in dynamic nonlinear models that seem to be simpler than corresponding catalogues in, e.g., Gallant (1987a) and Gallant and White (1988). Pötscher and Prucha (1991b) also provided improved consistency results for heteroskedastic-

---

2Results that ensure that $\alpha$-mixing or $\phi$-mixing is preserved by such transformations only seem to be available under conditions that are unnecessarily restrictive for a satisfactory general theory of inference in dynamic nonlinear models.
ity and autocorrelation robust variance covariance matrix estimators in case of near epoch dependent data. These improvements over results available in the literature related to less restrictive assumptions on the feasible rate of increase of the truncation lag parameter. A further novel feature was that the results provided rates of convergence for the variance covariance matrix estimators, which are essential for the optimal selection of the truncation lag parameter.


For reasons of keeping the discussion simple we will focus our discussion of an estimation theory for nonlinear models on the case where the data are i.i.d. Much of the handout consists of material from Pötscher and Prucha (1997), but with the material simplified as permitted by the i.i.d. assumption.

It should be noted that the monograph by Pötscher and Prucha (1997) was not written as a graduate text and was geared towards the professional econometrician and statistician. The subsequent handout summarizes important results and should only be viewed as to provide a basic structure for the discussion in class.

For textbook presentation of an estimation theory of nonlinear models see, e.g., Amemiya (1985), Bierens (1994) and Wooldridge (2002).
2 MODELS, DATA GENERATING PROCESSES, AND ESTIMATORS

We start with a brief review of the basic structure of the classical estimation problem, which can be described as follows: The researcher observes a set of data assumed to be generated by a stochastic process. The probability law of this process is determined by a model. This “true” model is assumed to belong to a class of models where each model is indexed by a parameter. This parameter may either characterize the probability law of the stochastic process completely or only partially (e.g., it may only characterize the first and second moments). Apart from knowing that the true model belongs to the given model class, the value of the true parameter is not known. The parameter may be an element of a finite or infinite dimensional space. The estimation problem is then to infer the value of the true parameter (or of certain components of interest) on the basis of the observed data. Specific estimators are often derived from general principles such as the maximum likelihood principle or the method of moments. Given a particular estimator it is then of interest to analyze its performance.

A crucial assumption in the estimation problem described above is that the data have been generated by a member of the class of models under consideration; that is, that the class of models under consideration contains the true model. The subsequent discussion will in essence continue to maintain this, admittedly, strong assumption. For an extension of an estimation theory towards misspecified models and a review of that literature see, e.g., the above cited monographs by Gallant, White, Pötscher and Prucha. We note that the assumption that the class of models under consideration contains the true model allows us to speak of true model parameters.

In the following we first formalize the above described general framework for M-estimators. We then illustrate this framework in terms of several estimators. In our discussion we will allow for the presence of nuisance parameters. However, to keep the discussion simple we will assume that the parameters are finite dimensional real vectors. Thus, at this stage of our discussion, we do not consider semi-parametric or non-parametric applications for which the parameter space is typically not a subset of a Euclidean space. Furthermore, to keep the presentation reasonably simple, we will focus the discussion on the case where the data generating process is i.i.d.. This rules out dynamic models. Again, for a more general estimation theory and review of the literature see, e.g., the above cited monographs by Gallant, White, Pötscher and Prucha.

More specifically, in the following we denote with \((z_i : i \in \mathbb{N})\) the data generating process defined on a probability space \((\Omega, \mathcal{A}, P)\) where the \(z_i\) are i.i.d. and take their values in \(Z \subseteq \mathbb{R}^{p_z}\), and we denote with \(B \subseteq \mathbb{R}^{p_\beta}\) and \(T \subseteq \mathbb{R}^{p_\tau}\) the space for the parameter vector of interest \(\beta\) and the space for the nuisance parameter vector \(\tau\), respectively. The sets \(Z, B\) and \(T\) are taken to be Borel sets. Now let \(Q_n(z_1, \ldots, z_n, \tau, \beta)\) be a real valued function defined on \(Z^n \times T \times B\) (where \(n\) denotes the sample size). Assume further that \(Q_n(z_1, \ldots, z_n, \tau, \beta)\) is
A\(\mathfrak{A}\)-measurable for all \((\tau, \beta) \in T \times B\). The M-estimators \(\hat{\beta}_n\) corresponding to the objective function \(Q_n\) for given estimators \(\hat{\tau}_n\) are then defined by:

\[
Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \hat{\beta}_n) = \inf_{\beta \in B} Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \beta),
\]

i.e., they minimize the objective function over \(B\).

The above framework covers typical M-estimators for linear and nonlinear models. We illustrate this in the following examples in terms of four examples, the nonlinear least squares (NLS) estimator, the normal full information maximum likelihood (NFIML) estimator and the nonlinear three stage least squares (N3SLS) estimator for an implicit nonlinear simultaneous equation system, and the probit/logit maximum likelihood (ML) estimators.

**Example 1:** Let \(g : X \times A \to \mathbb{R}\) be a Borel measurable function, where \(X \subseteq \mathbb{R}^{p_x}\) and \(A \subseteq \mathbb{R}^{p_a}\) are Borel sets. Assume that the endogenous variables \(y_i\) are generated according to the following model

\[
y_i = g(x_i, \alpha_0) + \epsilon_i, \quad i \in \mathbb{N}.
\]

Assume further that the processes of the exogenous variables \((x_i)\) and the disturbances \((\epsilon_i)\) are defined on \((\Omega, \mathfrak{A}, P)\) and take their values in \(X\) and \(\mathbb{R}\), respectively; let \(\alpha_0 \in A\) denote the true vector of regression parameters. The objective function of the NLS estimator is then given by

\[
Q_n(z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta),
\]

with

\[
q(z_i, \beta) = [y_i - g(x_i, \alpha)]^2
\]

where \(z_i = (y_i', x_i')'\) and \(\beta = \alpha\).

**Example 2:** Let \(f : Y \times X \times A \to E\) be a Borel measurable function, where \(Y \subseteq \mathbb{R}^{p_y}\), \(X \subseteq \mathbb{R}^{p_x}\), \(A \subseteq \mathbb{R}^{p_a}\), and \(E \subseteq \mathbb{R}^{p_e}\) are Borel sets (and \(p_y = p_e\)). Let the process of the endogenous variables \((y_i)\) be generated according to the following model

\[
f(y_i, x_i, \alpha_0) = \epsilon_i, \quad i \in \mathbb{N}.
\]

The processes of the exogenous variables \((x_i)\) and the disturbances \((\epsilon_i)\) are defined on \((\Omega, \mathfrak{A}, P)\) and take their values in \(X\) and \(E\), respectively; \(\alpha_0 \in A\) denotes the true vector of system parameters. We assume that the model has a well-defined reduced form, i.e., for each \((x, e, \alpha) \in X \times E \times A\) the equation

\[
f(y, x, \alpha) = e
\]

has a unique solution \(y = g(x, e, \alpha)\) where \(g\) is assumed to be measurable. This ensures that the process of endogenous variables \((y_i)\) is then

---

\(3\)In the following we will sometimes simply write \(Q_n(\tau, \beta)\) for \(Q_n(z_1, \ldots, z_n, \tau, \beta)\) whenever the meaning of this expression is evident from the context.
well-defined. We assume further that the vectors of disturbances \( \epsilon_i \) are distributed i.i.d. normal with zero mean and variance covariance matrix \( \Sigma_0 \) and that the process \( (\epsilon_i) \) is independent of the process \( (x_i) \). To define the NFIML estimator properly we assume further that \( f \) is continuously differentiable w.r.t. \( y \), that \( \partial f / \partial y \) is nonsingular, and that \( Y \) is open in \( \mathbb{R}^p \).

The objective function of the NFIML estimator, i.e., the normal log-likelihood function conditional on the exogenous variables is now (up to an additive constant and multiplied by \( -1/n \)) given by

\[
Q_n(z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta),
\]

with

\[
q(z_i, \beta) = -\ln |\det(\partial f_i / \partial y)| + (1/2) \ln \det(\Sigma) + (1/2) f_i \Sigma^{-1} f_i,
\]

where \( f_i \) and \( \partial f_i / \partial y \) are evaluated at \( (z_i, \alpha) \), \( z_i = (y_i', x_i')' \), and \( \beta \) is the vector composed of the elements of \( \alpha \) and the diagonal and upper diagonal elements of \( \Sigma \). Note that in (2.5) no nuisance parameter appears.

It may be helpful to write (2.4) in less compact notation. Let \( M = p_y = p_e \), then observe that the \( m \)-th equation of (2.4) can be written as \( (m = 1, \ldots, M) \)

\[
f_m^*(y_i, x_i, \alpha_{0m}) = f_m(y_i, x_i, \alpha_0) = \epsilon_{im}, \quad i \in \mathbb{N},
\]

where \( f_m \) and \( \epsilon_{im} \) denote the \( m \)-th element of \( f \) and \( \epsilon_i \), respectively, and \( \alpha_0 = [\alpha_{01}', \ldots, \alpha_{0M}']' \).

**Example 3:** Assume that the data are generated as in example 2. Now let \( a_i \) denote vectors of instruments such that the following moment conditions hold:

\[
E[\epsilon_{im} a_i] = 0
\]

for \( m = 1, \ldots, M \) and \( i = 1, \ldots, n \), or more compactly

\[
E[\epsilon_i \otimes a_i] = E[f(y_i, x_i, \alpha_0) \otimes a_i] = 0
\]

for \( i = 1, \ldots, n \). The objective function of the N3SLS estimator is then given by

\[
Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \beta)
\]

\[
= \left[n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right]^{-1} \left[ \sum_{i=1}^{n} a_i \hat{a}_i' \right]
\]

\[
\left[n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right]^{-1} \left[ \sum_{i=1}^{n} a_i \hat{a}_i' \right]
\]

with

\[
q(z_i, \beta) = f(y_i, x_i, \alpha) \otimes a_i,
\]

where \( \hat{\Sigma}_n \) is the two stage least squares estimator for \( \Sigma \), \( \hat{\tau}_n \) is the vector of diagonal and upper diagonal elements of \( [\hat{\Sigma}_n \otimes n^{-1} \sum_{i=1}^{n} a_i a_i']^{-1} \) which is assumed to exist, \( z_i = (y_i', x_i', a_i')' \), and \( \beta = \alpha \).
For further interpretation consider the case where $a_i = x_i$. Now let $X = [x'_1, \ldots, x'_n]'$ denote the matrix of observations on the exogenous variables and let $\varepsilon^{(m)} = [\varepsilon_1^{(m)}, \ldots, \varepsilon_{nm}]'$ denote the vector of disturbances in the $m$-th equation, then

$$n^{-1} \sum_{i=1}^{n} \varepsilon_i \otimes a_i = n^{-1} \sum_{i=1}^{n} \varepsilon_i \otimes x_i = \begin{bmatrix} n^{-1} \sum_{i=1}^{n} \varepsilon_{1i} x_i \\ \vdots \\ n^{-1} \sum_{i=1}^{n} \varepsilon_{Mi} x_i \end{bmatrix} = \begin{bmatrix} n^{-1} X' \varepsilon^{(1)} \\ \vdots \\ n^{-1} X' \varepsilon^{(M)} \end{bmatrix},$$

which puts the moment conditions in a more familiar form.

Example 4: Let $G : \mathbb{R} \to \mathbb{R}$ be a cumulative distribution function (cdf), let $(x_i)$ be a process of exogenous variables defined on $(\Omega, \mathcal{F}, P)$ that takes its values in $\mathbb{R}^{p_x}$ and let $A \subseteq \mathbb{R}^{p_x}$ be a Borel set (and $p_x = p_n$). Now consider the binary dependent variables $y_i$ defined by

$$P(\sigma_i = 1 \mid x_i) = G(x'_i \alpha),$$
$$P(\sigma_i = 0 \mid x_i) = 1 - G(x'_i \alpha).$$

The objective function of the (conditional) maximum likelihood (ML) is then given by

$$Q_n(z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta),$$

with

$$q(z_i, \beta) = y_i \ln[G(x'_i \alpha)] + (1 - y_i) \ln[1 - G(x'_i \alpha)]$$

where $z_i = (y'_i, x'_i)'$ and $\beta = \alpha$. Note, if $G$ is the cdf of a standardized normal distribution, then the above setup corresponds to a probit model, if $G$ is the cdf of a standard logistic distribution, then the above setup corresponds to a logit model.

Estimators corresponding to objective functions of, respectively, the form

$$Q_n(z_1, \ldots, z_n, \hat{\beta}) = n^{-1} \sum_{i=1}^{n} q(z_i, \hat{\beta})$$

and

$$Q_n(z_1, \ldots, z_n, \hat{\beta}) = \vartheta_n \left( n^{-1} \sum_{i=1}^{n} q(z_i, \hat{\beta}), \hat{\beta} \right),$$

where $\vartheta_n$ is some “distance” function, are usually referred to in the literature as least mean distance estimators and generalized method of moments estimators. We will focus our discussion on least mean distance estimators corresponding to objective functions of the form

$$Q_n(z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta)$$

8
For generalized method of moments estimators $\vartheta_n$ will often be a quadratic form in the moment vector $n^{-1} \sum_{i=1}^{n} q(z_i, \beta)$. We will focus our discussion on generalized method of moments estimators corresponding to objective functions of the form

$$Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \beta) = \left[ n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right] \left[ n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right]' \hat{\Xi}_n$$

(2.9)

where $\hat{\Xi}_n$ is a positive definite symmetric matrix and $\hat{\tau}_n$ is the vector of diagonal and upper diagonal elements of $\hat{\Xi}_n$. The NLS, NFIML and probit and logit ML estimators defined in the above examples are special cases of least mean distance estimators defined by (2.8), whereas the N3SLS estimator is a special case of a generalized method of moments estimator defined by (2.9). We also note that in general (quasi) maximum likelihood estimators can be viewed as least mean distance estimators defined by (2.8); in this case $q(z_i, \beta)$ corresponds to the negative of the conditional log-likelihood of unit (or in period) $i$. 
3 BASIC STRUCTURE OF THE CLASSICAL CONSISTENCY PROOF

In this section we describe the structure of the consistency proof for M-estimators in nonlinear econometric models, where no explicit expression exists for the estimator, as it has evolved from Jennrich (1969) and Malinvaud (1970). We shall refer to this proof as the classical consistency proof. The basic ideas date back to Doob (1934), Wald (1949) and Le Cam (1953). The consistency proofs in the articles on asymptotic inference in nonlinear econometric models listed in Section 1 all share this common structure.

We maintain the setup of Section 2. That is, as in Section 2 we denote with \( (z_i : i \in \mathbb{N}) \) the data generating process defined on a probability space \((\Omega, \mathfrak{A}, P)\) where the \( z_i \) are i.i.d. and take their values in \( Z \subseteq \mathbb{R}^p \), and we denote with \( B \subseteq \mathbb{R}^p \) and \( T \subseteq \mathbb{R}^p \) the space for the parameter vector of interest \( \beta \) and the space for the nuisance parameter vector \( \tau \), respectively. The sets \( Z, B \) and \( T \) are taken to be Borel sets. Let \( Q_n(z_1, \ldots, z_n, \tau, \beta) \) be a real valued criterion function defined on \( Z^n \times T \times B \) (where \( n \) denotes the sample size), where \( Q_n(z_1, \ldots, z_n, \tau, \beta) \) is \( \mathfrak{A} \)-measurable for all \((\tau, \beta) \in T \times B \). Furthermore, let \( \hat{\beta}_n \) be a corresponding M-estimator for the true parameter value \( \beta_0 \), and assume that \( \hat{\tau}_n \xrightarrow{p} \tau_0 \). The classical consistency proof then deduces the limiting behavior of \( \hat{\beta}_n \) from the limiting behavior of \( Q_n \). More specifically, under the setup adopted in Section 2 (and appropriate further assumptions) we will show that there exists a nonstochastic real valued function \( \bar{Q} \) defined on \( T \times B \) such that

\[
Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \beta) - \bar{Q}(\tau_0, \beta)
\]

converges to zero (in a sense specified later) as the sample size tends to infinity, and where \( \bar{Q}(\tau_0, \beta) \) is minimized at \( \beta_0 \).

Recall that we consider here the case where the data generating process is i.i.d. In this case the nonstochastic analogues of the objective function of the least mean distance estimator (2.8) and the generalized method of moments estimator (2.9) will be

\[
\bar{Q}(\beta) = \mathbb{E}q(z_i, \beta)
\]

and

\[
\bar{Q}(\tau_0, \beta) = [\mathbb{E}q(z_i, \beta)]' \Xi_0 [\mathbb{E}q(z_i, \beta)]
\]

where \( \Xi_0(\tau_0) = p \lim_{n \to \infty} \Xi_n(\hat{\tau}_n) \).

In essence, the structure of the classical consistency proof has two ingredients. In the case considered here, and assuming further that the parameter space \( B \subseteq \mathbb{R}^p \) is compact and that \( \bar{Q} \) is continuous (and no nuisance parameter is present) the two ingredients are:

- convergence of \( Q_n \) to \( \bar{Q} \) uniformly over the parameter space and
- the existence of a unique minimizer of \( \bar{Q} \) at the true parameter value.
The reason why it is necessary to require uniform and not just pointwise convergence can be easily demonstrated in terms of the following graph.

GRAPH

For explicit examples of functions that converge pointwise but not uniformly see, e.g., Amemiya (1985), pp. 108.

In more general cases where the nonstochastic analogue of $Q_n$, say $\bar{Q}_n$, also depends on the sample size essentially the same approach is employed subject to some modifications: As before, it is assumed that the difference between $Q_n$ and $\bar{Q}_n$ converges to zero uniformly over the parameter space. However, the assumption that the true parameter value is a unique minimizer of $\bar{Q}$ (together with continuity and compactness) is replaced by an assumption that ensures that the minimizers of $\bar{Q}_n$ are essentially unique, as well as that the functions $\bar{Q}_n$ do not become too flat at the minimizers. More formally, let $\bar{\tau}_n$ be a sequence of nonstochastic analogues of $\bar{\tau}_n$, and let $\bar{\beta}_n$ be a sequence of minimizers of $\bar{Q}_n(\bar{\tau}_n, \beta)$, then it is typically assumed that the sequence of minimizers $\bar{\beta}_n$ has the following property (where the existence of the minimizers is implicitly assumed):

**Definition 3.1**  
For a given sequence of functions $\bar{Q}_n : T \times B \to \mathbb{R}$, where $T \subseteq \mathbb{R}^p$ and $B \subseteq \mathbb{R}^q$ are Borel sets, and a given (nonstochastic) sequence $\bar{\tau}_n \in T$ the sequence of minimizers $\bar{\beta}_n$ of $\bar{Q}_n(\bar{\tau}_n, \beta)$ is called identifiably unique, if for every $\epsilon > 0$:

$$\liminf_{n \to \infty} \left[ \inf_{\beta \in B : |\beta - \bar{\beta}_n| \geq \epsilon} \bar{Q}_n(\bar{\tau}_n, \beta) - \bar{Q}_n(\bar{\tau}_n, \bar{\beta}_n) \right] > 0. \quad (3.1)$$

The above definition was introduced in White (1980) and Domowitz and White (1982).

**Remark:** Recall again that in our setup $\bar{Q}_n(\bar{\tau}_n, \beta)$ does not depend on the sample size, i.e.,

$$\bar{Q}_n(\bar{\tau}_n, \beta) \equiv \bar{Q}(\tau_0, \beta) \equiv \bar{R}(\beta)$$

In this case identifiable uniqueness of $\bar{\beta}_n \equiv \beta_0$ implies that $\beta_0$ is the unique minimizer of $\bar{R}$. If furthermore $B$ is compact and $\bar{R}$ is continuous, then identifiable uniqueness of $\beta_0$ is equivalent to uniqueness of $\beta_0$.

\[\text{4Of course, this definition also includes the case where no nuisance parameter is present.} \]

Also, we adopt the convention that the infimum over the empty set is plus infinity.
In the context of maximum likelihood estimation

\[ \bar{R}(\beta) = \bar{Q}(\beta) = \text{Eq}(z_i, \beta) \]

is the expected value of the log-likelihood function and the condition that \( \beta_0 \) is the unique minimizer of \( \bar{R} \) is equivalent to the condition that \( \beta_0 \) is identified in the usual sense.

