

SERIE RESEARCH MEMORANDA

NONLINEAR PARAMETRIC REGRESSION ANALYSIS

Herman J. Bierens

Researchmemorandum 1987-21

April '87



VRIJE UNIVERSITEIT
Faculteit der Economische Wetenschappen en Econometrie
A M S T E R D A M

NONLINEAR PARAMETRIC REGRESSION ANALYSIS *

Herman J. Bierens
Department of Econometrics
Free University, Amsterdam

December 1988

* Chapter 4 of:

PRINCIPLES OF NONLINEAR AND NONPARAMETRIC REGRESSION ANALYSIS

To be published by Marcel Dekker Inc., New York.



TABLE OF CONTENTS:

PREFACE

PART 1: PRELIMINARY MATHEMATICS

1. BASIC PROBABILITY THEORY
 - 1.1 Measure-theoretical foundation of probability theory
 - 1.2 Independence
 - 1.3 Borel measurable functions
 - 1.4 Mathematical expectation
 - 1.5 Characteristic functions
 - 1.6 Random functions
2. CONVERGENCE
 - 2.1 Weak and strong convergence of random variables
 - 2.2 Convergence of mathematical expectations
 - 2.3 Convergence of distributions
 - 2.4 Central limit theorems
 - 2.5 Further results on convergence of distributions and mathematical expectations, and laws of large numbers
 - 2.6 Convergence of random functions
 - 2.7 Uniform strong and weak laws of large numbers
3. INTRODUCTION TO CONDITIONING
 - 3.1 Definition of conditional expectation
 - 3.2. Basic properties of conditional expectations
 - 3.3 Identification of conditional expectations

PART 2: REGRESSION ANALYSIS UNDER INDEPENDENCE

4. NONLINEAR PARAMETRIC REGRESSION ANALYSIS
 - 4.1 Nonlinear regression models and the nonlinear least squares estimator
 - 4.2 Consistency and asymptotic normality: General theory
 - 4.2.1 Consistency
 - 4.2.2 Asymptotic normality
 - 4.3 Consistency and asymptotic normality of nonlinear least squares estimators in the i.i.d. case
 - 4.3.1 Consistency

- 4.3.2 Asymptotic normality
 - 4.3.3 Consistent estimation of the asymptotic variance matrix
 - 4.4 Consistency and asymptotic normality of the nonlinear least squares estimator under data heterogeneity
 - 4.4.1 Data heterogeneity
 - 4.4.2 Strong and weak consistency
 - 4.4.3 Asymptotic normality
 - 4.5 Testing parameter restrictions: The Wald test
5. TESTS FOR MODEL MISSPECIFICATION
 - 5.1 White's version of Hausman's test
 - 5.2 Newey's M-test
 - 5.2.1 Introduction
 - 5.2.2 The conditional M-test
 - 5.3 A consistent conditional M-test
 - 5.4 The integrated M-test
6. THE NADARAYA-WATSON KERNEL REGRESSION FUNCTION ESTIMATOR
 - 6.1 Introduction
 - 6.2 Asymptotic normality in the continuous case
 - 6.3 Uniform consistency in the continuous case
 - 6.4 Discrete and mixed continuous-discrete regressors
 - 6.4.1 The discrete case
 - 6.4.2 The mixed continuous-discrete case
 - 6.5 The choice of the kernel
 - 6.6 The choice of the window width
 - 6.7 An empirical application to specification of household expenditures systems and equivalence scales
 - 6.7.1 Introduction
 - 6.7.2 Model and data
 - 6.7.3 The results
 - 6.7.4 Sample selection
7. SAMPLE MOMENTS INTEGRATING NORMAL KERNEL ESTIMATORS
 - 7.1 Kernel density estimators
 - 7.1.1 Integral conditions
 - 7.1.2 Sample moments integrating kernel density estimators
 - 7.1.3 The Dirac-catastrophe and the modified SMINK density estimator