In the context of generalized method of moments estimation

\[ \bar{R}(\beta) = \bar{Q}(\tau_0, \beta) = [\text{Eq}(z_i, \beta)]' \Xi_0 [\text{Eq}(z_i, \beta)] \]

and \( \text{Eq}(z_i, \beta_0) = 0 \). Assuming that \( \Xi_0 \) is positive definite the condition that \( \beta_0 \) is the unique minimizer of \( \bar{R} \) is equivalent to the condition that \( \beta_0 \) is a unique solution of the moment condition

\[ \text{Eq}(z_i, \beta) = 0. \]

The following lemma gives basic conditions for the convergence behavior of M-estimators in case where

\[ R_n(\omega, \beta) = Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \beta), \]

\[ \bar{R}(\beta) = \bar{Q}(\tau_0, \beta). \]

**Theorem 3.2** Let \( R_n : \Omega \times B \to \mathbb{R} \), where \( B \subseteq \mathbb{R}^p \) is a Borel set, be a sequence of functions, and where \( R_n(\omega, \beta) \) is \( \mathfrak{A} \)-measurable for each \( \beta \in B \). Furthermore, let \( \bar{R} : B \to \mathbb{R} \) be a function such that

\[ \sup_{\beta \in B} |R_n(\omega, \beta) - \bar{R}(\beta)| \to 0 \text{ a.s. [i.p.] as } n \to \infty. \]  

(3.2)

Suppose \( B \) is compact, \( \bar{R} \) is continuous on \( B \), and \( \beta_0 \) is the unique minimizer of \( \bar{R}(\beta) \) on \( B \), then for any sequence \( \hat{\beta}_n \), such that

\[ R_n(\omega, \hat{\beta}_n) = \inf_{\beta \in B} R_n(\omega, \beta) \]

(3.3)

holds, we have \( \hat{\beta}_n \to \beta_0 \) a.s. [i.p.] as \( n \to \infty \).\(^5\)

**Proof:** For some arbitrary \( \varepsilon > 0 \) let

\[ \delta = \inf_{\{\beta \in B : |\beta - \beta_0| \geq \varepsilon\}} \bar{R}(\beta) - \bar{R}(\beta_0). \]  

\(^5\)Sufficient conditions for \( \hat{\beta}_n \) to be measurable are discussed in Lemma 3.3. A simple sufficient condition for the measurability of suprema (or infima) like in (3.2) is that the functions over which the supremum (or infimum) is taken are continuous on \( B \subseteq \mathbb{R}^p \) for all \( \omega \in \Omega \) and that \( B \) is compact.
Since \( \bar{R}(\beta) \) is continuous and \( \{ \beta \in B : |\beta - \beta_0| \geq \varepsilon \} \) is compact it follows that there exist a minimum of \( \bar{R}(\beta) \) on that compact set. Since \( \beta_0 \) is the unique minimizer of \( \bar{R}(\beta) \) on \( B \) it follows that \( \delta > 0 \). In preparation of arguments used below, note that given this construction, the function \( \bar{R}(\beta) \) exceeds \( \bar{R}(\beta_0) \) by at least \( \delta > 0 \) for any \( \beta \in B \) with \( |\beta - \beta_0| \geq \varepsilon \). Now consider the event

\[
\Omega_n = \{ \omega \in \Omega : \sup_{\beta \in B} |R_n(\omega, \beta) - \bar{R}(\beta)| < \delta/4 \}.
\]

Then for \( \omega \in \Omega_n \) we have

\[
\begin{align*}
\bar{R}(\hat{\beta}_n) &< R_n(\omega, \hat{\beta}_n) + \delta/4 < R_n(\omega, \beta_0) + \delta/4, \\
R_n(\omega, \beta_0) &< \bar{R}(\beta_0) + \delta/4,
\end{align*}
\]

and hence

\[
\bar{R}(\hat{\beta}_n) < \bar{R}(\beta_0) + \delta/2.
\]

Thus for \( \omega \in \Omega_n \) it follows in light of (*) that \( |\hat{\beta}_n(\omega) - \beta_0| < \varepsilon \), and hence

\[
P(\Omega_n) \leq P\left( |\hat{\beta}_n(\omega) - \beta_0| < \varepsilon \right) \to 1
\]

in light of (3.2).

\[\square\]

**Remark:** Theorem 3.2 implies that, given \( \bar{R} \) is continuous on the compact set \( B \) and \( \beta_0 \) is the unique minimizer \( \bar{R} \), the classical consistency proof reduces to the verification that the objective function \( R_n(\omega, \beta) = Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \beta) \) satisfies the uniform convergence condition (3.2). For generalized versions of the above theorem and a review of the literature see, e.g., Pötscher and Prucha (1997), ch. 3.

**Remark:** In the case of least mean distance estimators, where no nuisance parameter is present, the objective function and its nonstochastic analogue are of the form

\[
\begin{align*}
R_n(\omega, \beta) &= n^{-1} \sum_{i=1}^{n} q(z_i, \beta), \\
\bar{R}(\beta) &= n^{-1} \sum_{i=1}^{n} Eq(z_i, \beta) = Eq(z_i, \beta).
\end{align*}
\]

In this case we can write the uniform convergence condition (3.2) as

\[
\sup_{\beta \in B} \left| n^{-1} \sum_{i=1}^{n} [q(z_i, \beta) - Eq(z_i, \beta)] \right| \to 0 \text{ a.s. } /l.p. \text{ as } n \to \infty
\]

and thus (3.2) boils down to a uniform law of large numbers (ULLN) for \( q(z_i, \beta) \).
In the case of generalized method of moments estimators, where a nuisance parameter is present, the objective function and its nonstochastic analogue are of the form

\[ R_n(\omega, \beta) = \frac{1}{n} \sum_{i=1}^{n} q(z_i, \beta) \hat{\Xi}_n \]

\[ \bar{R}(\beta) = \left[ n^{-1} \sum_{i=1}^{n} Eq(z_i, \beta) \right] ' \Xi_0 \left[ n^{-1} \sum_{i=1}^{n} Eq(z_i, \beta) \right] \]

In this case it follows as a consequence of Lemma 3.3 in Pötscher and Prucha (1997) that the uniform convergence condition (3.2) holds if again \( q(z_i, \beta) \) satisfies a ULLN, \( Eq(z_i, \beta) \) is finite and continuous on \( B \), where \( B \) is compact, and \( \hat{\Xi}_n \to \Xi_0 \) a.s. /i.p./ as \( n \to \infty \).

We next present a ULLN for functions of i.i.d. random variables; for a proof see Jennrich (1969), Theorem 2. This theorem is also given in the handout on Asymptotic Theory.

**Theorem 3.3** Let \( z_i \) be a sequence of identically and independently distributed random vectors taking their values in \( Z \subseteq \mathbb{R}^p_z \), where \( Z \) is a Borel set. Let \( B \) be a compact subset of \( \mathbb{R}^p_\beta \), and let \( q \) be a real valued function on \( Z \times B \). Furthermore let \( q(\cdot, \beta) \) be Borel measurable for each \( \beta \in B \), and let \( q(z, \cdot) \) be continuous for each \( z \in Z \). If \( E \sup_{\beta \in B} |q(z_i, \beta)| < \infty \), then

\[ \sup_{\beta \in B} \left( n^{-1} \sum_{i=1}^{n} [q(z_i, \beta) - Eq(z_i, \beta)] \right) \overset{a.s.}{\to} 0 \text{ as } n \to \infty. \]  \tag{3.4}

Furthermore \( Eq(z_i, \beta) \) is finite and continuous on \( B \).

**Remark:** As an aside we note that the theorem also holds if the assumption that \( z_i \) is i.i.d. is replaced by the assumption that \( z_i \) is stationary and ergodic; see, e.g., Pötscher and Prucha (1986), Lemma A.2. For further discussions of ULLNs and a review of the literature see Pötscher and Prucha (1997), ch. 5. For a discussion of uniform convergence and equicontinuity concepts of random functions see Pötscher and Prucha (1996).

In Theorem 3.2 we have implicitly assumed that minimizers \( \hat{\beta}_n \) of the objective function exist. The following lemma gives sufficient conditions for the existence and measurability of minimizers. (Of course, the a.s. part of Theorem 3.2 holds even if the \( \hat{\beta}_n \) are not measurable.) With \( \mathcal{B}(M) \) we denote the Borel \( \sigma \)-field on a (metrizable) space \( M \).
Lemma 3.3 Let $Z \subseteq \mathbb{R}^p$, $T \subseteq \mathbb{R}^r$, and $B \subseteq \mathbb{R}^{p_3}$ be Borel sets. Furthermore assume that $B$ is compact. Let $Q_n(z_1, \ldots, z_n, \tau, \cdot) = \inf_{\beta \in B} Q_n(z_1, \ldots, z_n, \tau, \beta)$ for each $(z_1, \ldots, z_n, \tau) \in Z^n \times T$ and let $Q_n(\cdot, \beta)$ be a $\mathcal{B}(Z^n) \otimes \mathcal{B}(T)$-$\mathcal{B}(\mathbb{R})$-measurable function on $Z^n \times T$ for each $\beta \in B$. Then there exists a $\mathcal{B}(Z^n) \otimes \mathcal{B}(T)$-$\mathcal{B}(\mathbb{R})$-measurable function $\hat{\beta}_n = \hat{\beta}_n(z_1, \ldots, z_n, \tau)$ such that for all $(z_1, \ldots, z_n, \tau) \in Z^n \times T$

\[ Q_n(z_1, \ldots, z_n, \tau, \hat{\beta}_n) = \inf_{\beta \in B} Q_n(z_1, \ldots, z_n, \tau, \beta) \]

holds.

Often we encounter situations where we would like to establish that some stochastic function evaluated at some consistent estimator converges to some nonstochastic analogue evaluated at the true parameter value. Such a situation arises, e.g., if we want to show that the Hessian of the log-likelihood function evaluated at some consistent estimator converges to the expected value of the Hessian evaluated at the true parameter value. To economize on notation, we use again $R_n(\omega, \beta)$ for the stochastic function, $\bar{R}(\beta)$ for the nonstochastic analogue, and $\beta$ for the parameter vector. However, we emphasize that in the subsequent theorem $R_n$, $\bar{R}$ and $\beta$ should no longer be viewed as to represent necessarily the same quantities as in the above discussion.

Theorem 3.4 Let $R_n : \Omega \times B \to \mathbb{R}$, where $B \subseteq \mathbb{R}^{p_3}$ is a Borel set, be a sequence of functions, and where $R_n(\cdot, \beta)$ is $\mathfrak{A}$-measurable for each $\beta \in B$. Furthermore, let $\bar{R} : B \to \mathbb{R}$ be a function such that

\[ \sup_{\beta \in B} |R_n(\omega, \beta) - \bar{R}(\beta)| \to 0 \text{ a.s. [i.p.] as } n \to \infty. \] \hspace{1cm} (3.5)

Suppose $\bar{R}$ is continuous and $\hat{\beta}_n \to \beta_0$ a.s. [i.p.] as $n \to \infty$, then

\[ R_n(\omega, \hat{\beta}_n) \to \bar{R}(\beta_0) \text{ a.s. [i.p.] as } n \to \infty. \]

---

6The proof in Jennrich (1969) seems to be in error.
Proof: We prove the i.p. part. Clearly for every $\varepsilon > 0$ and $\delta > 0$ we have

$$P\left(\left| R_n(\omega, \hat{\beta}_n) - \bar{R}(\beta_0) \right| > \varepsilon \right) \leq P\left(\left| R_n(\omega, \hat{\beta}_n) - \bar{R}(\hat{\beta}_n) \right| > \varepsilon /2 \right) + P\left(\left| \bar{R}(\hat{\beta}_n) - \bar{R}(\beta_0) \right| > \varepsilon /2 \right)$$

$$= P\left(\left| R_n(\omega, \hat{\beta}_n) - \bar{R}(\hat{\beta}_n) \right| > \varepsilon /2, |\hat{\beta}_n - \beta_0| < \delta \right)$$

$$+ P\left(\left| R_n(\omega, \hat{\beta}_n) - \bar{R}(\hat{\beta}_n) \right| > \varepsilon /2, |\hat{\beta}_n - \beta_0| \geq \delta \right)$$

$$+ P\left(\left| \bar{R}(\hat{\beta}_n) - \bar{R}(\beta_0) \right| > \varepsilon /2 \right)$$

$$\leq P\left(\sup_{\beta \in B} |R_n(\omega, \beta) - \bar{R}(\beta)| > \varepsilon /2 \right)$$

$$+ P\left(|\hat{\beta}_n - \beta_0| \geq \delta \right)$$

$$+ P\left(\left| \bar{R}(\hat{\beta}_n) - \bar{R}(\beta_0) \right| > \varepsilon /2 \right).$$

Now the first term in the last inequality goes to zero in light of (3.5). The second term goes to zero since $\hat{\beta}_n \to \beta_0$ i.p. by assumption, and the last term goes to zero by Slutsky’s theorem again since $\hat{\beta}_n \to \beta_0$ i.p. by assumption and $\bar{R}$ is continuous.

We would again try to use a ULLN to establish the uniform convergence condition (3.5).
4 CONSISTENCY: CATALOGUES OF ASSUMPTIONS

We continue to maintain the general setup of Section 2: We denote with $(z_i : i \in \mathbb{N})$ the data generating process defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where the $z_i$ are i.i.d. and take their values in $Z \subseteq \mathbb{R}^{p_z}$, and we denote with $B \subseteq \mathbb{R}^{p_\beta}$ and $T \subseteq \mathbb{R}^{p_\tau}$ the space for the parameter vector of interest $\beta$ and the space for the nuisance parameter vector $\tau$, respectively. The sets $Z$, $B$ and $T$ are taken to be Borel sets. Now let $Q_n(z_1, ..., z_n, \tau, \beta)$ be a real valued function defined on $Z^n \times T \times B$ (where $n$ denotes the sample size). Assume further that $Q_n(z_1, ..., z_n, \tau, \beta)$ is $\mathcal{F}$-measurable for all $(\tau, \beta) \in T \times B$. The M-estimators $\hat{\beta}_n$ corresponding to the objective function $Q_n$ for given estimators $\hat{\tau}_n$ are then defined by:

$$Q_n(z_1, ..., z_n, \hat{\tau}_n, \hat{\beta}_n) = \inf_{\beta \in B} Q_n(z_1, ..., z_n, \hat{\tau}_n, \beta),$$

i.e., they minimize the objective function over $B$.

In the previous section we have discussed the basic structure of the classical consistency proof for M-estimators and have established basic modules that can be employed for consistency proofs of M-estimators. In the following we use those modules to give specific low-level catalogues of assumptions for the consistency of least mean distance and generalized method of moments estimators.

4.1 Least Mean Distance Estimators

4.1.1 General

As discussed, estimators corresponding to objective functions of the form

$$R_n(\omega, \beta) = Q_n(z_1, ..., z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta)$$

(4.1)

where $q(z, \beta)$ is a real valued function defined on $Z \times B$ are typically called least mean distance estimators. In the following we will give a general catalogue of assumptions for the consistency of least mean distance estimators for the case of i.i.d. data.

Assumption 4.1 (a) $q(., \cdot)$ is a real valued function on $Z \times B$, where $Z \subseteq \mathbb{R}^{p_z}$ and $B \subseteq \mathbb{R}^{p_\beta}$ are Borel sets. (b) $B$ is furthermore compact. (c) $q(., \beta)$ is Borel measurable for each $\beta \in B$, and $q(z, .)$ is continuous for each $z \in Z$. (d) $z_i$ is a sequence of identically and independently distributed random vectors that take their values in $\mathbb{R}^{p_z}$. (e) $E \sup_{\beta \in B} |q(z_i, \beta)| < \infty$.

We now have the following general result concerning the consistency of least mean distance estimators in case of i.i.d. data.
Theorem 4.1 Consider the least mean distance estimators \( \hat{\beta}_n \) defined as minimizers of (4.1). Suppose Assumption 4.1 holds and \( \beta_0 \) is a unique minimizer of \( Eq(z_i, \beta) \), then

\[ \hat{\beta}_n \to \beta_0 \text{ a.s. as } n \to \infty. \]

Proof: Given Assumption 4.1 it follows directly from the ULLN given as Theorem 3.3 that

\[ \sup_{\beta \in B} \left| n^{-1} \sum_{i=1}^{n} [q(z_i, \beta) - Eq(z_i, \beta)] \right| = \sup_{\beta \in B} \left| R_n (\omega, \beta) - \bar{R} (\beta) \right| \to 0 \text{ a.s. as } n \to \infty \]

with \( \bar{R} (\beta) = Eq(z_i, \beta) \), and furthermore that \( \bar{R} (\beta) \) is continuous. Since by assumption \( \beta_0 \) is a unique minimizer of \( \bar{R} (\beta) \) it follows further from Theorem 3.2 that \( \hat{\beta}_n \to \beta_0 \) a.s. as \( n \to \infty \).

Recall that we noted in Section 3 that the ULLN given as Theorem 3.3 also holds if the assumption that \( z_i \) is i.i.d. is replaced by the assumption that \( z_i \) is stationary and ergodic. Thus the above consistency also holds if the i.i.d. assumption is relaxed to the assumption that \( z_i \) is stationary and ergodic.

4.1.2 Nonlinear Least Squares

Consider the i.i.d. data process \( z_i = [y_i, x_i] \) with \( y_i \in \mathbb{R} \) and \( x_i \in \mathbb{R}^p \). Frequently it will be of interest to model \( E[y_i \mid x_i] \) as a function of the explanatory variables \( x_i \). Now let \( g(x_i, \beta) \) be a parametric model for \( E[y_i \mid x_i] \), where \( g: \mathbb{R}^p \times B \to \mathbb{R} \) and \( B \subseteq \mathbb{R}^\beta \). We have a correctly specified model for the conditional mean if

\[ E[y_i \mid x_i] = g(x_i, \beta_0) \]

for some \( \beta_0 \in B \).

Define \( \epsilon_i = y_i - g(x_i, \beta_0) \), then we can think of the data as being described by the following nonlinear regression model

\[ y_i = g(x_i, \beta_0) + \epsilon_i, \quad i \in \mathbb{N}, \quad (4.2) \]

where the disturbances \( \epsilon_i \in \mathbb{R} \) satisfy \( E[\epsilon_i \mid x_i] = 0 \). The objective function of the NLS estimator is now given by

\[ Q_n (z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta), \quad (4.3) \]

with

\[ q(z_i, \beta) = [y_i - g(x_i, \beta)]^2. \]

Remark: We will demonstrate that under a reasonable set of assumptions the NLS estimator for \( \beta_0 \) is consistent, i.e., nonlinear least squares yields a consistent estimate of the conditional mean. For further interpretation we note that
for arbitrary \( m(x_i) \), provided the existence of respective expectations, we have

\[
E \{y_i - m(x_i)\}^2 = E \{y_i - E[y_i \mid x_i]\}^2 + E \{E[y_i \mid x_i] - m(x_i)\}^2
\]

\[
+ 2E \{[y_i - E[y_i \mid x_i]] [E[y_i \mid x_i] - m(x_i)]\}
\]

\[
= E \{y_i - E[y_i \mid x_i]\}^2 + E \{E[y_i \mid x_i] - m(x_i)\}^2
\]

since

\[
E \{[y_i - E[y_i \mid x_i]] [E[y_i \mid x_i] - m(x_i)]\} = 0.
\]

Hence

\[
E \{y_i - m(x_i)\}^2 \geq E \{y_i - E[y_i \mid x_i]\}^2
\]

and the inequality is strict unless \( E \{E[y_i \mid x_i] - m(x_i)\}^2 = 0 \).

**Remark:** Given that \( E[y_i \mid x_i] = g(x_i, \beta_0) \) it follows that

\[
E \{y_i - g(x_i, \beta)\}^2
\]

is minimized for \( \beta = \beta_0 \). However, this does not ensure that \( \beta_0 \) is a unique minimizer. Consequently the following additional condition is typically postulated:

\[
E \{g(x_i, \beta_0) - g(x_i, \beta)\}^2 > 0 \text{ for } \beta \neq \beta_0.
\]

We will show in the proof of the consistency result that this condition indeed ensures that \( \beta_0 \) is a unique minimizer. For interpretation observe that in case \( g(x_i, \beta) = x_i \beta \) we have

\[
E \{g(x_i, \beta_0) - g(x_i, \beta)\}^2 = (\beta_0 - \beta)'EX_i x_i (\beta_0 - \beta) > 0 \text{ for } \beta \neq \beta_0
\]

if \( EX_i x_i \) has full rank.

We now specify a list of assumptions that ensures the consistency of the NLS estimator. We note that the assumptions do not maintain that the explanatory variables \( x_i \) and the disturbances \( \epsilon_i \) are independent.

**Assumption 4.2** (a) \( g(., .) \) is a real valued function on \( \mathbb{R}^{p_x} \times B \), (b) \( B \subseteq \mathbb{R}^{p_\beta} \) is compact. (c) \( g(., \beta) \) is Borel measurable for each \( \beta \in B \), and \( g(x_i, .) \) is continuous for each \( x_i \in \mathbb{R}^{p_x} \). (d) \( z_i = [y_i, x_i] \) with \( y_i \in \mathbb{R} \) and \( x_i \in \mathbb{R}^{p_x} \) is a sequence of identically and independently distributed random vectors. (e) \( E[y_i \mid x_i] = g(x_i, \beta_0) \) for \( \beta_0 \in B \). (f) \( E\{y_i - g(x_i, \beta_0)\}^2 < \infty \) and \( E \sup_{\beta \in B} g(x_i, \beta)^2 < \infty \).

We now have the following result concerning the consistency of the nonlinear least squares estimators in case of i.i.d. data.
Theorem 4.2 Consider the nonlinear least squares estimators \( \hat{\beta}_n \) defined as minimizers of (4.3). Suppose Assumption 4.2 holds and \( E \left[ g(x_i, \beta_0) - g(x_i, \beta) \right]^2 > 0 \) for \( \beta \neq \beta_0 \), then
\[
\hat{\beta}_n \to \beta_0 \text{ a.s. as } n \to \infty.
\]

Proof: To prove the theorem we first verify that Assumptions 4.2 imply Assumptions 4.1 for
\[
q(z, \beta) = [y - g(x, \beta)]^2
\]
with \( z = [y, x] \), \( z_i = [y_i, x_i] \) and \( Z = \mathbb{R}^{p_x} \), \( p_z = p_x + 1 \). Clearly Assumptions 4.2(a),(b),(c),(d) imply Assumptions 4.1(a),(b),(c),(d). Next observe that, using inequality 1.4.3 in Bierens (1994),
\[
E y_i^2 = E \left[ g(x_i, \beta_0) + \epsilon_i \right]^2
\leq 2 \left[ E g(x_i, \beta_0)^2 + E \epsilon_i^2 \right]
\leq 2 \sup_{\beta \in B} g(x_i, \beta)^2 + E \epsilon_i^2 < \infty
\]
and
\[
E \sup_{\beta \in B} |q(z_i, \beta)| = E \sup_{\beta \in B} \left| y_i - g(x_i, \beta) \right|^2
\leq E \sup_{\beta \in B} 2 \left[ y_i^2 + g(x_i, \beta)^2 \right]
= 2E y_i^2 + 2E \sup_{\beta \in B} g(x_i, \beta)^2 < \infty
\]
in light of Assumptions 4.2(f). Thus also Assumption 4.1(e) holds.

Now observe that
\[
E q(z_i, \beta) = E \left[ y_i - g(x_i, \beta) \right]^2 = E \left[ \left| y_i - g(x_i, \beta_0) \right| + \left| g(x_i, \beta_0) - g(x_i, \beta) \right| \right]^2
\leq E \left[ \epsilon_i + \left| g(x_i, \beta_0) - g(x_i, \beta) \right| \right]^2
= E \epsilon_i^2 + 2E \epsilon_i \left[ g(x_i, \beta_0) - g(x_i, \beta) \right] + E \left[ g(x_i, \beta_0) - g(x_i, \beta) \right]^2
\leq E \epsilon_i^2 + E \left[ g(x_i, \beta_0) - g(x_i, \beta) \right]^2
\]
since \( E \epsilon_i \left[ g(x_i, \beta_0) - g(x_i, \beta) \right] = 0 \) given that \( E[\epsilon_i | x_i] = 0 \) by Assumption 4.2(e). Furthermore, since by assumption \( E \left[ g(x_i, \beta_0) - g(x_i, \beta) \right]^2 \neq 0 \) for \( \beta \neq \beta_0 \) it follows that \( \beta_0 \) is the unique minimizer of \( E q(z_i, \beta) \). Thus all assumptions of Theorem 4.1 are satisfied and hence it follows from that theorem that \( \hat{\beta}_n \to \beta_0 \) a.s. as \( n \to \infty \).

For a trick that avoids having to maintain the existence of finite second moments for \( \epsilon_i \) see, e.g., Pötscher and Prucha (1997) ch. 4.1.
4.1.3 Maximum Likelihood

Consider the i.i.d. data process \( z_i = [y_i, x_i] \) with \( y_i \in Y \subseteq \mathbb{R}^{p_y} \) and \( x_i \in X \subseteq \mathbb{R}^{p_x} \), where \( Y \) and \( X \) are Borel sets. Let \( F(y \mid x) \) denote the conditional distribution of \( y \) given \( x = x_i \). We assume that this distribution can be described by a conditional density, \( f(y \mid x; \beta_0) \), with respect to some \((\sigma\text{-finite)} measure \( \mu \) defined on the \( Y \), and \( \beta_0 \in B \subseteq \mathbb{R}^{p_\beta} \).

The objective function of the conditional maximum likelihood estimator is then given by

\[
Q_n(z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta), \tag{4.4}
\]

where

\[
q(z_i, \beta) = - \ln [f(y_i \mid x_i; \beta)].
\]

now represents the negative conditional log-likelihood for observation \( i \). We maintain the following assumptions.

**Assumption 4.3** The data process \( z_i = [y_i, x_i] \) is i.i.d. with \( y_i \in Y \subseteq \mathbb{R}^{p_y} \) and \( x_i \in X \subseteq \mathbb{R}^{p_x} \), where \( Y \) and \( X \) are Borel sets. Let \( B \subseteq \mathbb{R}^{p_\beta} \) and consider the family of conditional densities \( \{ f(y \mid x; \beta) : \beta \in B, x \in X \} \) with respect to some \((\sigma\text{-finite}) measure \( \mu \). Then for all \( x \in X \) the (true) conditional distribution of \( y_i \) given \( x_i = x \) can be represented by the conditional density \( f(y \mid x; \beta_0) \) for some \( \beta_0 \in B \). Consider the real valued function \( q \) on \( Z \times B, Z = Y \times X \), defined as \( q(z, \beta) \equiv - \ln f(y \mid x; \beta) \) for all \( z = [y, x] \in Y \times X \subseteq \mathbb{R}^{p_z}, p_z = p_y + p_x \), and all \( \beta \in B \), then: (a) \( B \subseteq \mathbb{R}^{p_\beta} \) is compact. (b) \( q(\cdot, \beta) \) is Borel measurable for each \( \beta \in B \), and \( q(z, \cdot) \) is continuous for each \( z \in Z \). (c) \( \mathbb{E} \sup_{\beta \in B} |q(z, \beta)| < \infty \).

We now have the following result concerning the consistency of conditional maximum likelihood estimators in case of i.i.d. data. Of course, the result also covers (unconditional) maximum likelihood estimators as a special case.

**Theorem 4.3** Consider the conditional maximum likelihood estimators \( \hat{\beta}_n \) defined as minimizers of (4.4). Suppose Assumption 4.3 holds and \( \beta_0 \) is a unique minimizer of \( Eq(z_i, \beta) \), then

\[
\hat{\beta}_n \to \beta_0 \quad \text{a.s. as} \quad n \to \infty.
\]

**Proof:** Clearly Assumptions 4.3 implies Assumptions 4.1 and thus the theorem follows from Theorem 4.1. \[\blacksquare\]
Remark: The above theorem assumes that $\beta_0$ is a unique minimizer of $Eq(z_i, \beta)$. We note that, as shown by the next lemma, for maximum likelihood estimators a sufficient condition for a unique maximum is that the parameter is identified. For simplicity consider the case where $z_i = y_i$, and thus

$$Eq(z_i, \beta) = -E \ln [f(y_i; \beta)] = -\int \ln [f(y; \beta)] [f(y; \beta_0)] d\mu.$$  

Lemma: (Information Inequality) Suppose Assumption 4.3 holds and $z_i = y_i$. If $\beta_0$ is identified in the sense that if $\beta \neq \beta_0$, then $f(y; \beta) \neq f(y; \beta_0)$ on a set with positive measure $\mu$, then $\beta_0$ is the unique minimizer of $q(z_i, \beta)$.

Proof: Recall that for any non-constant random variable $v$ and any strictly concave function $s(v)$ we have by the strict version of Jensen inequality that $s(E(v)) > E[s(v)]$. In particular, $\ln[E(v)] > E[\ln(v)]$ or $E[-\ln(v)] > -\ln[E(v)]$. Hence for $\beta \neq \beta_0$ we have

$$K[f(\cdot, \beta_0), f(\cdot, \beta)] = \int \ln \left[ \frac{f(y; \beta_0)}{f(y; \beta)} \right] [f(y; \beta_0)] d\mu$$

$$= \int -\ln \left[ \frac{f(y; \beta)}{f(y; \beta_0)} \right] [f(y; \beta_0)] d\mu$$

$$= E \left\{ -\ln \left[ \frac{f(y_i; \beta)}{f(y_i; \beta_0)} \right] \right\}$$

$$> -\ln \left\{ E \left[ \frac{f(y_i; \beta)}{f(y_i; \beta_0)} \right] \right\}$$

$$= -\ln \int \left[ \frac{f(y; \beta)}{f(y; \beta_0)} \right] [f(y; \beta_0)] d\mu$$

$$= -\ln \int [f(y; \beta)] d\mu = \ln(1) = 0.$$  

The claim now follows since

$$Eq(z_i, \beta) - Eq(z_i, \beta_0) = E \ln f(y_i, \beta_0) - E \ln f(y_i, \beta)$$

$$= K[f(\cdot, \beta_0), f(\cdot, \beta)]$$

and as just shown $K[f(\cdot, \beta_0), f(\cdot, \beta)] > 0$ for $\beta \neq \beta_0$.  

Remark: The quantity $K(\cdot, \cdot)$ is know as the Kullback-Leibler divergence. The proof of the above lemma shows that the Kullback-Leibler divergence is zero iff $f(y; \beta) = f(y; \beta_0)$ $\mu$-a.e. The result can be generalized to conditional densities; see, e.g., Manski (1988), ch. 5.1, and Wooldridge (2002), Appendix 13A.

Remark: If in the above theorem the family of conditional densities

$$\{f(y \mid x; \beta) : \beta \in B, x \in X\}$$
does not contain the true density, but there exists a $\beta_0$ that is unique minimizer of $Eq(z_i, \beta)$, then the above theorem still holds. Of course, in this case we can no longer interpret $\beta_0$ as a true parameter, but only as a minimizer of the limiting objective function $Eq(z_i, \beta)$.

### 4.2 Generalized Method of Moments Estimators

As discussed, estimators corresponding to objective functions of the form

$$R_n(\omega, \beta) = Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \beta) = \left[ n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right]' \hat{\Xi}_n \left[ n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right]$$

(4.5)

where $q(z, \beta)$ is a real vector valued function defined on $Z \times B$ taking its values in $\mathbb{R}^{pq}$, $\hat{\Xi}_n$ is a $pq \times pq$ positive semi-definite symmetric matrix and $\hat{\tau}_n$ is the vector of diagonal and upper diagonal elements of $\hat{\Xi}_n$, are typically called generalized method of moments estimators. In the following we will give a general catalogue of assumptions for the consistency of generalized method of moments estimators for the case of i.i.d. data.

Consider the following analogue of Assumption 4.1 where $q$ is now a vector valued function. In the following $\| \|\|$ denotes the Euclidean vector norm. Observe that for any vector $a = [a_1, \ldots, a_p]'$ we have $|a| \leq \|a\| = \left[ \sum_{i=1}^{p} a_i^2 \right]^{1/2}$ and $\|a\| = \left[ \sum_{i=1}^{p} a_i^2 \right]^{1/2} \leq p^{1/2} \sum_{i=1}^{p} |a_i|$, where the last inequality follows from the inequality (1.4.4) in Bierens (1994). Hence the convergence of vectors is equivalent to the convergence of the elements.

**Assumption 4.4** (a) $q : Z \times B \rightarrow \mathbb{R}^{pq}$, where $Z \subseteq \mathbb{R}^p$, and $B \subseteq \mathbb{R}^p$ are Borel sets. (b) $B$ is furthermore compact. (c) $q(\cdot, \beta)$ is Borel measurable for each $\beta \in B$, and $q(z, \cdot)$ is continuous for each $z \in Z$. (d) $z_i$ is a sequence of identically and independently distributed random vectors that take their values in $\mathbb{R}^{pq}$. (e) $E \sup_{\beta \in B} \|q(z_i, \beta)\| < \infty$.

We now have the following general result concerning the consistency of generalized method of moments estimators in case of i.i.d. data.

**Theorem 4.4** Consider the generalized method of moments estimators $\hat{\beta}_n$ defined as minimizers of (4.5). Suppose Assumption 4.4 holds, $\hat{\Xi}_n \rightarrow \Xi_0$ a.s. [i.p.] as $n \rightarrow \infty$, where the $pq \times pq$ (stochastic real valued) matrices $\hat{\Xi}_n$ are symmetric positive semi-definite and the $pq \times pq$ (real) matrix $\Xi_0$ is symmetric positive definite, and suppose $\beta_0$ is the unique solution of $Eq(z_i, \beta) = 0$, then

$$\hat{\beta}_n \rightarrow \beta_0 \text{ a.s. [i.p.] as } n \rightarrow \infty.$$
Proof: Given Assumption 4.4 it follows directly from the ULLN given as Theorem 3.3 that

$$\sup_{\beta \in B} \left\| n^{-1} \sum_{i=1}^{n} [q(z_i, \beta) - Eq(z_i, \beta)] \right\| \to 0 \text{ a.s. as } n \to \infty$$

and that $Eq(z_i, \beta)$ is continuous. Let $R(\beta) = Eq(z_i, \beta)'\Xi_0 Eq(z_i, \beta)$, then clearly $R(\beta)$ is continuous. Furthermore, as remarked above, since $\Xi_n \to \Xi_0$ a.s. /i.p./ as $n \to \infty$, it follows, e.g., from Lemma 3.3 in Pötscher and Prucha (1997) that

$$\sup_{\beta \in B} \left| R_n(\omega, \beta) - R(\beta) \right| \to 0 \text{ a.s. /i.p./ as } n \to \infty.$$

Finally observe that $R(\beta_0) = 0$ and $R(\beta) > 0$ for $\beta \neq \beta_0$ since $\Xi_0$ is positive definite. Thus $\beta_0$ is the unique minimizer of $R(\beta)$. It now follows from Theorem 3.2 that $\hat{\beta}_n \to \beta_0$ a.s. /i.p./ as $n \to \infty$.

Recall again that we noted in Section 3 that the ULLN given as Theorem 3.3 also holds if the assumption that $z_i$ is i.i.d. is replaced by the assumption that $z_i$ is stationary and ergodic. Thus the above consistency also holds if the i.i.d. assumption is relaxed to the assumption that $z_i$ is stationary and ergodic.
5 BASIC STRUCTURE OF THE ASYMPTOTIC NORMALITY PROOF

In this section we describe the basic structure underlying the derivation of the asymptotic distribution of M-estimators in nonlinear econometric models. As remarked in Section 1, the basic methods used in this derivation date back to Doob (1934), Cramér (1946), LeCam (1953), Huber (1967) and Jennrich (1969), to mention a few; for a more extensive bibliography see Norden (1972, 1973) and the references in Section 1. The asymptotic normality proofs in the articles on nonlinear econometric models listed in Section 1 all share this common structure. The basic idea is to express the estimator as a linear function of the score vector by means of a Taylor series expansion and then to derive the asymptotic distribution of the estimator from the asymptotic distribution of the score vector.

We maintain the basic setup as described in Section 2. That is, as in Section 2 we denote with \((z_i : i \in \mathbb{N})\) the data generating process defined on a probability space \((\Omega, \mathcal{F}, P)\) where the \(z_i\) are i.i.d. and take their values in \(Z \subseteq \mathbb{R}^{p_z}\), and we denote with \(B \subseteq \mathbb{R}^{p\beta}\) and \(T \subseteq \mathbb{R}^{p\tau}\) the space for the parameter vector of interest \(\beta\) and the space for the nuisance parameter vector \(\tau\), respectively. The sets \(Z, B\) and \(T\) are taken to be Borel sets. With \(Q_n(z_1, \ldots, z_n, \tau, \beta)\) we denote again the real valued criterion function defined on \(Z^n \times T \times B\), where \(Q_n(z_1, \ldots, z_n, \tau, \beta)\) is assumed to be \(\mathcal{F}\)-measurable for all \((\tau, \beta) \in T \times B\). In the preceding sections \(\hat{\beta}_n\) denoted the corresponding M-estimator for the true parameter value \(\beta_0\). We note that in the following \(\hat{\beta}_n\) need not necessarily be a minimizer of \(Q_n\), but only an approximate solution of a set of corresponding first order conditions. As before we assume that \(\hat{\tau}_n \overset{p}{\to} \tau_0\).

**Notation:** It proves helpful to introduce more compact notation for vectors and matrices of derivatives. Let \(f\) be a \(s \times 1\) vector of real valued functions defined on an open subset of \(\mathbb{R}^p \times \mathbb{R}^r\), let \(x = (x_1, \ldots, x_p)' \in \mathbb{R}^p\) and \(y = (y_1, \ldots, y_r)' \in \mathbb{R}^r\). Then \(\nabla_x f = (\partial f/\partial x_1, \ldots, \partial f/\partial x_p)'\) is the \(s \times p\) matrix of first order partial derivatives w.r.t. \(x\) and \(\nabla_x f' = (\nabla_x f)'\). If \(s = 1\), then \(\nabla_{xy} f = \nabla_y (\nabla_x f)\) denotes the \(p \times r\) matrix of second order partial derivatives. More generally, if \(s \geq 1\), then \(\nabla_{xy} f = \nabla_y (\text{vec}(\nabla_x f))\) denotes the \(ps \times r\) matrix of second order partial derivatives. \(\nabla_{xx} f\) and \(\nabla_{yy} f\) are defined analogously.

5.1 Outline of the Asymptotic Normality Proof

We start by giving an informal outline of the basic structure of the asymptotic normality proof. For simplicity we focus here on least mean distance estimators \(\hat{\beta}_n\) corresponding to the objective function

\[ Q_n(z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta). \]
The true parameter vector $\beta_0$ is assumed to be the unique minimizer of the nonstochastic analogue $Eq(z_i, \beta)$. We assume further that $\beta_0$ is an interior point of the parameter space, that all functions are sufficiently differentiable and that interchanging differentiation and taking expectations is permissible.

Now consider the first order condition for $\hat{\beta}_n$ after multiplication of the objective function with $n^{1/2}$, i.e.,

$$n^{1/2} \nabla_{\beta'} Q_n(z_1, \ldots, z_n, \hat{\beta}_n) = n^{-1/2} \sum_{i=1}^{n} \nabla_{\beta'} q(z_i, \hat{\beta}_n) = 0.$$ 

Applying the mean value theorem yields

$$0 = n^{1/2} \nabla_{\beta'} Q_n(z_1, \ldots, z_n, \hat{\beta}_n) = n^{1/2} \nabla_{\beta'} Q_n(z_1, \ldots, z_n, \tilde{\beta}_n) + \nabla_{\beta \beta} Q_n(z_1, \ldots, z_n, \tilde{\beta}_n) n^{1/2}(\tilde{\beta}_n - \beta_0)$$

where $\tilde{\beta}_n$ denote some values between $\hat{\beta}_n$ and $\beta_0$. (For a more formal application of the mean value theorem see the proof of the basic asymptotic normality results given below.) Assuming that $\nabla_{\beta \beta} Q_n(z_1, \ldots, z_n, \tilde{\beta}_n)$ is nonsingular we have

$$n^{1/2}(\tilde{\beta}_n - \beta_0) = - \left[ \nabla_{\beta \beta} Q_n(z_1, \ldots, z_n, \tilde{\beta}_n) \right]^{-1} n^{1/2} \nabla_{\beta'} Q_n(z_1, \ldots, z_n, \beta_0).$$

(5.1)

Under typical assumptions we will have

$$\nabla_{\beta \beta} Q_n(z_1, \ldots, z_n, \tilde{\beta}_n) = n^{-1} \sum_{i=1}^{n} \nabla_{\beta \beta} q(z_i, \tilde{\beta}_n) \overset{P}{\rightarrow} A_0$$

(5.2)

where $A_0 = E\nabla_{\beta \beta} q(z_i, \beta_0)$ is a real symmetric positive definite matrix. Recall that $Eq(z_i, \beta)$ is minimized at $\beta_0$. Hence, under the maintained assumptions,

$$\nabla_{\beta'} Eq(z_i, \beta_0) = E \left[ \nabla_{\beta'} q(z_i, \beta_0) \right] = 0.$$

Since $\nabla_{\beta'} q(z_i, \beta_0)$ is furthermore i.i.d., in light that $z_i$ is i.i.d., it follows (under an additional mild moment condition) that $\nabla_{\beta'} q(z_i, \beta_0)$ satisfies a CLT, i.e.,

$$n^{1/2} \nabla_{\beta'} Q_n(z_1, \ldots, z_n, \beta_0) = n^{-1/2} \sum_{i=1}^{n} \nabla_{\beta'} q(z_i, \beta_0) \overset{D}{\rightarrow} \zeta$$

(5.3)

where $\zeta \sim N(0, \Lambda_0)$. Then combining (5.1), (5.2), (5.3) it follows that

$$n^{1/2}(\tilde{\beta}_n - \beta_0) \overset{D}{\rightarrow} N(0, B_0 A_0^{-1})$$

with $B_0 = \Lambda_0$. 

26
5.2 Basic Theorems for Asymptotic Normality

We now give a formal basic theorem concerning the asymptotic normality of M-estimators in the i.i.d. case that covers both least mean distance and GMM estimators. We shall use the following assumption to formally derive the asymptotic distribution of \( \hat{\beta}_n \).