- 7.1.4 How to choose the window width of the modified SMINK density estimator
- 7.2 Regression
 - 7.2.1 SMINK estimators of a regression function
 - 7.2.2 How to choose the window width parameters
- 7.3 A numerical example
- 7.4 Proofs
- 8. NONLINEAR REGRESSION WITH DISCRETE EXPLANATORY VARIABLES
 - 8.1. The earnings function
 - 8.2. The functional form of a regression model with discrete explanatory variables
 - 8.3. The choice of the linear separator
 - 8.4. Estimating and testing the regression function
 - 8.4.1 Estimation
 - 8.4.2 Model specification testing
 - 8.4.3 The selection of the polynomial order
 - 8.5 Proofs
 - 8.5.1 Proof of theorem 8.4.2
 - 8.5.2 Proof of theorem 8.4.3

PART 3: TIME SERIES

- 9. CONDITIONING AND DEPENDENCE
 - 9.1 Conditional expectations relative to a Borel field
 - 9.1.1 Definition and basic properties
 - 9.1.2 Martingales
 - 9.1.3 Martingale convergence theorems
 - 9.1.4 A martingale difference central limit theorem
 - 9.2 Measures of dependence
 - 9.2.1 Mixingales
 - 9.2.2 Uniform and strong mixing
 - 9.2.3 ν -Stability
 - 9.3 Weak laws of large numbers for dependent random variables
 - 9.4 Proper heterogeneity and uniform laws for functions of infinitely many random variables
- 10. FUNCTIONAL SPECIFICATION OF TIME SERIES MODELS
 - 10.1 Linear times series regression models
 - 10.1.1 Introduction

- 10.1.2 The Wold decomposition
- 10.1.3 Linear vector time series models
- 10.2 ARMA memory index models
 - 10.2.1 Introduction
 - 10.2.2 Finite conditioning of univariate rational-valued time series
 - 10.2.3 Infinite conditioning of univariate rational-valued time series
 - 10.2.3 The multivariate case
 - 10.2.4 The nature of the ARMA memory index parameters and the response functions
 - 10.2.5 Discussion
- 10.3 Nonlinear ARMAX models
- 11. ARMAX MODELS: ESTIMATION AND TESTING
 - 11.1 Estimation of linear ARMAX models
 - 11.1.1 Introduction
 - 11.1.2 Consistency and asymptotic normality
 - 11.2 Estimation of nonlinear ARMAX models
 - 11.3 A consistent $N(0,1)$ model specification test
 - 11.4 A consistent Hausman-type model specification test
 - 11.5 An autocorrelation test
- 12. NONPARAMETRIC TIME SERIES REGRESSION
 - 12.1 Assumptions and preliminary lemmas
 - 12.2 Consistency and asymptotic normality of time series kernel regression function estimators

PREFACE

This book deals with statistical inference of nonlinear regression models from two opposite points of view, namely the case where the functional form of the model is completely specified as a known function of regressors and unknown parameters, and the opposite case where the functional form of the model is completely unknown. First it is assumed that the response function of the regression model under review belongs to a certain well-specified parametric family of functional forms, by which estimation of the model merely amounts to estimation of the unknown parameters. For this class of models we review the asymptotic properties of the nonlinear least squares estimator for independent data as well as for time series.

In practice assumptions on the functional form are often made on the basis of computational convenience rather than on the basis of precise a priori knowledge of the empirical phenomenon under review. Therefore the linear regression model is still the most popular model specification in applied research. However, even if the specification of the functional form is based on sound theoretical considerations there is quite often a large range of functional forms that are theoretically admissible, so that there is no guarantee that the actually chosen functional form is true. Functional specification of a parametric nonlinear regression model should therefore always be verified by conducting model misspecification tests. Various model misspecification tests will therefore be discussed, in particular consistent tests which have asymptotic power 1 against all deviations from the null hypothesis that the model is correct.

The opposite case of parametric regression is nonparametric regression. Nonparametric regression analysis is concerned with estimation of a regression model without specifying in advance its functional form. Thus the only source of information about the functional form of the model is the data set itself. In this book we shall review various nonparametric regression approaches, with special emphasis on the kernel method, under various distributional assumptions.

This book is divided into three parts. In the first part we review the elements of abstract probability theory we need in part 2. Part 2 is devoted to the asymptotic theory of para-

metric and nonparametric regression analysis in the case of independent data generating processes. In part 3 we extend the analysis involved to time series.