Assumption 5.1 \(^7\) (a) The parameter spaces \( T \) and \( B \) are compact subsets of \( \mathbb{R}^p \) and \( \mathbb{R}^q \), respectively.
(b) \( Q_n : \mathbb{Z}^n \times T \times B \to \mathbb{R} \) where \( Q_n(z_1, \ldots, z_n, \tau, \beta) \) is \( \mathfrak{A} \)-measurable for all \( (\tau, \beta) \in T \times B \) and \( Q_n(z_1, \ldots, z_n, \ldots) \) is a.s. twice continuously partially differentiable at every point \( (\tau, \beta) \) in the interior of \( T \times B \) (where the exceptional null set does not depend on \( (\tau, \beta) \)).
(c) The estimators \( (\hat{\tau}_n, \hat{\beta}_n) \) take their values in \( T \times B \), the true parameters \( (\tau_0, \beta_0) \) lie in the interior of \( T \times B \),
\[
\begin{align*}
\hat{\beta}_n & \overset{p}{\to} \beta_0 \text{ as } n \to \infty, \\
n^{1/2}(\hat{\tau}_n - \tau_0) & = O_p(1).
\end{align*}
\]
(d) The sequence \( \hat{\beta}_n \) satisfies
\[
n^{1/2}\nabla_{\beta}Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \hat{\beta}_n) = o_p(1).
\]
(I.e., \( \hat{\beta}_n \) satisfies the normalized first order conditions up to an error of magnitude \( o_p(1) \).)
(e) For all sequences of random vectors \( (\hat{\tau}_n, \hat{\beta}_n) \) with \( \hat{\tau}_n \overset{p}{\to} \tau_0 \) and \( \hat{\beta}_n \overset{p}{\to} \beta_0 \) we have
\[
\nabla_{\beta\beta}Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \hat{\beta}_n) \overset{p}{\to} A_0
\]
as \( n \to \infty \), where \( A_0 \) is a real symmetric positive definite matrix.
(f) For all sequences \( (\hat{\tau}_n, \hat{\beta}_n) \) as in (e) we have
\[
\nabla_{\tau\beta}Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \hat{\beta}_n) \overset{p}{\to} 0.
\]
(g) There exists a real matrix \( D_0 \) such that
\[
-n^{1/2}\nabla_{\beta}Q_n(z_1, \ldots, z_n, \tau_0, \beta_0) = D_0 \zeta_n + o_p(1)
\]
where \( \zeta_n \) and \( \zeta \) are random vectors satisfying \( \zeta_n \overset{D}{\to} \zeta \).

\(^7\) We note that because of Assumption 5.1(c) there exists a sequence of sets \( \Omega_n \in \mathfrak{A} \) with \( P(\Omega_n) \to 1 \) as \( n \to \infty \) such that \( (\hat{\tau}_n, \hat{\beta}_n) \) and \( (\tilde{\tau}_n, \tilde{\beta}_n) \) belong to the interior of \( T \times B \) for \( \omega \in \Omega_n \). Therefore the derivatives of \( Q_n \) evaluated at \( (\hat{\tau}_n, \hat{\beta}_n) \) and \( (\tilde{\tau}_n, \tilde{\beta}_n) \) considered below are well-defined at least for \( \omega \in \Omega_n \). As usual, in the sequel we will often use the notation \( \xi_n = o_p(a_n) \) or \( \xi_n = O_p(a_n) \) even if the variables \( \xi_n \) are only well defined on sets \( \Omega_n \in \mathfrak{A} \) with \( P(\Omega_n) \to 1 \) as \( n \to \infty \).
Remark: A few remarks regarding the assumptions maintained by the above theorem.

(a) Clearly, if $\hat{\beta}_n$ is an interior minimizer of $Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \beta)$, or more generally, if $\hat{\beta}_n$ is a solution of the first order conditions, then Assumption 5.1(d) is trivially satisfied.

(b) Assumption 5.1 also covers the case where no nuisance parameter is present. In particular, to incorporate this case into the framework of Assumption 5.1 we may formally view the objective function as a function on $T \times B$, where $T$ can be chosen as an arbitrary subset of $\mathbb{R}^{p_q}$ with $\text{int}(T) \neq \emptyset$, and by setting $\hat{\tau}_n = \tau_0$, where $\tau_0$ is an arbitrary element of $\text{int}(T)$.

(c) The convergence of $\nabla_\beta Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \hat{\beta}_n)$ and $\nabla_{\beta \tau} Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \hat{\beta}_n)$ postulated in Assumption 5.1(e,f) will typically be established by establishing some uniform convergence of $\nabla_\beta Q_n(z_1, \ldots, z_n, \tau, \beta)$ and $\nabla_{\beta \tau} Q_n(z_1, \ldots, z_n, \tau, \beta)$ and then applying Theorem 3.4.

(d) The convergence in distribution postulated in Assumption 5.1(g) for the score vector evaluated at the true parameters will typically be established with the help of a CLT.

We now have the following basic theorem regarding the asymptotic normality of $\hat{\beta}_n$. The proof will be given in the next section.

Theorem 5.1 Given Assumption 5.1 holds, then

$$n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{D} A_0^{-1}D_0\zeta.$$  

If $\zeta \sim N(0, \Lambda_0)$, then

$$n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{D} N(0, A_0^{-1}B_0A_0^{-1})$$

with $B_0 = D_0A_0D_0'$.

The above theorem covers both least mean distance and GMM estimators. We next give a more specialized result that is geared towards GMM estimators. For GMM estimators the objective function is of the general form

$$Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \hat{\beta}_n) = S_n(z_1, \ldots, z_n, \hat{\beta}_n)^T\hat{\Xi}_nS_n(z_1, \ldots, z_n, \hat{\beta}_n)$$

with $S_n : Z^n \times B \rightarrow \mathbb{R}^{p_q}$ and $\hat{\Xi}_n = \hat{\Xi}_n(\hat{\tau}_n) \xrightarrow{p} \Xi_0 = \Xi_0(\tau_0)$, and where $\hat{\Xi}_n$ and $\Xi_0$ are $p_q \times p_q$ symmetric positive semi-definite/definite matrices, and $\hat{\tau}_n$ and $\tau_0$ are vectors of length $p_q(p_q + 1)/2$ that contain the diagonal and upper-diagonal elements of $\hat{\Xi}_n$ and $\Xi_0$, respectively. Utilizing the specific structure of the
above objective function it is possible to establish asymptotic normality under
the weaker assumption that \( \hat{\tau}_n - \tau_0 = o_p(1) \) as compared to the assumption
that \( n^{1/2}(\hat{\tau}_n - \tau_0) = O_p(1) \) as was maintained above. Consider the following
assumption.

**Assumption 5.2** (a) The parameter space \( B \) is a compact subsets of \( \mathbb{R}^p \).
(b) \( S_n : \mathbb{Z}^n \times B \to \mathbb{R}^p \) where \( S_n(z_1, \ldots, z_n, \beta) \) is \( \mathfrak{A} \)-measurable for all \( \beta \in B \)
and \( S_n(z_1, \ldots, z_n, \cdot) \) is a.s. continuously partially differentiable at every point \( \beta \) in the interior of \( B \) (where the exceptional null set does not depend on \( \beta \)).
(c) The estimators \( \hat{\beta}_n \) take their values in \( B \), the true parameter \( \beta_0 \) lies
in the interior of \( B \), and \( \hat{\beta}_n \to \beta_0 \) as \( n \to \infty \).
(d) The sequence \( \hat{\beta}_n \) satisfies
\[
n^{1/2} \nabla_\beta S_n(z_1, \ldots, z_n, \hat{\beta}_n) \hat{\Xi}_n S_n(z_1, \ldots, z_n, \hat{\beta}_n) = o_p(1).
\]
(I.e., \( \hat{\beta}_n \) satisfies the normalized first order conditions up to an error of magnitude \( o_p(1) \).
(e) For all sequences of random vectors \( \tilde{\beta}_n \) and \( \hat{\beta}_n \to \beta_0 \) we have
\[
\nabla_\beta S_n(z_1, \ldots, z_n, \tilde{\beta}_n) \to G_0
\]
as \( n \to \infty \), where \( G_0 \) is a real \( p_\beta \times p_\beta \) matrix with full column rank.
(g) The normalized score vector satisfies
\[
n^{1/2} S_n(z_1, \ldots, z_n) \to \zeta
\]
where \( \zeta \) is a \( p_\beta \times 1 \) real valued random vector.

The proof of the following theorem concerning asymptotic normality of GMM
estimators is given in the next section.

**Theorem 5.2** Given Assumption 5.2 holds, then
\[
n^{1/2}(\hat{\beta}_n - \beta_0) \to -[G_0^\prime \Xi_0 G_0]^{-1} G_0^\prime \Xi_0 \zeta.
\]
If \( \zeta \sim N(0, \Lambda_0) \), then
\[
n^{1/2}(\hat{\beta}_n - \beta_0) \to N(0, A_0^{-1} B_0 A_0^{-1}).
\]
with \( A_0 = G_0^\prime \Xi_0 G_0 \), where \( A_0 \) is nonsingular, and \( B_0 = G_0^\prime \Xi_0 A_0 \Xi_0 G_0 \).
5.3 Proof of Basic Theorems for Asymptotic Normality

In this section we formally prove Theorems 5.1 and 5.2.

**Proof of Theorem 5.1.** Assumption 5.1(c) implies the existence of a sequence \( \Omega_n \) with \( P(\Omega_n) \to 1 \) as \( n \to \infty \) such that on respective events \( \Omega_n \) the line segments between \((\hat{\tau}_n, \hat{\beta}_n)\) and \((\tau_0, \beta_0)\) lie in the interior of \( T \times B \). We also may assume that the sets \( \Omega_n \) are disjoint from the exceptional null set in Assumption 5.1(b). For ease of notation we use \( Q_n(\omega, \tau, \beta) \) as shorthand for \( Q_n(z_1, ..., z_n, \tau, \beta) \). Then for \( \omega \in \Omega_n \) we have by the mean value theorem applied to \( \nabla_\beta Q_n \):

\[
\nabla_\beta Q_n(\omega, \tilde{\tau}_n, \tilde{\beta}_n) = \nabla_\beta Q_n(\omega, \tau_0, \beta_0) + \nabla_{\beta\beta} Q_n(\omega, \{\tilde{\tau}_n^i\}, \{\tilde{\beta}_n^i\})(\hat{\beta}_n - \beta_0) + \nabla_{\beta\tau} Q_n(\omega, \{\tilde{\tau}_n^i\}, \{\tilde{\beta}_n^i\})(\hat{\tau}_n - \tau_0).
\]

Here \( \nabla_{\beta\beta} Q_n(\omega, \{\tilde{\tau}_n^i\}, \{\tilde{\beta}_n^i\}) \) denotes the matrix whose \( j \)-th row is the \( j \)-th row of \( \nabla_{\beta\beta} Q_n \) evaluated at \((\omega, \tilde{\tau}_n^i, \tilde{\beta}_n^i)\), where \((\tilde{\tau}_n^i, \tilde{\beta}_n^i)\) is the mean value arising from the application of the mean value theorem to the \( j \)-th component of \( \nabla_{\beta} Q_n(\omega, \tilde{\tau}_n, \tilde{\beta}_n) \). (Lemma 3 in Jennrich (1969) implies that the mean values actually can be chosen to be measurable.) It now follows from Assumption 5.1(d) that

\[
o_p(1) = n^{1/2} \nabla_{\beta\tau} Q_n(\omega, \tau_0, \beta_0) + \nabla_{\beta\beta} Q_n(\omega, \{\tilde{\tau}_n^i\}, \{\tilde{\beta}_n^i\}) n^{1/2}(\hat{\beta}_n - \beta_0) + \nabla_{\beta\tau} Q_n(\omega, \{\tilde{\tau}_n^i\}, \{\tilde{\beta}_n^i\}) n^{1/2}(\hat{\tau}_n - \tau_0).
\]

Assumption 5.1(f) implies that

\[
\nabla_{\beta\tau} Q_n(\omega, \tilde{\tau}_n^j, \tilde{\beta}_n^j) = o_p(1)
\]

for \( j = 1, ..., p_\beta \). Since the \( j \)-th row of \( \nabla_{\beta\tau} Q_n(\omega, \{\tilde{\tau}_n^i\}, \{\tilde{\beta}_n^i\}) \) coincides with the \( j \)-th row of \( \nabla_{\beta\tau} Q_n(\omega, \tilde{\tau}_n, \tilde{\beta}_n) \) it follows that

\[
\nabla_{\beta\tau} Q_n(\omega, \{\tilde{\tau}_n^i\}, \{\tilde{\beta}_n^i\}) = o_p(1).
\]

Since by Assumption 5.1(c) \( n^{1/2}(\hat{\tau}_n - \tau_0) = O_p(1) \) it follows further that

\[
\nabla_{\beta\tau} Q_n(\omega, \{\tilde{\tau}_n^i\}, \{\tilde{\beta}_n^i\}) n^{1/2}(\hat{\tau}_n - \tau_0) = o_p(1).
\]

In light of this it follows from (5.5) that

\[
A_n n^{1/2}(\hat{\beta}_n - \beta_0) = -n^{1/2} \nabla_{\beta\tau} Q_n(\omega, \tau_0, \beta_0) + o_p(1),
\]

30
where \( A_n = \nabla_{\beta\beta} Q_n(\omega, \{\tilde{\beta}^i_n\}, \{\tilde{\beta}_n^j\}) \)

Multiplying (5.6) by the Moore-Penrose inverse of \( A_n \) and rearranging terms we get

\[
n^{1/2}(\hat{\beta}_n - \bar{\beta}_n) = [I - A_n^+ A_n] n^{1/2} (\hat{\beta}_n - \bar{\beta}_n) + A_n^+ n^{1/2} \nabla_{\beta} Q_n(\omega, \tau_0, \beta_0) + A_n^+ o_p(1).
\] (5.7)

Assumption 5.1(e) implies that

\[
\nabla_{\beta\beta} Q_n(\omega, \tilde{\beta}^i_n, \tilde{\beta}^j_n) - A_0 = o_p(1)
\]

for every \( j = 1, \ldots, p_\beta \). Since the \( j \)-th row of \( \nabla_{\beta\beta} Q_n(\omega, \{\tilde{\beta}^i_n\}, \{\tilde{\beta}_n^j\}) \) coincides with the \( j \)-th row of \( \nabla_{\beta\beta} Q_n(\omega, \tilde{\beta}^i_n, \tilde{\beta}_n^j) \) it follows that

\[
A_n \xrightarrow{p} A_0
\]

as \( n \to \infty \). Since the matrix \( A_0 \) is non-singular by Assumption 5.1(e) it follows further that

\[
A_n^+ \xrightarrow{p} A_0^{-1}
\]

as \( n \to \infty \). In light of this it follows that \([I - A_n^+ A_n] n^{1/2} (\hat{\beta}_n - \bar{\beta}_n) \) is zero on \( \omega \)-sets of probability tending to one. Observing that \( A_n^+ = O_p(1) \) it follows that \( A_n^+ o_p(1) = o_p(1) \). That is, both the first and last term on the r.h.s. of (5.7) are \( o_p(1) \). This gives

\[
n^{1/2}(\hat{\beta}_n - \beta_0) = -A_n^+ n^{1/2} \nabla_{\beta} Q_n(\omega, \tau_0, \beta_0) + o_p(1)
\]

By Assumption 5.1(g) we have

\[
n^{1/2} \nabla_{\beta} Q_n(\omega, \tau_0, \beta_0) = D_0 \zeta_n + o_p(1)
\]

with \( \zeta_n \xrightarrow{D} \zeta \). Hence clearly \( n^{1/2} \nabla_{\beta} Q_n(\omega, \tau_0, \beta_0) = O_p(1) \) and thus

\[
n^{1/2}(\hat{\beta}_n - \beta_0) = -A_0^{-1} n^{1/2} \nabla_{\beta} Q_n(\omega, \tau_0, \beta_0) + o_p(1)
\]

\[= A_0^{-1} D_0 \zeta_n + o_p(1).\]

The remaining parts of the lemma are obvious.

**Proof of Theorem 5.2.** Assumption 5.2(c) implies the existence of a sequence \( \Omega_n \in \mathcal{A} \) with \( P(\Omega_n) \to 1 \) as \( n \to \infty \) such that on respective events \( \Omega_n \) the line segments between \( \hat{\beta}_n \) and \( \beta_0 \) lie in the interior of \( B \). We also may assume that the sets \( \Omega_n \) are disjoint from the exceptional null set in Assumption 5.2(b). For ease of notation we use \( S_n(\omega, \beta) \) as shorthand for \( S_n(\mathbf{z}_1, \ldots, \mathbf{z}_n, \beta) \). Then for \( \omega \in \Omega_n \) we have by the mean value theorem applied to \( S_n \)

\[
S_n(\omega, \hat{\beta}_n) = S_n(\omega, \beta_0) + \nabla_{\beta} S_n(\omega, \{\hat{\beta}_n^j\})(\hat{\beta}_n - \beta_0).
\] (5.8)
Here \( \nabla_{\beta'}S_n(\omega, \{\tilde{\beta}_n^j\}) \) denotes the matrix whose \( j \)-th row is the \( j \)-th row of \( \nabla_{\beta'}S_n \) evaluated at \( (\omega, \tilde{\beta}_n^j) \), where \( \tilde{\beta}_n^j \) is the mean value arising from the application of the mean value theorem to the \( j \)-th component of \( S_n(\omega, \tilde{\beta}_n) \). (Lemma 3 in Jennrich (1969) implies that the mean values actually can be chosen to be measurable.) It now follows from Assumption 5.1(d) that

\[
o_p(1) = n^{1/2} \nabla_{\beta'}S_n(\omega, \hat{\beta}_n) \tilde{\xi}_n S_n(\omega, \beta_0) + \nabla_{\beta'}S_n(\omega, \hat{\beta}_n) \tilde{\xi}_n \nabla_{\beta'}S_n(\omega, \{\tilde{\beta}_n^j\}) n^{1/2}(\hat{\beta}_n - \beta_0). \tag{5.9}
\]

Rearranging terms in (5.9) yields

\[
A_n n^{1/2}(\hat{\beta}_n - \beta_0) = -n^{1/2} \nabla_{\beta'}S_n(\omega, \hat{\beta}_n) \tilde{\xi}_n S_n(\omega, \beta_0) + o_p(1), \tag{5.10}
\]

where

\[
A_n = \nabla_{\beta'}S_n(\omega, \hat{\beta}_n) \tilde{\xi}_n \nabla_{\beta'}S_n(\omega, \{\tilde{\beta}_n^j\})
\]

Multiplying (5.10) by the Moore-Penrose inverse of \( A_n \) and rearranging terms we get

\[
n^{1/2}(\hat{\beta}_n - \tilde{\beta}_n) = [I - A_n^+ A_n] n^{1/2}(\hat{\beta}_n - \tilde{\beta}_n) \tag{5.11}
\]

\[
- A_n^+ n^{1/2} \nabla_{\beta'}S_n(\omega, \hat{\beta}_n) \tilde{\xi}_n S_n(\omega, \beta_0) + A_n^+ o_p(1).
\]

Assumption 5.2(c) implies that

\[
\nabla_{\beta}S_n(\omega, \hat{\beta}_n) - G_0 = o_p(1)
\]

and

\[
\nabla_{\beta}S_n(\omega, \{\tilde{\beta}_n^j\}) - G_0 = o_p(1)
\]

and for \( j = 1, \ldots, p_\beta \). Since the \( j \)-th row of \( \nabla_{\beta'}S_n(\omega, \{\tilde{\beta}_n^j\}) \) coincides with the \( j \)-th row of \( \nabla_{\beta'}S_n(\omega, \tilde{\beta}_n^j) \) it follows that

\[
\nabla_{\beta}S_n(\omega, \{\tilde{\beta}_n^j\}) - G_0 = o_p(1).
\]

Given Assumption 5.2(c) it follows further that

\[
A_n \xrightarrow{p} G_0 \tilde{\xi}_0 G_0
\]

as \( n \rightarrow \infty \). Since \( G \) has full column rank by Assumption 5.2(e) and \( \tilde{\xi}_0 \) is nonsingular by Assumption 5.2(c) it follows further that

\[
A_n^+ \xrightarrow{p} [G_0^\top \tilde{\xi}_0 G_0]^{-1}
\]

as \( n \rightarrow \infty \). In light of this it follows that \([I - A_n^+ A_n] n^{1/2}(\hat{\beta}_n - \tilde{\beta}_n) \) is zero on \( \omega \)-sets of probability tending to one. Observing that \( A_n^+ = O_p(1) \) it follows that \( A_n^+ o_p(1) = o_p(1) \). That is, both the first and last term on the r.h.s. of (5.11) are \( o_p(1) \). This gives

\[
n^{1/2}(\hat{\beta}_n - \beta_0) = -A_n^+ \nabla_{\beta'}S_n(\omega, \hat{\beta}_n) \tilde{\xi}_n n^{1/2} S_n(\omega, \beta_0) + o_p(1)
\]

32
By Assumption 5.2(g) we have

\[ n^{1/2} S_n(\omega, \beta_0) \xrightarrow{D} \zeta. \]

Using the above observations regarding the probability limits of \( A_n^+ \), \( \nabla \beta^* S_n(\omega, \hat{\beta}_n) \) and \( \Xi_0 \) it follows that

\[ n^{1/2} (\hat{\beta}_n - \beta_0) \xrightarrow{D} - [G_0' \Xi_0 G_0]^{-1} G_0' \Xi_0 \zeta. \]

The remaining parts of the lemma are obvious.
6 ASYMPTOTIC NORMALITY: CATALOGUES OF ASSUMPTIONS

We continue to maintain the general setup of Section 2: We denote with \((z_i : i \in \mathbb{N})\) the data generating process defined on a probability space \((\Omega, \mathcal{A}, \mathbb{P})\) where the \(z_i\) are i.i.d. and take their values in \(Z \subseteq \mathbb{R}^p_z\), and we denote with \(B \subseteq \mathbb{R}^p_\beta\) and \(T \subseteq \mathbb{R}^p_\tau\) the space for the parameter vector of interest \(\beta\) and the space for the nuisance parameter vector \(\tau\), respectively. The sets \(Z, B, T\) are taken to be Borel sets.

Now let \(Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \hat{\beta}_n)\) be a real valued function defined on \(Z^n \times T \times B\) (where \(n\) denotes the sample size). Assume further that \(Q_n(z_1, \ldots, z_n, \tau, \beta)\) is \(\mathcal{A}\)-measurable for all \((\tau, \beta) \in T \times B\). The M-estimators \(\hat{\beta}_n\) corresponding to the objective function \(Q_n\) for given estimators \(\hat{\tau}_n\) are then defined by:

\[
Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \hat{\beta}_n) = \inf_{\beta \in B} Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \beta),
\]

i.e., they minimize the objective function over \(B\).

In the previous section we have discussed the basic structure of the asymptotic normality proof for M-estimators and have established basic modules that can be employed to establish asymptotic normality of M-estimators. In the following we use those modules to give specific low-level catalogues of assumptions for the asymptotic normality of least mean distance and generalized method of moments estimators. One of the assumptions maintained by the basic modules that allows us to establish asymptotic normality is that \(\hat{\beta}_n\) is consistent. In Section 4 we have given specific low-level catalogues of assumptions for the consistency of least mean distance and generalized method of moments estimators. For purposes of convenience we will restate those catalogues in the following, and infer consistency of \(\hat{\beta}_n\) from those catalogues by appealing to the theorems regarding consistency proven in Section 4.

We use the following notation. Let \(A = (a_{ij})\) be some matrix, then \(\|A\|\) denotes the Euclidean matrix norm, which is defined as

\[
\|A\| = [\text{tr}(A'A)]^{1/2} = \left[ \sum_i \sum_j a_{ij}^2 \right]^{1/2}.
\]

Note that this matrix norm is submultiplicative, i.e., for two conformable matrices \(A\) and \(B\) we have

\[
\|AB\| \leq \|A\| \|B\|.
\]

If \(A\) is a vector, then \(\|A\|\) is just the usual Euclidean vector norm.

In the following we shall also make use of the following observations without further explicit reference: Let \(f : Z \times B' \to \mathbb{R}^p_f\), where \(Z \subseteq \mathbb{R}^p_z\) and \(B' \subseteq \mathbb{R}^p_{\beta'}\) are Borel sets, suppose

\[
E \sup_{\beta \in B'} \|f(z_i, \beta)\|^2 < \infty
\]

then:
• \[ E \sup_{\beta \in B'} \| f(z_i, \beta) f(z_i, \beta)' \| < \infty \]
  since
  \[
  \| f(z_i, \beta) f(z_i, \beta)' \| = \left\{ \text{tr} \left[ f(z_i, \beta) f(z_i, \beta)' f(z_i, \beta) f(z_i, \beta)' \right] \right\}^{1/2}
  = \left\{ \text{tr} \left[ f(z_i, \beta)' f(z_i, \beta) f(z_i, \beta)' f(z_i, \beta) \right] \right\}^{1/2}
  = f(z_i, \beta)' f(z_i, \beta) = \| f(z_i, \beta) \| ^2.
  \]

• \[ E \sup_{\beta \in B'} \| f(z_i, \beta) \| < \infty \]
  since
  \[ \| f(z_i, \beta) \| \leq 1 + \| f(z_i, \beta) \| ^2. \]

\[ E \| f(z_i, \beta_0) f(z_i, \beta_0)' \| < \infty, \]
\[ E \| f(z_i, \beta_0) \| < \infty, \]
and so \( E f(z_i, \beta_0) f(z_i, \beta_0)' \) and \( E f(z_i, \beta_0) \) are well defined for any \( \beta_0 \in B' \).

### 6.1 Least Mean Distance Estimators

#### 6.1.1 General

As discussed, estimators corresponding to objective functions of the form

\[ R_n(\omega, \beta) = Q_n(z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \tag{6.1} \]

where \( q(z, \beta) \) is a real valued function defined on \( Z \times B \) are typically called least mean distance estimators. In the following we will give a general catalogue of assumptions for the asymptotic normality of least mean distance estimators for the case of i.i.d. data

**Assumption 6.1**

(a) \( q(\ldots) \) is a real valued function on \( Z \times B \), where \( Z \subseteq \mathbb{R}^p \) and \( B \subseteq \mathbb{R}^p \) are Borel sets. (b) \( B \) is furthermore compact. (c) \( q(\ldots, \beta) \) is Borel measurable for each \( \beta \in B \), and \( q(z, \ldots) \) is continuous for each \( z \in Z \). (d) \( z_i \) is a sequence of identically and independently distributed random vectors that take their values in \( \mathbb{R}^p \). (e) \( E \sup_{\beta \in B} |q(z_i, \beta)| < \infty \). (f) \( \beta_0 \) is a unique minimizer of \( Eq(z_i, \beta) \).
Assumption 6.1 collects the assumptions maintained for the consistency result for least mean distance estimators given as Theorem 4.1. We need the following additional assumption.

Assumption 6.1

(a) $\beta_0$ is an interior point of $B$. (b) $q(z, \cdot)$ is twice continuously differentiable at every interior point $\beta \in B$ for each $z \in Z$. (c) $E \sup_{\beta \in B'} \| \nabla_{\beta} q(z_i, \beta) \| < \infty$ and $E \sup_{\beta \in B'} \| \nabla_{\beta} q(z_i, \beta) \|^2 < \infty$ where $B'$ is a closed ball around $\beta_0$ that lies in the interior of $B$. (d) $E \nabla_{\beta} q(z_i, \beta_0)$ is nonsingular (and symmetric in light of (b)).

Remark: The reason for introducing the set $B'$ in formulating the dominance conditions in part (c) of the above assumption is to ensure that the derivatives are well defined on the space over which the supremum is being taken.

We now have the following general result concerning the asymptotic normality of least mean distance estimators in case of i.i.d. data.

Theorem 6.1 Consider the least mean distance estimators $\hat{\beta}_n$ defined as minimizers of (6.1). Suppose Assumption 6.1 and 6.1* hold, then

$$n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{D} N(0, A_0^{-1}B_0A_0^{-1})$$

where

$$A_0 = E \nabla_{\beta} q(z_i, \beta_0), \quad B_0 = E \nabla_{\beta} q(z_i, \beta_0) \nabla_{\beta} q(z_i, \beta_0).$$

and $A_0$ positive definite. Furthermore, let

$$\hat{A}_n = n^{-1} \sum_{i=1}^{n} \nabla_{\beta} q(z_i, \hat{\beta}_n), \quad \hat{B}_n = n^{-1} \sum_{i=1}^{n} \nabla_{\beta} q(z_i, \hat{\beta}_n) \nabla_{\beta} q(z_i, \hat{\beta}_n),$$

then

$$\hat{A}_n \xrightarrow{p} A_0, \quad \hat{B}_n \xrightarrow{p} B_0,$$

and thus

$$\hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1} \xrightarrow{p} A_0^{-1}B_0A_0^{-1}.$$

Proof: To prove the theorem we first verify that Assumptions 6.1 and 6.1* imply Assumptions 5.1 for $Q_n(z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta)$ and $\hat{\beta}_n$, which are defined as minimizers of $Q_n(z_1, \ldots, z_n, \beta)$ over $B$. Since no nuisance parameter is present we can ignore all conditions with respect the nuisance parameter and the corresponding estimator; compare the remark after Theorem 5.1. We first note that in light of Assumption 6.1 it follows from Theorem 4.1 that $\hat{\beta}_n \xrightarrow{p} \beta_0$ as $n \to \infty$, where $\beta_0$ is an interior point of $B$ by Assumption 6.1*(a). By definition $\hat{\beta}_n \in B$, and thus the relevant part of Assumption 5.1(c) holds.
Assumptions 5.1(a), (b) hold in light of Assumptions 6.1(a), (b), (c) and Assumption 6.1*(b).

Since \( \hat{\beta}_n \overset{p}{\to} \beta_0 \) as \( n \to \infty \), where \( \beta_0 \) is an interior point of \( B \), it follows that \( \hat{\beta}_n \) lies in the interior of \( B \) except for \( \omega \)-sets whose probability tends to zero. That is

\[
n^{1/2} \nabla' q_n(z_1, \ldots, z_n, \hat{\beta}_n) = n^{-1/2} \sum_{i=1}^{n} \nabla' q(z_i, \hat{\beta}_n) = 0
\]

except on \( \omega \)-sets whose probability tends to zero. Thus also Assumption 5.1(d) holds.

Next consider

\[
-n^{1/2} \nabla' q_n(z_1, \ldots, z_n, \beta_0) = -n^{-1/2} \sum_{i=1}^{n} \nabla' q(z_i, \beta_0).
\]

In light of Assumption 6.1*(c) we have \( E \sup_{\beta \in B'} \| \nabla' q(z_i, \beta) \| < \infty \) and hence we are allowed to interchange differentiation and integration (see, e.g., Theorem 16.8 in Billingsley (1979)) to obtain

\[
\nabla' E q(z_i, \beta_0) = E \nabla' q(z_i, \beta_0) = 0.
\]

The last equality holds since by Assumption 6.1(f) the true parameter vector \( \beta_0 \) is the unique minimizer of \( E(z_i, \beta_0) \). Consequently it follows that the score vectors \( \nabla' q(z_i, \beta_0) \) are i.i.d. with zero mean and variance covariance matrix \( A_0 = E \nabla' q(z_i, \beta_0) \nabla' q(z_i, \beta_0) \). Let

\[
\zeta_n = -n^{-1/2} \sum_{i=1}^{n} \nabla' q(z_i, \beta_0)
\]

then \( \zeta_n \overset{D}{\to} N(0, A_0) \) Hence Assumption 5.1(g) is satisfied with \( \zeta \sim N(0, A_0) \) and \( D_0 = I \). Observe that here \( B_0 = D_0 A_0 D'_0 = A_0 \).

Next observe that all assumptions maintained by the uniform law of large numbers given as Theorem 3.3 are satisfied for \( \nabla' q(z_i, \beta) \) and \( \nabla' q(z_i, \beta) \nabla' q(z_i, \beta) \). The needed condition that

\[
E \sup_{\beta \in B'} \| \nabla' q(z_i, \beta) \| < \infty \) and \( E \sup_{\beta \in B'} \| \nabla' q(z_i, \beta) \nabla' q(z_i, \beta) \| < \infty
\]

is satisfied in light of Assumption 6.1*(c). Hence

\[
\sup_{\beta \in B'} \left| \frac{1}{n} \sum_{i=1}^{n} \nabla' q(z_i, \beta) - E \nabla' q(z_i, \beta) \right| \overset{P}{\to} 0, \quad (6.2)
\]

\[
\sup_{\beta \in B'} \left| \frac{1}{n} \sum_{i=1}^{n} \nabla' q(z_i, \beta) \nabla' q(z_i, \beta) - E \nabla' q(z_i, \beta) \nabla' q(z_i, \beta) \right| \overset{P}{\to} 0,
\]

37
as \( n \to \infty \).

Now let \( \beta_n \) be any sequence of random vectors with \( \beta_n \xrightarrow{p} \beta_0 \) as \( n \to \infty \). Furthermore define \( \beta_n = \beta_n \) if \( \beta_n \in B' \) and \( \beta_n = \beta_0 \) if \( \beta_n \notin B' \). Then clearly \( \beta_n \xrightarrow{p} \beta_0 \) as \( n \to \infty \), and furthermore

\[
    n^{-1} \sum_{i=1}^{n} \nabla_{\beta} q(z_i, \tilde{\beta}_n) - n^{-1} \sum_{i=1}^{n} \nabla_{\beta} q(z_i, \beta_n) = o_p(1),
\]

(6.3)

\[
    n^{-1} \sum_{i=1}^{n} \nabla_{\beta} q(z_i, \tilde{\beta}_n) \nabla q(z_i, \tilde{\beta}_n) - n^{-1} \sum_{i=1}^{n} \nabla_{\beta} q(z_i, \beta_n) \nabla q(z_i, \beta_n) = o_p(1).
\]

Because of (6.2) it follows from Theorem 3.4 that

\[
    n^{-1} \sum_{i=1}^{n} \nabla_{\beta} q(z_i, \tilde{\beta}_n) \xrightarrow{p} A_0, \quad n^{-1} \sum_{i=1}^{n} \nabla_{\beta} q(z_i, \beta_n) \nabla q(z_i, \beta_n) \xrightarrow{p} B_0.
\]

(6.4)

Combining (6.3) and (6.4) we have

\[
    \nabla_{\beta} Q_n(z_1, \ldots, z_n, \tilde{\beta}_n) = n^{-1} \sum_{i=1}^{n} \nabla_{\beta} q(z_i, \tilde{\beta}_n) \xrightarrow{p} A_0,
\]

(6.5)

\[
    n^{-1} \sum_{i=1}^{n} \nabla_{\beta} q(z_i, \tilde{\beta}_n) \nabla q(z_i, \tilde{\beta}_n) \xrightarrow{p} B_0.
\]

From the first line in (6.5), and since \( A_0 \) is nonsingular by Assumption 6.1*(d), it follows that also Assumption 5.1(e) holds.

We have thus verified that under the assumptions maintained by the theorem all conditions of Assumption 5.1 hold. It now follows from Theorem 5.1 that

\[
    n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{p} N(0, A_0^{-1} B_0 A_0^{-1})
\]

as claimed. The remaining claim that \( \hat{A}_n \xrightarrow{p} A_0 \) and \( \hat{B}_n \xrightarrow{p} B_0 \) follows as a special case of (6.5).

6.1.2 Nonlinear Least Squares

For reasons of convenience we briefly repeat the basic setup for the nonlinear least squares estimator given in Section 4. Consider the i.i.d. data process \( z_i = [y_i, x_i] \) with \( y_i \in \mathbb{R} \) and \( x_i \in \mathbb{R}^p \). Frequently it will be of interest to model \( E[y_i | x_i] \) as a function of the explanatory variables \( x_i \). Now let \( g(x_i, \beta) \) be a parametric model for \( E[y_i | x_i] \), where \( g : \mathbb{R}^p \times B \to \mathbb{R} \) and \( B \subseteq \mathbb{R}^p \).

We have a correctly specified model for the conditional mean if

\[
    E[y_i | x_i] = g(x_i, \beta_0)
\]

for some \( \beta_0 \in B \).

Define \( \epsilon_i = y_i - g(x_i, \beta_0) \), then we can think of the data as being described by the following nonlinear regression model

\[
    y_i = g(x_i, \beta_0) + \epsilon_i, \quad i \in \mathbb{N},
\]

(6.6)
where the disturbances \( \epsilon_i \in \mathbb{R} \) satisfy \( E[\epsilon_i \mid x_i] = 0 \). The objective function of the NLS estimator is now given by

\[
Q_n(z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta),
\]

where

\[
q(z_i, \beta) = \frac{1}{2} [y_i - g(x_i, \beta)]^2
\]

with

\[
\nabla_{\beta}^2 q(z_i, \beta) = -\nabla_{\beta}^2 g(x_i, \beta) [y_i - g(x_i, \beta)],
\]

Assuming the existence of the derivatives note that

\[
\nabla g(x_i, \beta) \nabla g(x_i, \beta) = \nabla g(x_i, \beta) \nabla g(x_i, \beta) - \nabla g(x_i, \beta) \nabla g(x_i, \beta) [y_i - g(x_i, \beta)].
\]

We now specify a list of assumptions that ensures the asymptotic normality of the NLS estimator. We note that the assumptions do not maintain that the explanatory variables \( x_i \) and the disturbances \( \epsilon_i \) are independent.

**Assumption 6.2** (a) \( g(.,.) \) is a real valued function on \( \mathbb{R}^{p_x} \times B \), (b) \( B \subseteq \mathbb{R}^{p_y} \) is compact. (c) \( g(., \beta) \) is Borel measurable for each \( \beta \in B \), and \( g(x, .) \) is continuous for each \( x \in \mathbb{R}^{p_x} \). (d) \( z_i = [y_i, x_i] \) with \( y_i \in \mathbb{R} \) and \( x_i \in \mathbb{R}^{p_x} \) is a sequence of identically and independently distributed random vectors. (e) \( E[y_i \mid x_i] = g(x_i, \beta_0) \) for \( \beta_0 \in B \). (f) \( E[y_i - g(x_i, \beta_0)]^2 < \infty \) and \( E \sup_{\beta \in B} g(x_i, \beta)^2 < \infty \). (g) \( E[y_i - g(x_i, \beta_0) - g(x_i, \beta)]^2 > 0 \) for \( \beta \neq \beta_0 \).

Assumption 6.2 collects the assumptions maintained for the consistency result for the nonlinear least squares estimator given as Theorem 4.2. We need the following additional assumption.

**Assumption 6.2** * (a) \( \beta_0 \) is an interior point of \( B \). (b) \( g(x, .) \) is twice continuously differentiable at every interior point \( \beta \in B \) for each \( x \in \mathbb{R}^{p_x} \). (c) \( E \sup_{\beta \in B'} \| \nabla_{\beta} g(x_i, \beta) \| |y_i - g(x_i, \beta)|^2 < \infty \), \( E \sup_{\beta \in B'} \| \nabla_{\beta} g(x_i, \beta) \|^2 < \infty \), and \( E \sup_{\beta \in B'} \| \nabla_{\beta} g(x_i, \beta) \|^2 |y_i - g(x_i, \beta)|^2 < \infty \) where \( B' \) is a closed ball around \( \beta_0 \) that lies in the interior of \( B \). (d) \( E \nabla_{\beta} g(x_i, \beta_0) \nabla_{\beta} g(x_i, \beta_0) \) is non-singular.

We now have the following result concerning the asymptotic normality of the nonlinear least squares estimator in case of i.i.d. data.
Theorem 6.2 Consider the nonlinear least squares estimators \( \hat{\beta}_n \) defined as minimizers of (6.7). Suppose Assumption 6.2 and 6.2* hold, then

\[
n^{1/2}(\hat{\beta}_n - \beta_0) \overset{D}{\to} N(0, A_0^{-1}B_0A_0^{-1})
\]

where

\[
A_0 = E \left[ \nabla_\beta g(x_i, \hat{\beta}_n) \nabla_\beta g(x_i, \beta_0) \right],
\]

\[
B_0 = E \left[ (y_i - g(x_i, \hat{\beta}_n))^2 \nabla_\beta g(x_i, \hat{\beta}_n) \nabla_\beta g(x_i, \beta_0) \right],
\]

and \( A_0 \) positive definite. Furthermore, let

\[
\hat{A}_n = n^{-1} \sum_{i=1}^n \nabla_\beta g(x_i, \hat{\beta}_n) \nabla_\beta g(x_i, \hat{\beta}_n),
\]

\[
\hat{B}_n = n^{-1} \sum_{i=1}^n (y_i - g(x_i, \hat{\beta}_n))^2 \nabla_\beta g(x_i, \hat{\beta}_n) \nabla_\beta g(x_i, \hat{\beta}_n),
\]

then

\[
\hat{A}_n \overset{p}{\to} A_0, \quad \hat{B}_n \overset{p}{\to} B_0,
\]

and thus

\[
\hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1} \overset{p}{\to} A_0^{-1}B_0A_0^{-1}.
\]

Proof: To prove the theorem we verify that Assumptions 6.2 and 6.2* imply Assumptions 6.1 and 6.1* with

\[
q(z, \beta) = [y - g(x, \beta)]^2
\]

and \( z = [y, x] \), \( z_i = [y_i, x_i] \) and \( Z = R^{p_\beta} \), \( p_\beta = p_x + 1 \).