The selection of the topics mainly reflexes my own interest in the subject. Instead of providing an encyclopedic survey of the literature, I have chosen for a setup which aims to fill the gap between intermediate statistics (including linear time series analysis) and the level necessary to get access to the recent literature on nonlinear and nonparametric regression analysis, with emphasis on my own contributions. The ultimate goal is to provide the student with the tools for his own independent research in this area, by showing what tools I and others have used and what they have been used for. Thus, this book may be viewed as an account of my own struggle with the material involved. I think this book is particularly suitable for self-tuition (at least it aims to be), and may prove useful in a graduate course in mathematical statistics and advanced econometrics.

Acknowledgements:

The first five chapters of this book have been disseminated in draft form as working papers. I am grateful to Anil Bera, Alexander Georgiev and Jan Magnus for suggesting additional references, and in particular to Lourens Broersma, Johan Smits and Ton Steerneman who suggested various improvements.

A large body of the material in chapter 6 has been published earlier in Truman F. Bewley (ed.), *Advances in Econometrics, Fifth World Congress*, Cambridge University Press. I am indebted to Cambridge University Press for granting permission to reprint it.

4. NONLINEAR PARAMETRIC REGRESSION ANALYSIS

In this chapter we consider the asymptotic properties, i.e. weak and strong consistency and asymptotic normality, of the least squares estimators of the parameters of a nonlinear regression model. Throughout we assume that the data generating process is independent, but we distinguish between the identically distributed and the nonidentically distributed case. Also, we consider the Wald test of nonlinear restrictions on the parameters.

Notable work in this area has been done by Jennrich (1969), Malinvaud (1970a,b), White (1980b,1982) and Bierens (1981), among others. The present approach is a further elaboration of the approach in Bierens (1981, section 3.1).

4.1 Nonlinear regression models and the nonlinear least squares estimator

Consider an independent data generating process $\{(Y_j, X_j)\}$, $j=1,2,\dots$ where $Y_j \in \mathbb{R}$ is the dependent variable and $X_j \in \mathbb{R}^k$ is a vector of regressors. The central problem in regression analysis is how to determine the conditional expectation functions

$$g_j(X_j) = E(Y_j | X_j), \quad j=1,2,\dots \quad (4.1.1)$$

We shall call this function g_j the *response function*. Note that this response function exists and is a Borel measurable real function on $\mathbb{R}^k$, provided

$$E|Y_j| < \infty \quad (4.1.2)$$

(cf. section 3.1). Defining

$$U_j = Y_j - g_j(X_j), \quad j=1,2,\dots \quad (4.1.3)$$

we get the tautological regression model

$$Y_j = g_j(X_j) + U_j, \quad j=1,2,\dots \quad (4.1.4)$$

where by construction the errors U_j satisfy

$$E(U_j | X_j) = 0 \text{ a.s., } j=1,2,\dots \quad (4.1.5)$$

In parametric regression analysis it is assumed that:

(I) g_j does not depend on j :

$$g_j = g \text{ for } j=1,2,\dots; \quad (4.1.6)$$

(II) the function g belongs to a parametric family of functions $f(.,\theta)$;

$$g(.) \in \{f(.,\theta) : \theta \in \Theta\}, \quad (4.1.7)$$

where Θ is a subset of a Euclidean space, called the *parameter space*, and $f(x,\theta)$ is known for each $x \in R^k$ and each $\theta \in \Theta$. Quite often one also assumes that the errors U_j are independent of X_j and even that the U_j 's are normally distributed. These assumptions, however, are not necessary for the asymptotic theory of regression analysis and therefore we shall not depend on them.

Condition (I) naturally holds if the data generating process is i.i.d. However, if some or all of the components of X_j are control variables set by the analyst in a statistical experiment, we can no longer assume that the X_j 's are identically distributed. In that case the response function may depend on the observation index j .

The condition that the response function belongs to a certain parametric family of known functions is crucial to parametric regression analysis, but is also disputable. In practice we almost never know the exact functional form of a regression model. The choice of the parametric family (4.1.7) is therefore almost always a matter of convenience rather than a matter of precise a priori knowledge. For that reason the linear regression model is still the most popular functional specification in empirical research, due to its convenience in estimating the parameters by ordinary least squares (OLS). So why should we bother about nonlinear regression models?