In proving Theorem 4.2 to establish that \( \hat{\beta}_n \overset{p}{\to} \beta_0 \) as \( n \to \infty \) we have verified that

Assumptions 6.2 implies Assumption 6.1. Assumptions 6.2*(a),(b) follow immediately from Assumptions 6.1*(a),(b). Next observe that using (6.8) we have

\[
\|\nabla_\beta q(z_i, \beta)\|^2 = \|\nabla_\beta g(x_i, \beta)\|^2 [y_i - g(x_i, \beta)]^2,
\]

\[
\|\nabla_\beta g(z_i, \beta)\| \leq \|\nabla_\beta g(x_i, \beta)\|^2 + \|\nabla_\beta g(x_i, \beta)\| |y_i - g(x_i, \beta)|.
\]

Given this Assumption 6.2*(c) clearly implies Assumption 6.1*(c). By Assumption 6.2(c) we have \( E[y_i - g(x_i, \beta_0) \mid x_i] = 0 \), and consequently

\[
E \nabla_{\beta \beta} g(x_i, \beta_0) [y_i - g(x_i, \beta_0)]
\]

\[
= EE \{ \nabla_{\beta \beta} g(x_i, \beta_0) [y_i - g(x_i, \beta_0)] \mid x_i \}
\]

\[
= E \nabla_{\beta \beta} g(x_i, \beta_0) E [y_i - g(x_i, \beta_0) \mid x_i] = 0.
\]
Using this it follows from (6.8) that
\[ E\nabla_{\beta\beta}q(z_i, \beta_0) = E\nabla_{\beta}g(x_i, \beta_0)\nabla_{\beta}g(x_i, \beta_0) - E\nabla_{\beta\beta}g(x_i, \beta_0) [y_i - g(x_i, \beta_0)] \]
\[ = E\nabla_{\beta}g(x_i, \beta_0)\nabla_{\beta}g(x_i, \beta_0). \]

Nonsingularity of \( E\nabla_{\beta\beta}q(z_i, \beta_0) \) follows from Assumption 6.2(d). Thus also Assumption 6.1(d) holds. Having verified that all assumptions of Theorem 6.1 hold under the assumptions of Theorem 6.2, the claims of the latter theorem follow from the former.

**Remark:** Suppose the assumptions of Theorem 6.2 hold. Assume furthermore that the disturbances are conditional homoskedastic, i.e., \( E[(y_i - g(x_i, \beta_0))^2 \mid x_i] = \sigma^2 \), then \( B_0 = \sigma^2 A_0 \) and hence
\[ n^{1/2}(\hat{\beta}_n - \beta_0) \overset{D}{\rightarrow} N(0, \sigma^2 A_0^{-1}). \]

Observing that \( E[y_i - g(x_i, \beta_0)^2 \mid x_i] = \sigma^2 \) implies that \( E[y_i - g(x_i, \beta_0)^2] = \sigma^2 \) it is seen that
\[ \hat{\sigma}^2_n = n^{-1} \sum_{i=1}^{n} (y_i - g(x_i, \hat{\beta}_n))^2 \overset{P}{\rightarrow} \sigma^2. \]

Next consider as a further special case the linear regression model, i.e., \( g(x_i, \beta) = x_i \beta \). Then \( \nabla_{\beta}g(x_i, \beta) = x'_i \), \( A_0 = Ex'_ix_i \), and \( \hat{A}_n = n^{-1} \sum_{i=1}^{n} x'_ix_i \). This shows that, as expected, we obtain our previous result concerning the asymptotic distribution of the OLS estimator as a special case of the above theorem.

### 6.1.3 Maximum Likelihood

For reasons of convenience we briefly repeat the basic setup for the maximum likelihood estimator given in Section 4. Consider the i.i.d. data process \( z_i = [y_i, x_i] \) with \( y_i \in Y \subseteq \mathbb{R}^p \) and \( x_i \in X \subseteq \mathbb{R}^p \), where \( Y \) and \( X \) are Borel sets. Let \( F(y \mid x) \) denote the conditional distribution of \( y_i \) given \( x_i = x \). We assume that this distribution can be described by a conditional density,
\[ f(y \mid x; \beta_0), \]
with respect to some (\( \sigma \)-finite) measure \( \mu \) defined on the \( Y \), and \( \beta_0 \in B \subseteq \mathbb{R}^p \).

The objective function of the conditional maximum likelihood estimator is then given by
\[ Q_n(z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta), \tag{6.9} \]
where
\[ q(z_i, \beta) = -\ln [f(y_i \mid x_i; \beta)]. \]
now represents the negative conditional log-likelihood for observation $i$. We maintain the following assumptions.

**Assumption 6.3** The data process $z_i = [y_i, x_i]$ is i.i.d. with $y_i \in Y \subseteq \mathbb{R}^{p_y}$ and $x_i \in X \subseteq \mathbb{R}^{p_x}$, where $Y$ and $X$ are Borel sets. Let $B \subseteq \mathbb{R}^{p_x}$ and consider the family of conditional densities $\{f(y | x; \beta) : \beta \in B, x \in X\}$ with respect to some ($\sigma$-finite) measure $\mu$. Then for all $x \in X$ the (true) conditional distribution of $y_i$ given $x_i = x$ can be represented by the conditional density $f(y | x; \beta_0)$ for some $\beta_0 \in B$. Consider the real valued function $q$ on $Z \times B$, $Z = Y \times X$, defined as $q(z, \beta) \equiv -\ln f(y | x; \beta)$ for all $z = [y, x] \in Y \times X \subseteq \mathbb{R}^{p_y + p_x}$, and all $\beta \in B$, then: (a) $B \subseteq \mathbb{R}^{p_x}$ is compact. (b) $q(., \beta)$ is Borel measurable for each $\beta \in B$, and $q(z, \cdot)$ is continuous for each $z \in Z$. (c) $E\sup_{\beta \in B} |q(z_i, \beta)| < \infty$. (d) $\beta_0$ is a unique minimizer of $E_q(z_i, \beta)$.

Assumption 6.3 collects the assumptions maintained for the consistency result for conditional maximum likelihood estimators given as Theorem 4.3. We need the following additional assumption.

**Assumption 6.3** *(a) $\beta_0$ is an interior point of $B$. (b) $q(z, \cdot)$ is twice continuously differentiable at every interior point $\beta \in B$ for each $z \in Z$. (c) $E\sup_{\beta \in B} \|\nabla_{\beta} q(z_i, \beta)\| < \infty$ and $E\sup_{\beta \in B} \|\nabla_{\beta} q(z_i, \beta)\|^2 < \infty$ where $B'$ is a closed ball around $\beta_0$ that lies in the interior of $B$. (d) $E\nabla_{\beta} q(z_i, \beta_0)$ is nonsingular (and symmetric in light of (b)).

We now have the following result concerning the asymptotic normality of conditional maximum likelihood estimators in case of i.i.d. data. Of course, the result also covers (unconditional) maximum likelihood estimators as a special case.

**Theorem 6.3** Consider the conditional maximum likelihood estimators $\hat{\beta}_n$, defined as minimizers of (6.9). Suppose Assumption 6.3 and 6.3* hold, then

$$n^{1/2}(\hat{\beta}_n - \beta_0) \overset{D}{\to} N(0, A_0^{-1}B_0A_0^{-1})$$

where

\[
A_0 = E\nabla_{\beta} q(z_i, \beta_0) = -E\nabla_{\beta} \ln [f(y_i | x_i; \beta_0)], \quad B_0 = E\nabla_{\beta} q(z_i, \beta_0)\nabla_{\beta} q(z_i, \beta_0) = E\nabla_{\beta} \ln [f(y_i | x_i; \beta_0)] \nabla_{\beta} \ln [f(y_i | x_i; \beta_0)],
\]
and $A_0$ positive definite. Furthermore, let

$$
\hat{A}_n = n^{-1} \sum_{i=1}^{n} \nabla_{\beta \beta} q(z_i, \hat{\beta}_n) \\
\hat{B}_n = n^{-1} \sum_{i=1}^{n} \nabla_{\beta} q(z_i, \hat{\beta}_n) \nabla_{\beta} g(z_i, \hat{\beta}_n)
$$

$$
= n^{-1} \sum_{i=1}^{n} \nabla_{\beta} \ln \left[ f(y_i \mid x_i; \hat{\beta}_n) \right] \nabla_{\beta} \ln \left[ f(y_i \mid x_i; \hat{\beta}_n) \right],
$$

then

$$\hat{A}_n \xrightarrow{p} A_0, \quad \hat{B}_n \xrightarrow{p} B_0,$$

and thus

$$\hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1} \xrightarrow{p} A_0^{-1} B_0 A_0^{-1}.$$

If additionally

$$\int \sup_{\beta \in B'} \| \nabla_{\beta} f(y \mid x ; \beta) \| \, d\mu(y) < \infty, \quad (6.10)$$

$$\int \sup_{\beta \in B'} \| \nabla_{\beta \beta} f(y \mid x ; \beta) \| \, d\mu(y) < \infty,$$

for all values of $x$ outside a set $N$ with $P_x(N) = 0$, where $P_x$ is the distribution of $x$, then

$$A_0 = B_0.$$

**Proof:** To prove the theorem we verify that Assumptions 6.3 and 6.3* imply Assumptions 6.1 and 6.1* with

$$q(z, \beta) = -\ln \left[ f(y \mid x ; \beta) \right]$$

and $z = [y, x] \in Z$, $Z = Y \times X \subseteq \mathbb{R}^p$, $p_z = p_y + p_x$, and $z_i = [y_i, x_i]$.

In proving Theorem 4.3 to establish that $\hat{\beta}_n \xrightarrow{p} \beta_0$ as $n \to \infty$ we have verified that Assumptions 6.3 implies Assumption 6.1. Assumptions 6.3* is identical to Assumption 6.1* with $q(z, \beta) = -\ln \left[ f(y \mid x ; \beta) \right]$. Thus all parts of Theorem 6.3, except for the last claim, follow immediately from Theorem 6.1.

To prove the last claim that

$$E \nabla_{\beta \beta} q_t(z_t, \beta_0) = E \left[ \nabla_{\beta \beta} q_t(z_t, \beta_0) \nabla_{\beta \beta} q_t(z_t, \beta_0) \right], \quad (6.11)$$

observe that

$$\nabla_{\beta \beta} q_t(z_t, \beta_0) = f_t^{-2} \nabla_{\beta} f_t \nabla_{\beta} f_t - f_t^{-1} \nabla_{\beta \beta} f_t$$

$$= \nabla_{\beta} q_t(z_t, \beta_0) \nabla_{\beta} q_t(z_t, \beta_0) - f_t^{-1} \nabla_{\beta \beta} f_t,$$
with \( f_t = f(y_t | x_t; \beta_0) \). Under the maintained assumptions both \( E \nabla_{\beta\beta} q_t(z_t, \beta_0) \) and \( E \nabla_{\beta\beta} q_t(z_t, \beta_0) \) exist and are finite. Clearly then also the expected value of the expression \( f_t^{-1} \nabla_{\beta\beta} f_t \) exists and is finite. Hence for (6.11) to hold it remains to be shown that \( E f_t^{-1} \nabla_{\beta\beta} f_t \) is zero. For this it clearly suffices to show that the integral of \( f_t^{-1} \nabla_{\beta\beta} f_t \) w.r.t. \( f(y_t | x_t; \beta_0) \) is zero for all values of \( x_t \) outside the set \( N \) with \( P_{x}(N) = 0 \), where \( P_x \) is the distribution of \( x_t \). This integral, dropping subscripts \( t \), clearly reduces to

\[
\int \nabla_{\beta\beta} f(y | x; \beta_0) d\mu(y).
\]

Since

\[
\int f(y | x; \beta) d\mu(y) = 1
\]

for all \( \beta \) clearly

\[
\nabla_{\beta\beta} \int f(y | x; \beta) d\mu(y) = 0,
\]

\[
\nabla_{\beta\beta} \int f(y | x; \beta) d\mu(y) = 0.
\]

Hence we obtain the desired result if the operations of differentiation and integration can be interchanged. However this interchange is permitted in light of (6.10), which completes the proof.

**Remark:** As remarked, in case the (conditional) likelihood function does not correspond to the (conditional) density we call the optimizer of the (conditional) likelihood function the (conditional) quasi-maximum likelihood estimator. This case is covered by Theorem 6.1 with \( q(z, \beta) = -\ln [ f(y | x; \beta) ] \). In this case in general \( A_0 \neq B_0 \), and hence the variance covariance matrix of the limiting distribution of the (conditional) quasi-maximum likelihood estimator does not simplify and is given by the “sandwich formula” \( A_0^{-1} B_0 A_0^{-1} \).

### 6.2 Generalized Method of Moments Estimators

As discussed, estimators corresponding to objective functions of the form

\[
R_n(\omega, \beta) = Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \beta) = \left[ n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right]^{'} \hat{\Xi}_n \left[ n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right]
\]

where \( q(z, \beta) \) is a real vector valued function defined on \( Z \times B \) taking its values in \( \mathbb{R}^{p_q} \), \( \hat{\Xi}_n \) is a \( p_q \times p_q \) positive semi-definite symmetric matrix and \( \hat{\tau}_n \) is the vector of diagonal and upper diagonal elements of \( \hat{\Xi}_n \), are typically called generalized method of moments estimators. In the following we will give a general catalogue of assumptions for the asymptotic normality of generalized method of moments estimators for the case of i.i.d. data.
Assumption 6.4 (a) $q : Z \times B \to \mathbb{R}^p$, where $Z \subseteq \mathbb{R}^p$ and $B \subseteq \mathbb{R}^p$ are Borel sets. (b) $B$ is furthermore compact. (c) $q(., \beta)$ is Borel measurable for each $\beta \in B$, and $q(z, .)$ is continuous for each $z \in Z$. (d) $z_i$ is a sequence of identically and independently distributed random vectors that take their values in $\mathbb{R}^p$. (e) $E \sup_{\beta \in B} ||q(z_i, \beta)|| < \infty$. (f) $\hat{z}_n \overset{P}{\to} z_0$ as $n \to \infty$, where the $p_q \times p_q$ (stochastic real valued) matrices $\hat{z}_n$ are symmetric positive semi-definite and the $p_q \times p_q$ (real) matrix $z_0$ is symmetric positive definite. (g) $\beta_0$ is the unique solution of $Eq(z_i, \beta) = 0$.

Assumption 6.4 collects the assumptions maintained for the consistency result for generalized method of moments estimators given as Theorem 4.4. We need the following additional assumption.

Assumption 6.4* (a) $\beta_0$ is an interior point of $B$. (b) $q(z, .)$ is continuously differentiable at every interior point $\beta \in B$ for each $z \in Z$. (c) $E \sup_{\beta \in B} ||\nabla \beta q(z_i, \beta)|| < \infty$ and $E \sup_{\beta \in B} ||q(z_i, \beta)||^2 < \infty$ where $B'$ is a closed ball around $\beta_0$ that lies in the interior of $B$. (d) $E\nabla \beta q(z_i, \beta_0)$ has full column rank.

We now have the following general result concerning the asymptotic normality of generalized method of moments estimators in case of i.i.d. data.

Theorem 6.4 Consider the generalized method of moments estimators $\hat{\beta}_n$ defined as minimizers of (6.12). Suppose Assumption 6.4 and 6.4* hold, then

$$n^{1/2}(\hat{\beta}_n - \beta_0) \overset{D}{\to} N(0, A_0^{-1}B_0A_0^{-1})$$

where

$$A_0 = G_0^\prime \Xi_0 G_0, \quad B_0 = G_0^\prime \Xi_0 \Lambda_0 \Xi_0 G_0,$$

and

$$G_0 = E\nabla \beta q(z_i, \beta_0), \quad \Lambda_0 = Eq(z_i, \beta_0)q(z_i, \beta_0)^\prime,$$

and $A_0$ positive definite. Furthermore, let

$$\hat{A}_n = \hat{G}_n^\prime \hat{\Xi}_n \hat{G}_n, \quad \hat{B}_n = \hat{G}_n^\prime \hat{\Xi}_n \hat{\Lambda}_n \hat{\Xi}_n \hat{G}_n,$$

with

$$\hat{G}_n = n^{-1} \sum_{i=1}^n \nabla \beta q(z_i, \hat{\beta}_n), \quad \hat{\Lambda}_n = n^{-1} \sum_{i=1}^n q(z_i, \hat{\beta}_n)q(z_i, \hat{\beta}_n)^\prime$$

then

$$\hat{G}_n \overset{P}{\to} G_0, \quad \hat{\Lambda}_n \overset{P}{\to} \Lambda_0,$$

and thus

$$\hat{A}_n \overset{P}{\to} A_0, \quad \hat{B}_n \overset{P}{\to} B_0,$$

and furthermore

$$\hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1} \overset{P}{\to} A_0^{-1}B_0A_0^{-1}. $$

45
Proof: To prove the theorem we first verify that Assumptions 6.4 and 6.4* imply Assumptions 5.2 for $S_n(z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^n q(z_i, \beta)$ and $\hat{\beta}_n$, which are defined as minimizers over $B$ of the objective function $Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \beta)$ defined in (6.12).

We first note that in light of Assumption 6.4 it follows from Theorem 4.4 that $\hat{\beta}_n \overset{p}{\to} \beta_0$ as $n \to \infty$, where $\beta_0$ is an interior point of $B$ by Assumption 6.4*(a). By definition $\hat{\beta}_n \in B$. Given this and given Assumption 6.4(f) clearly all conditions postulated in Assumption 5.2(c) hold.

Assumptions 5.2(a),(b) hold in light of Assumptions 6.4(a),(b),(c) and Assumption 6.4*(b). Since $\hat{\beta}_n \overset{p}{\to} \beta_0$ as $n \to \infty$, where $\beta_0$ is an interior point of $B$, it follows that $\hat{\beta}_n$ lies in the interior of $B$ except for $\omega$-sets whose probability tends to zero. That is

$$n^{1/2} \nabla_{\beta} S_n(z_1, \ldots, z_n, \hat{\beta}_n) = n^{-1} \sum_{i=1}^n \nabla_{\beta} q(z_i, \hat{\beta}_n) \Xi_n S_n(z_1, \ldots, z_n, \hat{\beta}_n) = 0$$

except on $\omega$-sets whose probability tends to zero. Thus also Assumption 5.2(d) holds.

Next consider

$$n^{1/2} S_n(z_1, \ldots, z_n, \beta_0) = n^{-1/2} \sum_{i=1}^n q(z_i, \beta_0).$$

In light of Assumption 6.4(d),(e),(g) and 6.4*(c) the random vectors $q(z_i, \beta_0)$ are seen to be i.i.d. with mean vector $E q(z_i, \beta_0) = 0$ and finite variance-covariance matrix $\Lambda_0 = E q(z_i, \beta_0)q(z_i, \beta_0)'$. It now follows from the CLT for i.i.d. random vectors that

$$n^{1/2} S_n(z_1, \ldots, z_n, \beta_0) \overset{D}{\to} \zeta$$

where $\zeta \sim N(0, \Lambda_0)$. Hence Assumption 5.2(g) is satisfied with $\zeta \sim N(0, \Lambda_0)$.

Next observe that all assumptions maintained by the uniform law of large numbers given as Theorem 3.3 are satisfied for $\nabla_{\beta} q(z_i, \beta)$ and $q(z_i, \beta)q(z_i, \beta)'$. The needed condition that

$$E \sup_{\beta \in B'} \| \nabla_{\beta} q(z_i, \beta) \| < \infty \quad \text{and} \quad E \sup_{\beta \in B'} \| q(z_i, \beta)q(z_i, \beta)' \| < \infty$$

is satisfied in light of Assumption 6.4*(c). Hence

$$\sup_{\beta \in B'} n^{-1} \sum_{i=1}^n \nabla_{\beta} q(z_i, \beta) - E \nabla_{\beta} q(z_i, \beta) \overset{p}{\to} 0,$$

$$\sup_{\beta \in B'} n^{-1} \sum_{i=1}^n q(z_i, \beta)q(z_i, \beta)' - E q(z_i, \beta)q(z_i, \beta)' \overset{p}{\to} 0,$$

as $n \to \infty$. 

46
Now let \( \tilde{\beta}_n \) be any sequence of random vectors with \( \tilde{\beta}_n \xrightarrow{p} \beta_0 \) as \( n \to \infty \). Furthermore define \( \tilde{\beta}_n = \tilde{\beta}_n \) if \( \tilde{\beta}_n \in B' \) and \( \tilde{\beta}_n = \beta_0 \) if \( \tilde{\beta}_n \notin B' \). Then clearly \( \tilde{\beta}_n \xrightarrow{p} \beta_0 \) as \( n \to \infty \), and furthermore

\[
 n^{-1} \sum_{i=1}^{n} \nabla_{\beta'} q(z_i, \tilde{\beta}_n) - n^{-1} \sum_{i=1}^{n} \nabla_{\beta'} q(z_i, \tilde{\beta}_n) = o_p(1), \quad (6.14)
\]

\[
 n^{-1} \sum_{i=1}^{n} q(z_i, \tilde{\beta}_n)q(z_i, \tilde{\beta}_n)' - n^{-1} \sum_{i=1}^{n} q(z_i, \tilde{\beta}_n)q(z_i, \tilde{\beta}_n)' = o_p(1).
\]

Because of (6.13) it follows from Theorem 3.4 that

\[
 n^{-1} \sum_{i=1}^{n} \nabla_{\beta'} q(z_i, \tilde{\beta}_n) \xrightarrow{p} G_0, \quad n^{-1} \sum_{i=1}^{n} q(z_i, \tilde{\beta}_n)q(z_i, \tilde{\beta}_n)' \xrightarrow{p} \Lambda_0. \quad (6.15)
\]

Combining (6.14) and 6.15 we have

\[
 \nabla_{\beta'} S_n(z_1, \ldots, z_n, \tilde{\beta}_n) = n^{-1} \sum_{i=1}^{n} \nabla_{\beta'} q(z_i, \tilde{\beta}_n) \xrightarrow{p} G_0, \quad (6.16)
\]

\[
 n^{-1} \sum_{i=1}^{n} q(z_i, \tilde{\beta}_n)q(z_i, \tilde{\beta}_n)' \xrightarrow{p} \Lambda_0.
\]

\( G_0 \) has full column rank by Assumption 6.4* (d). This shows that also Assumption 5.2(e) holds.

We have thus verified that under the assumptions maintained by the theorem all conditions of Assumption 5.2 hold. It now follow from Theorem 5.2 that \( n^{1/2}(\tilde{\beta}_n - \beta_0) \xrightarrow{D} N(0, \Lambda_0^{-1}B_0\Lambda_0^{-1}) \) as claimed. The claims that \( \tilde{G}_n \xrightarrow{p} G_0 \) and \( \tilde{\Lambda}_n \xrightarrow{p} \Lambda_0 \) follows as a special case of (6.16). The remaining claims are trivial. \( \blacksquare \)

**Remark:** We next discuss how we should choose the weighting matrix \( \hat{\Xi}_n \) “optimally”, given a certain set of moment conditions. For purposes of this discussion let \( \hat{\beta}_n(\hat{\Xi}_n) \) denote the GMM estimator corresponding to \( \hat{\Xi}_n \) and let the corresponding limiting variance covariance be denoted by \( \Psi(\hat{\Xi}_n) \), which is in general, according to the above theorem, given by

\[
 \Psi(\hat{\Xi}_n) = (G_0'\hat{\Xi}_nG_0)^{-1}G_0'\hat{\Xi}_n\Lambda_0\hat{\Xi}_nG_0(G_0'\hat{\Xi}_nG_0)^{-1}. \quad (6.17)
\]

Intuitively it seems reasonable to try to use the inverse of the variance covariance matrix of the moment vector, i.e., \( \Lambda_0^{-1} \) as a weighting matrix. Since \( \Lambda_0 \) is unoberved we may use some consistent estimator instead, e.g.,

\[
 \hat{\Lambda}_n = n^{-1} \sum_{i=1}^{n} q(z_i, \tilde{\beta}_n)q(z_i, \tilde{\beta}_n)'.
\]
where \( \hat{\beta}_n \) is some consistent estimator for \( \beta \). Such a consistent estimator can, in particular, be obtained by estimating the model in a first stage with the weights matrix equal to the identity matrix, that is \( \hat{\beta}_n \) is defined as the minimizer of

\[
\left( n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right)' \left( n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right). 
\]

Now let \( \hat{\beta}_n(\tilde{\Lambda}_n^{-1}) \) denote the GMM estimator corresponding to \( \tilde{\Lambda}_n^{-1} \) and let the corresponding limiting variance covariance be denoted by \( \Psi(\tilde{\Lambda}_n^{-1}) \). Since \( \tilde{\Lambda}_n \xrightarrow{p} \Lambda_0 \) it then follows as a special case of (6.17) that

\[
\Psi(\tilde{\Lambda}_n^{-1}) = (G'_0 \Lambda_0^{-1} G_0)^{-1} G'_0 \Lambda_0^{-1} \Lambda_0 G_0 G'_0 \Lambda_0^{-1} G_0)^{-1} = (G'_0 \Lambda_0^{-1} G_0)^{-1}.
\]

It follows furthermore from the lemma given below that

\[
\Psi(\hat{\beta}_n(\tilde{\Lambda}_n^{-1})) - \Psi(\hat{\beta}_n(\Lambda_n^{-1}))
\]

is positive definite, and thus \( \hat{\beta}_n(\tilde{\Lambda}_n^{-1}) \) is the the asymptotically efficient GMM estimator, for the given set of moment conditions. By construction the objective function of \( \hat{\beta}_n(\tilde{\Lambda}_n^{-1}) \) is given by

\[
\left( n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right)' \left( n^{-1} \sum_{i=1}^{n} q(z_i, \hat{\beta}_n(\tilde{\Lambda}_n^{-1}))' \right)^{-1} \left( n^{-1} \sum_{i=1}^{n} q(z_i, \hat{\beta}_n(\tilde{\Lambda}_n^{-1})) \right).
\]

A further estimator is one that corresponds to the objective function

\[
\left( n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right)' \left( n^{-1} \sum_{i=1}^{n} q(z_i, \beta)' q(z_i, \beta) \right)^{-1} \left( n^{-1} \sum_{i=1}^{n} q(z_i, \beta) \right)
\]

which is called the continuous updating GMM estimator. The properties of this estimator are currently analyzed within the context of the so-called “weak instrument” literature.

In the above argumentation we have utilized the following well known lemma, which is frequently used in comparing respective variance covariance matrices:

**Lemma 6.4** Let \( G \) be a \( p \times g \) matrix of full column rank, let \( \Xi \) be a nonsingular \( p \times p \) matrix, and let \( \Lambda \) be a \( p \times p \) symmetric positive definite matrices, then

\[
(G' \Xi G)^{-1} G' \Xi \Lambda \Xi' G(G' \Xi G)^{-1} - (G' \Lambda^{-1} G)^{-1} \tag{6.18}
\]

is positive semidefinite.

**Proof:** Instead of proving (6.18) it is more convenient to prove that

\[
\Delta = G' \Lambda^{-1} G - (G' \Xi G)(G' \Xi \Lambda \Xi' G)^{-1}(G' \Xi G)
\]
is positive semidefinite. Define $D = \Lambda^{1/2} \Xi' G$, then

$$D(D'D)^{-1}D' = \Lambda^{1/2} \Xi' G (G' \Xi \Lambda \Xi' G)^{-1} G' \Xi \Lambda^{1/2}$$

and hence

$$\Delta = G' \Lambda^{-1/2} M \Lambda^{-1/2} G$$

with $M = [I_p - D(D'D)^{-1}D']$. Since $M$ is symmetric and idempotent we have

$$\Delta = P' P$$

with $P = M \Lambda^{-1/2} G$ which shows that $\Delta$ is positive semidefinite.
7 NUMERICAL OPTIMIZATION

7.1 Introduction

M-estimators are defined as minimizers (maximizers) of some objective function \( Q_n(z_1, \ldots, z_n, \beta) \). W.o.l.g. we will consider the case where the estimator is defined as a minimizer. As discussed, for nonlinear specifications it will often be difficult to get an explicit expression for the M-estimator, and we will have to resort to numerical methods in finding the optimizer of the objective function. Since the numerical optimization is performed for given observations on \( z_1, \ldots, z_n \) and for a given sample size \( n \) it proves convenient to use \( Q(\beta) \) for \( Q_n(z_1, \ldots, z_n, \beta) \). For additional literature on numerical optimization, see, e.g., Judge et al. (1985), Appendix B; Quandt (1983), Wooldridge (2002), ch. 12.7.

Most numerical optimization methods follow an iterative scheme. To start the optimization routine we need some initial (starting) parameter vector, say, \( \beta_0 \). Analogously, let \( \beta_g \) denote the parameter vector corresponding to the \( g \)-th iteration, and let \( Q(\beta_g) \) denote the corresponding value of the objective function. The following steps are then typically performed at the \( g \)-th iteration:

1. Determine a vector \( \xi_g \), called a step, and add \( \xi_g \) to \( \beta_{g-1} \) to determine a new trial parameter vector \( \beta_g \), i.e.,
   \[
   \beta_g = \beta_{g-1} + \xi_g.
   \]  
   (7.1)

2. Compute the objective function \( Q(\beta_g) \) and check if \( Q(\beta_g) < Q(\beta_{g-1}) \). If this is not the case modify the step until \( Q(\beta_g) < Q(\beta_{g-1}) \).

3. Check whether one or more termination criteria are satisfied, e.g., if for a prespecified small \( \varepsilon > 0 \):
   
   (a) \( (\beta_g - \beta_{g-1})' (\beta_g - \beta_{g-1}) < \varepsilon \)
   
   (b) \( Q(\beta_{g-1}) - Q(\beta_g) < \varepsilon \)
   
   (c) \( [\partial Q(\beta_g)/\partial \beta]' [\partial Q(\beta_g)/\partial \beta] < \varepsilon \)
   
   (d) The maximum number of iterations has been attained.

   If the termination criteria are not satisfied, start a new iteration though steps 1-3. If the termination criteria are satisfied stop the iterative process.

   Different numerical optimization routines differ in the determination of the step \( \xi_g \). In computing a step the numerical routine typically first determines a search direction, say \( \delta_g \), as well as a step length, say \( \lambda_g \geq 0 \), and then defines the step as
   
   \[
   \xi_g = \lambda_g \delta_g.
   \]  
   (7.2)

   Clearly, given \( \beta_{g-1} \) we seek a direction \( \delta_g \) in which the function declines. However, given a direction \( \delta_g \) in which the function declines, we have to be careful
of not going too far in that direction, since in going to far we may end up going uphill again. That is \( \delta_g \) and \( \lambda_g \) have to be chosen such that

\[
Q(\beta_{g-1} + \lambda_g \delta_g) < Q(\beta_{g-1}).
\]

Now consider the directional derivative

\[
\frac{\partial Q(\beta_{g-1} + \lambda \delta_g)}{\partial \lambda} \bigg|_{\lambda=0} = \left. \frac{\partial Q(\beta_{g-1} + \lambda \delta_g)}{\partial \beta} \right|_{\lambda=0} \delta_g = \left[ \frac{\partial Q(\beta_{g-1})}{\partial \beta} \right] \delta_g.
\]

To ensure that \( Q(\beta_{g-1} + \lambda \delta_g) \) is indeed a decreasing function of \( \lambda \) at least for \( \lambda \) sufficiently small, and thus to ensure that a \( \lambda_g \) exists such that \( Q(\beta_{g-1} + \lambda_g \delta_g) < Q(\beta_{g-1}) \), it hence necessary that the direction \( \delta_g \) is chosen such that

\[
\left[ \frac{\partial Q(\beta_{g-1})}{\partial \beta} \right] \delta_g < 0.
\]

Now suppose we choose

\[
\delta_g = -P_g \left[ \frac{\partial Q(\beta_{g-1})}{\partial \beta} \right] \tag{7.3}
\]

where \( P_g \) is a positive definite matrix, then

\[
\left[ \frac{\partial Q(\beta_{g-1})}{\partial \beta} \right] \delta_g = - \left[ \frac{\partial Q(\beta_{g-1})}{\partial \beta} \right] P_g \left[ \frac{\partial Q(\beta_{g-1})}{\partial \beta} \right] \leq 0
\]

where the inequality is strict as long as the gradient is non-zero. Combining (7.1)-(7.3) yields

\[
\beta_g = \beta_{g-1} - \lambda_g P_g \left[ \frac{\partial Q(\beta_{g-1})}{\partial \beta} \right]. \tag{7.4}
\]

It seems that most widely used numerical optimization algorithms in econometrics are based on iteration schemes of the form (7.4). Such algorithm are referred to as gradient methods. Different gradient methods differ in the selection of the direction matrix \( P_g \) (and step length \( \lambda_g \)).

### 7.2 Examples of Gradient Methods

#### 7.2.1 Method of Steepest Descent

This method corresponds to

\[
P_g = I_K
\]

where \( K \) is the dimension of \( \beta \). Although it can be shown that this method selects the direction of the initially steepest descent, this method may converge very slowly if the minimum is in a long and narrow valley. This method is not recommended in most cases.
7.2.2 Newton-Raphson Method

This method corresponds to

\[ P_g = \left[ \frac{\partial^2 Q(\beta_{g-1})}{\partial \beta \partial \beta'} \right]^{-1}, \]

i.e., \( P_g \) is equal to the inverse of the Hessian matrix evaluated at \( \beta_{g-1} \). This method can be motivated as follows: Consider the following Taylor expansion of the score vector around \( \beta_{g-1} \):

\[ \frac{\partial Q(\beta)}{\partial \beta'} = \frac{\partial Q(\beta_{g-1})}{\partial \beta'} + \frac{\partial^2 Q(\beta_{g-1})}{\partial \beta \partial \beta'} (\beta - \beta_{g-1}) + r \]

where \( r \) is a remainder term. Ignoring the remainder term our objective is to find \( \beta_g \) such that

\[ 0 = \frac{\partial Q(\beta_g)}{\partial \beta'} = \frac{\partial Q(\beta_{g-1})}{\partial \beta'} + \frac{\partial^2 Q(\beta_{g-1})}{\partial \beta \partial \beta'} (\beta_g - \beta_{g-1}) \]

which yields

\[ \beta_g = \beta_{g-1} - \left[ \frac{\partial^2 Q(\beta_{g-1})}{\partial \beta \partial \beta'} \right]^{-1} \frac{\partial Q(\beta_{g-1})}{\partial \beta'} \]

The remainder term is exactly equal to zero if the objective function is linear quadratic. Provided the Hessian is positive definite we achieve the minimum in one step and \( \beta_g \) is equal to the unique minimizer. In general the Hessian will only be positive definite in a neighborhood of the minimum. Hence, hence is \( \beta_{g-1} \) is far from the minimizer of the objective function, the Hessian evaluated at \( \beta_{g-1} \) may not be positive definite, and thus the algorithm may not find a step \( \xi_g \) that leads to a reduction in the objective function. For possible modifications that ensure that the direction matrix is positive definite see, e.g., Quandt (1983).

Another drawback of the Newton-Raphson is that it requires the computation of second order derivatives, which may be computationally time consuming, especially if analytic expressions for those methods are not available. Of course, the updating equation can be generalized by multiplying the inverse Hessian by a step length \( \lambda_g \).

7.2.3 Berndt, Hall, Hall and Hausman (BHHH) Method

Suppose the objective function is a (negative) likelihood function and given by

\[ Q(\beta) = n^{-1} \sum_{i=1}^{n} q_i(\beta), \]

where \( q_i(\beta) \) denotes the (negative) likelihood function corresponding to the \( i \)-th observation. Then we know from our discussion of the maximum likelihood
estimator that at the true parameter value $\beta^0$

$$E \frac{\partial^2 Q(\beta^0)}{\partial \beta \partial \beta'} = n^{-1} \sum_{i=1}^{n} E \frac{\partial^2 q_i(\beta^0)}{\partial \beta \partial \beta'} = n^{-1} \sum_{i=1}^{n} E \frac{\partial q_i(\beta^0)}{\partial \beta'} \frac{\partial q_i(\beta^0)}{\partial \beta}. $$

This motivated Berndt, Hall, Hall and Hausman (1974) to suggest an algorithm where the Hessian is replaced by the outer product of the first order derivatives, i.e.,

$$P_g = \left[ n^{-1} \sum_{i=1}^{n} \frac{\partial q_i(\beta_{g-1})}{\partial \beta'} \frac{\partial q_i(\beta_{g-1})}{\partial \beta} \right]^{-1}. $$

The corresponding updating equation is given by

$$\beta_g = \beta_{g-1} - \lambda_g \left[ n^{-1} \sum_{i=1}^{n} \frac{\partial q_i(\beta_{g-1})}{\partial \beta'} \frac{\partial q_i(\beta_{g-1})}{\partial \beta} \right]^{-1} \left[ n^{-1} \sum_{i=1}^{n} \frac{\partial q_i(\beta_{g-1})}{\partial \beta'} \right]. $$

By construction the direction matrix is always positive semidefinite and the algorithm only requires the computation of first order derivatives. Although the estimator is motivated by the properties of the likelihood function, the algorithm can also be applied for other estimators.

### 7.2.4 Gauss-Newton Method

Consider the nonlinear regression model discussed above. The objective function of the nonlinear least squares estimator is given by

$$Q(\beta) = n^{-1} \sum_{i=1}^{n} q_i(\beta)$$

where

$$q_i(\beta) = [y_i - g(x_i, \beta)]^2 / 2.$$ 

As demonstrated above

$$\frac{\partial q(z_i, \beta)}{\partial \beta} = - \frac{\partial g(x_i, \beta)}{\partial \beta} [y_i - g(x_i, \beta)],$$

$$\frac{\partial^2 q(z_i, \beta)}{\partial \beta \partial \beta'} = \frac{\partial g(x_i, \beta)}{\partial \beta} \frac{\partial g(x_i, \beta)}{\partial \beta'} - \frac{\partial^2 g(x_i, \beta)}{\partial \beta \partial \beta'} [y_i - g(x_i, \beta)].$$

Now let $\beta^0$ denote the true parameter value, then we have $E \left[ [y_i - g(x_i, \beta)] \mid x_i \right] = 0$ by the assumptions maintained for the nonlinear regression model. Consequently

$$E \frac{\partial^2 q(z_i, \beta^0)}{\partial \beta \partial \beta'} = E \frac{\partial g(x_i, \beta^0)}{\partial \beta'} \frac{\partial g(x_i, \beta^0)}{\partial \beta}.$$ 

This motivates the specification of the direction matrix as

$$P_g = \left[ n^{-1} \sum_{i=1}^{n} \frac{\partial g_i(\beta_{g-1})}{\partial \beta'} \frac{\partial g_i(\beta_{g-1})}{\partial \beta} \right]^{-1}. $$
where \( g_i(\beta) = g(x_i, \beta) \). Clearly by construction \( P_g \) is then positive semidefinite, and the corresponding updating equation is given by

\[
\beta_g = \beta_{g-1} + \lambda_g \left[ n^{-1} \sum_{i=1}^{n} \frac{\partial g_i(\beta_{g-1})}{\partial \beta} \right]^{-1} \left[ n^{-1} \sum_{i=1}^{n} \frac{\partial g_i(\beta_{g-1})}{\partial \beta} [y_i - g_i(\beta_{g-1})] \right].
\]

This optimization algorithm is known as the Gauss or Gauss-Newton method.

### 7.3 Miscellaneous Remarks

To be successful in numerically optimizing an objective functions it is generally essential to initiate the iteration scheme form “good” starting values. Unfortunately, finding good starting values is more of an art than a science. Some thoughts:

- Suppose your model is of the form
  
  \[
y_i = \alpha f(x_i, \gamma) + u_i,
  \]
  
  or
  
  \[
y_i = \alpha + f(x_i, \gamma) + u_i,
  \]
  
  where \( \alpha \) is a scalar. In this case it is often helpful to first fix the starting values for \( \gamma \) as constants, and only estimate \( \alpha \). You can then use the estimate of \( \alpha \) together with the initially fixed values of \( \gamma \) as new starting values to estimate \( \alpha \) and \( \gamma \) jointly. This avoids the problem that the starting values are such that the regression line corresponding to the starting values lies totally outside the scatter of observations.

- Let \( \beta = [\alpha', \gamma'] \). Often we are more confident about the values of some parameters, say, \( \gamma \), than others parameters, say \( \alpha \). In this case it often makes sense to fix the starting values of \( \gamma \) as constants and initially only optimize the objective function w.r.t. \( \alpha \) (for given values of \( \gamma \)). This reduces the dimensionality of the optimization problem. Once optimizers for \( \alpha \) have been found we may start a new optimization scheme from the previously obtained estimates of \( \alpha \) and the initially fixed values of \( \gamma \).

- Ideas for starting values can often also be gotten by first estimating an approximate version of the model that is linear (or partially linear) in parameter.

- Guidance for starting values can also be obtained from estimation results reported by other researchers.

- The numerical optimization algorithm may stall in a local optimum. One way of guarding against stopping at a local optimum is to start from different starting values, and to check whether or not the same optimum is achieved regardless from where the algorithm starts.
M-estimators are defined as global optimizers of the objective function that defines the estimator. Statistically it is irrelevant how the global optimum is found.
A APPENDIX: FURTHER DISCUSSION OF EXAMPLES OF LMD AND GMM ESTIMATION

In the following we give a further discussion of examples of ML and GMM estimators. The examples are illustrative and are not intended to provide a comprehensive discussion of some of the underlying modelling approaches. Such a discussion will be given later in the econometrics sequence.

A.1 LMD Estimation

As in the main body of the handout consider the i.i.d. data process \( z_i = [y_i, x_i] \) with \( y_i \in Y \subseteq \mathbb{R}^p \) and \( x_i \in X \subseteq \mathbb{R}^{p_x} \), where \( Y \) and \( X \) are Borel sets. Let \( F(y \mid x) \) denote the conditional distribution of \( y_i \) given \( x_i = x \). We assume that this distribution can be described by a conditional density,

\[
f(y \mid x; \beta_0),
\]

with respect to some (\( \sigma \)-finite) measure \( \mu \) defined on the \( Y \), and \( \beta_0 \in B \subseteq \mathbb{R}^{p_\beta} \) denotes the true parameter vector. The objective function of the conditional maximum likelihood estimator is then given by

\[
Q_n(z_1, \ldots, z_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta),
\]

where

\[
q(z_i, \beta) = -\ln \left[ f(y_i \mid x_i; \beta) \right].
\]

now represents the negative conditional log-likelihood for observation \( i \).

A.1.1 ML Estimation of Linear and Nonlinear Regression Models

Suppose the scalar dependent variable \( y_i \) is generated according to the following linear regression model

\[
y_i = x_i \alpha_0 + \varepsilon_i, \quad i = 1, \ldots, n,
\]

where \([x_i, \varepsilon_i]\) is i.i.d., \( \varepsilon_i \) is independent of \( x_i \) and distributed \( N(0, \sigma_0^2) \). Then \([y_i, x_i]\) is i.i.d., \( y_i \) conditional on \( x_i \) is distributed \( N(x_i, \alpha_0, \sigma_0^2) \), and the negative conditional log-likelihood function for observation \( i \) is given by

\[
q(z_i, \beta) = -\ln \left[ f(y_i \mid x_i; \beta) \right]
= \frac{1}{2} \ln(2\pi) + \frac{1}{2} \ln(\sigma^2) + \frac{1}{2\sigma^2} (y_i - x_i \alpha)^2
\]
where \( z_i = [y_i, x_i], \beta = [\alpha', \sigma^2]' \) and \( \beta_0 = [\alpha_0', \sigma_0^2]' \). As demonstrated in the first part of the course, the ML estimator coincides with the OLS estimator and is given by

\[
\hat{\beta}_n = [\hat{\alpha}_n, \hat{\sigma}_n^2]'
\]

with

\[
\hat{\alpha}_n = \left( \sum_{i=1}^{n} x_i' x_i \right)^{-1} \left( \sum_{i=1}^{n} x_i' y_i \right) = (X'X)^{-1}X'y,
\]

\[
\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i \hat{\alpha}_n)^2 = \frac{1}{n} \hat{\varepsilon}' \hat{\varepsilon},
\]

where \( X = [x_1', \ldots, x_n']', y = [y_1, \ldots, y_n] \), and \( \hat{\varepsilon} = y - X \hat{\alpha}_n \).

In case \( y_i \) is generated by the nonlinear regression model

\[
y_i = g(x_i, \alpha_0) + \varepsilon_i, \quad i = 1, \ldots, n,
\]

where \( g \) is known, we would replace \( x_i \alpha \) by \( g(x_i, \alpha) \) in the expression for \( q(z_i, \beta) \). In this case it is in general no longer possible to give an explicit expression for \( \hat{\alpha}_n \).

Now suppose we optimize the negative log-likelihood function w.r.t. \( \sigma^2 \) for given values of \( \alpha \), then it is readily seen the maximizing value is given by

\[
\hat{\sigma}^2(\alpha) = \frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i, \alpha))^2.
\]

Substitution of this expression back into the negative log-likelihood function yields (ignoring constants)

\[
\ln \left[ \frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i, \alpha))^2 \right].
\]

This function is typically referred to as the negative concentrated log-likelihood function. Of course, minimizing this function w.r.t. \( \alpha \) is equivalent with minimizing

\[
\frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i, \alpha))^2,
\]

which shows that for the setting under consideration the ML estimator and the NLS estimator for \( \alpha \) are identical.

**A.1.2 ML Estimation of Discrete Response Models**

Consider the latent variable model \((i = 1, \ldots, n)\)

\[
y^*_i = x_i \beta_0 + \varepsilon_i.
\]
Assume that $[x_i, \varepsilon_i]$ is i.i.d. and let $G: \mathbb{R} \rightarrow \mathbb{R}$ be a cumulative distribution function (c.d.f.) of $\varepsilon_i$ given $x_i$. Assume furthermore that the distribution of $\varepsilon_i$ given $x_i$ is symmetric around zero. The latent variable $y^*_i$ is unobserved. Instead we observe the binary variable $y_i$ defined as

$$y_i = \begin{cases} 1 & \text{if } y^*_i > 0 \\ 0 & \text{if } y^*_i \leq 0 \end{cases}.$$ 

Then

$$P(y_i = 1 \mid x_i) = P(y^*_i > 0 \mid x_i) = P(x_i\beta_0 + \varepsilon_i > 0 \mid x_i)$$

$$= P(\varepsilon_i > -x_i\beta_0 \mid x_i) = 1 - G(-x_i\beta_0)$$

$$P(y_i = 0 \mid x_i) = 1 - G(x_i\beta_0).$$

The p.d.f. of $y_i$ given $x_i$ is thus given by

$$f(y_i \mid x_i; \beta_0) = \begin{cases} G(x_i\beta_0)^y[1 - G(x_i\beta_0)]^{1-y} & \text{for } y = 0, 1 \\ 0 & \text{else} \end{cases}.$$ 

The negative log-likelihood corresponding to the $i$-th observation conditional on $x_i$ is thus given by

$$q(z_i, \beta) = -\ln f(y_i \mid x_i; \beta) = -y_i \ln [G(x_i\beta)] - (1 - y_i) \ln [1 - G(x_i\beta)].$$

with $z_i = [y_i, x_i]$.

The probit model is a special case of the above model with $G$ the c.d.f. of the standardized normal distribution. The logit model is a special case with

$$G(v) = \frac{\exp(v)}{1 + \exp(v)}.$$ 

### A.1.3 ML Estimation of Censored Regression Models

Consider the standard censored Tobit model (or type I Tobit model) where the random variable $y^*_i$ is generated as follows ($i = 1, \ldots, n$):

$$y^*_i = \begin{cases} x_i \alpha_0 + \varepsilon_i & \text{if } y^*_i > 0 \\ 0 & \text{if } y^*_i \leq 0 \end{cases},$$

where $[x_i, \varepsilon_i]$ is i.i.d., $\varepsilon_i$ is independent of $x_i$ and distributed $N(0, \sigma_0^2)$. It is assumed that $z_i = [y_i, x_i]$ is observed, but that $y^*_i$ is unobserved if $y^*_i \leq 0$. The parameters of interest are $\beta_0 = [\alpha_0, \sigma_0^2]'$. As will be explained in more detail later in the econometrics sequence, $E(y_i \mid x_i) \neq x_i\alpha_0$ and thus an OLS
regression of $y_i$ on $x_i$ yields biased and inconsistent estimates for $\alpha_0$. One approach to estimate the model is by ML. It can be shown, see, e.g., Amemiya (1985), p. 363, or Wooldridge (2002), p. 526, that the negative conditional log-likelihood function is given by

$$q(z_i, \beta) = -1[y_i = 0] \ln[1 - \Phi(x_i \alpha/\sigma)] - 1[y_i > 0] \{\ln \phi[(y_i - x_i \alpha)/\sigma] - \ln(\sigma^2)/2\}$$

with $\beta = [\alpha', \sigma^2]'$, and $1[.]$ denotes the indicator function, and $\phi$ and $\Phi$ denote, respectively, the p.d.f. and c.d.f. of the standardized normal distribution.

### A.1.4 ML Estimation of a Count Data Model

The discussion in this section is based to a large extent on Winkelmann (2003), ch. 7.3. As Winkelmann states, a natural application of count data modeling arises when one is interested in finding out what determines the number of trips taken by a person (or household) over a specific time period. Such data occur, for example, in empirical studies in the field of environmental economics or regional economics. In the former, trip frequency can be used to estimate the value of a recreational site. In regional economics, and urban planning in particular, one is interested in the number of trips to a particular shopping cite, and how the number is affected by distance, characteristics of the shopping cite and the location and attributes of alternative sites in the region.

Winkelmann list various studies of the demand for recreational trips. The prevailing approach is the so-called “travel cost method”. The goal of these models is to estimate a conventional downward sloping demand function. The “quantity demanded” is the number of trips taken to the site during a given period of time, and the “price” is the travel cost of reaching the site. Price variation derives from the fact that individuals live at different distances from the site. Those living nearby have lower cost and would be expected to undertake more trips. Formally, let $y_i$ be the number of trips to a single site by individual (household) $i$. It is then often assumed that $y_i$, conditional on some co-variates $x_i$, follows a Poisson distribution. The conditional density is completely determined by the conditional mean $\mu(x_i) = E(y_i \mid x_i)$ and the conditional Poisson density is given by

$$f(y \mid x) = \frac{e^{-\mu(x)} \mu(x)^y}{y!}, \quad y = 0, 1, 2, \ldots$$

The conditional mean is typically specified as

$$\mu(x_i) = \exp(x_i \beta_0).$$

The explanatory variables will consist of variables measuring cost as well as various socioeconomic characteristics, including income. The model can be estimated by conditional ML. The negative conditional log-likelihood function is
given by
\[ q(z_i, \beta) = -\ln f(y_i | x_i; \beta) = -y_i[\ln \mu(x_i)] + \mu(x_i) + \ln(y_i!) \]
\[ = -y_i[\mu(x_i; \beta) + \exp(x_i; \beta) + \ln(y_i!)] \]

with \( z_i = [y_i, x_i] \). The term \( \ln(y_i!) \) does not depend on \( \beta \) and can be dropped.

For further discussions on count data model see, e.g., also Cameron and Trivedi (1998) and Wooldridge (2002), ch. 19.


**A.1.5 NLS Estimation of a Linear Regression Model with Heteroskedastic Disturbances**

Suppose the scalar dependent variable \( y_i \) is generated according to the following two variable linear regression model
\[ y_i = a_0 + b_0 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \]
where \([x_i, \varepsilon_i]\) is i.i.d. and
\[ E[\varepsilon_i | x_i] = 0, \]
\[ E[\varepsilon_i^2 | x_i] = (c_0 + d_0 x_i)^2 = c_0^2(1 + \delta_0 x_i)^2 \]
with \( \delta_0 = d_0/c_0 \). That is, the disturbances \( \varepsilon_i \) are conditionally heteroskedastic.

Consider the transformed model
\[ \frac{y_i}{1 + \delta_0 x_i} = a_0 \frac{1}{1 + \delta_0 x_i} + b_0 \frac{x_i}{1 + \delta_0 x_i} + \frac{\varepsilon_i}{1 + \delta_0 x_i}, \quad i = 1, \ldots, n. \]

By construction, the transformed disturbances \( \varepsilon_i / (1 + \delta_0 x_i) \) have mean zero and are homoskedastic. This suggests that one way of estimating the model is by nonlinear least squares. The objective function of the nonlinear least squares estimator corresponding to the transformed model is then readily seen to be a least mean distance estimator with
\[ q(z_i, \beta) = \frac{(y_i - a - bx_i)^2}{(1 + \delta x_i)^2} \]
where \( z_i = [y_i, x_i] \), \( \beta_0 = [a_0, b_0, \delta_0] \), and \( \beta = [a, b, \delta] \).
A.2 GMM Estimation

As in the main body of this handout in general GMM estimators $\hat{\beta}_n$ are defined as to correspond to objective functions of the form

$$ R_n(\omega, \beta) = Q_n(z_1, \ldots, z_n, \hat{\tau}_n, \beta) = n^{-1} \sum_{i=1}^{n} q(z_i, \beta) ' \hat{\Xi}_n n^{-1} \sum_{i=1}^{n} q(z_i, \beta) $$

where $q(z, \beta)$ is a real vector valued function defined on $Z \times B$ taking its values in $\mathbb{R}^{p_q}, \hat{\Xi}_n$ is a $p_q \times p_q$ positive semi-definite symmetric matrix and $\hat{\tau}_n$ is the vector of diagonal and upper diagonal elements of $\hat{\Xi}_n$. Here $q(.)$ corresponds to a set of moment conditions

$$ E q(z_i, \beta_0) = 0 $$

where $\beta_0$ is the true parameter vector.

In case we have as many moment conditions as unknown parameters it will typically be possible to select $\hat{\beta}_n$ such that

$$ n^{-1} \sum_{i=1}^{n} q(z_i, \hat{\beta}_n) = 0, $$

and $\hat{\beta}_n$ can then be viewed as a Method of Moments estimator.

A.2.1 GMM Estimation of Linear and Nonlinear Regression Models

Suppose the data $z_i = [y_i, x_i]$ are generated according to the linear regression model specified in the corresponding subsection on ML estimators in this appendix. Under those assumptions we have the following $K + 1$ moment conditions

$$ E q(z_i, \beta_0) = 0 $$

with $\beta_0 = [\alpha_0^t, \sigma_0^2]^t$. In this case we have as many moment conditions as unknown parameters and the GMM estimator $\hat{\beta}_n = [\hat{\alpha}_n^t, \hat{\sigma}_n^2]^t$ becomes the Method of Moments estimator that solves

$$ n^{-1} \sum_{i=1}^{n} q(z_i, \hat{\beta}_n) = \begin{bmatrix} n^{-1} \sum_{i=1}^{n} \hat{x}_i (y_i - x_i \hat{\alpha}_n) \\ n^{-1} \sum_{i=1}^{n} \hat{x}_i (y_i - x_i \hat{\alpha}_n)^2 - \hat{\sigma}_n^2 \end{bmatrix} = 0. $$

Of course, this yields again the OLS estimator.

In case $y_i$ is generated by the nonlinear regression model

$$ y_i = g(x_i, \alpha_0) + \varepsilon_i, \quad i = 1, \ldots, n, $$

we can still utilize the above moment conditions. However in this case the number of moment conditions will generally differ from the number of unknown
parameters. The objective function of the GMM estimator for $\beta_0$ with $\hat{\Xi}_n = (\sum_{i=1}^n x_i x_i')^{-1}$ is given by

$$n^{-1} \sum_{i=1}^n q(z_i, \beta)' \hat{\Xi}_n [n^{-1} \sum_{i=1}^n q(z_i, \beta)] = n^{-1} \sum_{i=1}^n x_i' (y_i - g(x_i, \alpha))' [n^{-1} \sum_{i=1}^n x_i x_i]^{-1} [n^{-1} \sum_{i=1}^n x_i' (y_i - g(x_i, \alpha))]$$

where $X = [x_1', \ldots, x_n']'$, $y = [y_1, \ldots, y_n]$, and

$$\epsilon(\alpha) = \begin{bmatrix} y_1 - g(x_1, \alpha) \\ \vdots \\ y_n - g(x_n, \alpha) \end{bmatrix}.$$ 

The choice of $\hat{\Xi}_n = (n^{-1} \sum_{i=1}^n x_i x_i)^{-1}$ is motivated by $VC(n^{-1} \sum_{i=1}^n x_i' \epsilon_i) = \sigma^2 EX|x_i$. Observe that the objective function of the GMM estimator for $\alpha_0$ differs from that of the NLS estimator, which is given by

$$n^{-1} \epsilon(\alpha)' \epsilon(\alpha) = \frac{1}{n} \sum_{i=1}^n (y_i - g(x_i, \alpha))^2.$$ 

### A.2.2 GMM Estimation of Simultaneous Linear and Nonlinear Regression Models

Many economic variables are determined by systems of equations. A typical example would be a demand and supply equation which jointly determines the equilibrium price and quantity. For an example regarding the supply and demand of labor and specific examples see, e.g., Wooldridge (2002), ch. 8.1.

We start our discussion by considering the following system of $G$ linear simultaneous equations

$$y_g = Y_g \beta_g^0 + X_g \gamma_g^0 + \varepsilon_g$$

$$W_g = Y_g X_g^T$$

where $Y_g$ is the matrix of observations containing the $G_g$ endogenous r.h.s. variables and $X_g$ is the matrix of observations containing the $K_g$ exogenous
variables in the $g$-th equation. Let

\[
\mathbf{Y} = (\mathbf{y}_g) = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = [y_{1g}, \ldots, y_{Gg}],
\]

\[
\mathbf{X} = (\mathbf{x}_{ik}) = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = [x_{1k}, \ldots, x_{Kg}],
\]

\[
\mathbf{E} = (\mathbf{\varepsilon}_{ig}) = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} = [\varepsilon_{1g}, \ldots, \varepsilon_{Gg}],
\]

denote, respectively, the $n \times G$, $n \times K$ and $n \times G$ matrices of all endogenous variables, exogenous variables, and disturbances in the system. Also let $\beta_0^g = [\beta_{1g}^0, \ldots, \beta_{Gg}^0]'$ and $\gamma_0^g = [\gamma_{1g}^0, \ldots, \gamma_{Kg}^0]'$ denote the true parameter vectors in the $g$-th equation.

Without loss of generality we assume for the remainder of this section that the equation of interest is the first equation and that the variables are labeled such that

\[
\mathbf{Y} = [y_{11}, \mathbf{Y}_1^*],
\]

\[
\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_1^*].
\]

The first equation can then be written in scalar notation as

\[
y_{i1} = \sum_{g=2}^{G_1} y_{ig} \beta_{1g}^0 + \sum_{k=1}^{K_1} x_{ik} \gamma_{1k}^0 + \varepsilon_{i1}. \quad (*)
\]

Assume that the processes $(x_i, \varepsilon_i)$ and hence the process $(y_i, x_i)$ is i.i.d. Assume further that $x_i$ and $\varepsilon_i$ are independent and that $E \varepsilon_i = 0$ and $E \varepsilon_i' \varepsilon_i = \Sigma = (\sigma_{gl})$. Under those assumptions clearly $E \mathbf{X}_i' \varepsilon_{ig} = 0$ for $g = 1, \ldots, G$, and in particular we have the following moment conditions involving the parameters of the first equation:

\[
E q(\mathbf{z}_i, \beta_0) = E \mathbf{X}_i' \varepsilon_{i1} = E \begin{bmatrix} \mathbf{x}_{i1} \\ \vdots \\ \mathbf{x}_{iK} \end{bmatrix} \begin{bmatrix} y_{i1} - \sum_{g=2}^{G_1} y_{ig} \beta_{1g}^0 - \sum_{k=1}^{K_1} x_{ik} \gamma_{1k}^0 \end{bmatrix} = 0
\]

where $\mathbf{z}_i = [y_{i1}, y_{i2}, \ldots, y_{iG_1}, x_{i1}, \ldots, x_{iK}]$ and $\beta_0 = \delta_1^0$. Towards defining the corresponding GMM estimator it proves helpful to rewrite the average of the moment conditions as

\[
n^{-1} \sum_{i=1}^{n} E q(\mathbf{z}_i, \beta_0) = n^{-1} E \mathbf{X}_1' \varepsilon_{11} = n^{-1} E \mathbf{X}_1' [\mathbf{y}_{11} - \mathbf{W}_1 \delta_1^0] = 0.
\]
The objective function of the GMM estimator for $\beta_0 = \delta_1^o$ with $\hat{\Xi}_n = \sum_{i=1}^n x'_i x_i^{-1} = (X'X)^{-1}$ is now given by

$$n^{-1} \sum_{i=1}^n q(z_i, \beta)' \hat{\Xi}_n \left[ n^{-1} \sum_{i=1}^n q(z_i, \beta) \right]$$

$$= [y_{1,1} - \hat{W}_1 \delta_1^o]'X(X'X)^{-1}X[y_{1,1} - \hat{W}_1 \delta_1^o].$$

The choice of $\hat{\Xi}_n = \sum_{i=1}^n x'_i x_i^{-1}$ is motivated by $VC(n^{-1} \sum_{i=1}^n x'_i \xi_1) = \sigma_1^o E x'_i x_i$. It is readily seen that this objective function is minimized at

$$\hat{\delta}_{1,n} = [\hat{W}_1' \hat{W}_1]^{-1} \hat{W}_1' y_{1,1},$$

$$\hat{W}_1 = X(X'X)^{-1}X' \hat{W}_1.$$ 

The estimator $\hat{\delta}_{1,n}$ is known as the 2SLS estimator. An in depth discussion of this estimator will be given later in the econometrics sequence.

The estimation approach can be naturally extended to nonlinear simultaneous equation models. Suppose the first equation is of the following explicit nonlinear form:

$$y_{1,1} = g(W_1, \delta_1^o) + \varepsilon_1.$$ 

Then the nonlinear 2SLS estimator is defined as the minimizer of the objective function.

$$[y_{1,1} - g(W_1, \delta_1)]'X(X'X)^{-1}X[y_{1,1} - g(W_1, \delta_1)].$$ 

If the first equation is of the following implicit nonlinear form:

$$f(y_{1,1}, W_1, \delta_1^o) = \varepsilon_1,$$

then the nonlinear 2SLS estimator is defined as the minimizer of the objective function.

$$f(y_{1,1}, W_1, \delta_1)'X(X'X)^{-1}X f(y_{1,1}, W_1, \delta_1).$$

### A.2.3 Education Example

The following example is an adaptation of Angrist and Krueger (1992) and the discussion of this model in Hall (2001). Angrist and Krueger (1992) investigate the impact of age at school entry on education attainment using the model

$$y_i = a_0 + b_0 a_i + \varepsilon_i,$$

where $y_i$ is the number of years of education completed and $a_i$ is the age of school entry by by student $i$. Within this model, the marginal response of attainment to age of entry is captured by $\beta$, and so this represents the parameter of interest. Estimation of this parameter is complicated by a correlation between the explanatory variable and the error which arises because many children who
start school at a younger age do so because they show above average learning potential. This correlation means the ordinary least squares estimator is inconsistent. However, the error is anticipated to be uncorrelated with the quarter of birth. Let \( d_i = [d_{i1}, d_{i2}, d_{i3}, d_{i4}] \) be a vector of dummy variables with \( d_{ji} = 1 \) if student \( i \) was born in quarter \( j = 1, \ldots, 4 \) and \( d_{ji} = 0 \) otherwise. Then this logic leads to the population moment condition

\[
E(q(z_i, \beta_0)) = E(d_i (y_i - a_0 - b_0 a_i)) = 0
\]

where \( z_i = [y_i, a_i, d_i] \) and \( \beta_0 = [a_0, b_0]^t \).


References


[127] Kabaila, P. Parameter values of ARMA models minimizing the one step ahead prediction error when the true system is not in the model set. *Journal of Applied Probability* 20 (1983): 405-408.