There are various reasons why nonlinear regression analysis makes sense. First, it is possible to test whether a given (linear) functional specification is correct (see chapter 5). If we have to reject the linear specification we are then forced to look for a nonlinear specification that is closer to

the true response function. Second, sometimes a linear regression model is logically inconsistent with the nature of the data. If for example the range of Y_j is $[0, \infty)$ and the range of the components of X_j is $(-\infty, \infty)$, then the linear response function may take negative values, whereas $E(Y_j | X_j) \geq 0$ a.s. (cf. theorem 3.2.1.V). Another example is the case where Y_j is a binary variable only taking the values 0 and 1. Then $E(Y_j | X_j)$ is just the conditional probability of the event $Y_j = 1$ relative to X_j , and must therefore lie in the interval $[0, 1]$. A possible specification in this case is a logit-type model:

$$Y_j = 1/[1 + \exp(-\theta_0 - \theta_1' X_j)] + U_j = f(X_j, \theta) + U_j,$$

$$\text{with } E(U_j | X_j) = 0 \text{ a.s.}$$

where $\theta = (\theta_0, \theta_1')' \in \mathbb{R} \times \mathbb{R}^k = \Theta$. Note that in this case U_j cannot be independent of X_j .

Third, there are cases where theory is more explicit about the functional form of a model. An example is the CES production function in economics, introduced by Arrow, Chenery, Minhas and Solow (1961). This production function takes the form

$$y = (\theta_1 x_1^{-\theta_3} + \theta_2 x_2^{-\theta_3})^{-1/\theta_3},$$

where $\theta_1 > 0$, $\theta_2 > 0$, $\theta_3 \geq -1$, y is the output level of an industry, x_1 is the total labor force employed in this industry and x_2 is the stock of capital in this industry. Adding an error term to the right hand side yields a nonlinear regression model.

Fourth, there are flexible functional specifications that allow the functional form to be partly determined by the data. A typical example of such a flexible functional specification is the Box-Cox (1964) transformation. The Box-Cox transformation of a variable v is given by

$$\eta(v|\lambda) = (v^\lambda - 1)/\lambda.$$

For $\lambda = 0$ we have $\eta(v|0) = \ln(v)$, and clearly $\eta(v|1) = v - 1$. This transformation is particularly suitable when the regressors are positively valued. Consider for example the case $X_j \in \mathbb{R}$, $P(X_j > 0) = 1$. Specifying the regression model as

$$Y_j = \theta_1 + \theta_2 \eta(X_j | \theta_3) + U_j$$

yields a model that contains the linear model and the semi-loglinear model ($Y_j = \theta_1 + \theta_2 \ln(X_j) + U_j$) as special cases. By estimating the parameters θ_1 , θ_2 and θ_3 and testing the hypothesis $\theta_3 = 0$ against $\theta_3 = 1$ we actually let the data choose between these alternatives.

From now on we take the parametric family of functional forms as given (though not necessarily as true). Moreover, we confine ourselves to functions that are Borel measurable in the explanatory variables and the parameters and are continuous in the parameters. Furthermore, the parameter space Θ is assumed to be a compact Borel set.

Assumption 4.1.1. Given the data generating process $\{(Y_j, X_j)\}$, $j=1,2,\dots$ in $\mathbb{R} \times \mathbb{R}^k$ with $E|Y_j| < \infty$, the conditional expectation of Y_j relative to X_j equals $f(X_j, \theta_0)$ a.s. for $j=1,2,\dots$, where $f(x, \theta)$ is a known Borel measurable real function on $\mathbb{R}^k \times \Theta$ with Θ a compact Borel set in $\mathbb{R}^m$ containing θ_0 . For each $x \in \mathbb{R}^k$ the function $f(x, \theta)$ is continuous on Θ .

We recall that compactness of a set implies that every open covering contains a finite subcovering, and that a subset of a Euclidean space is compact if and only if it is closed and bounded. Cf. Royden (1968). Assuming that the data generating process is observable for $j=1,2,\dots,n$, the nonlinear least squares estimator for θ_0 is now defined as a measurable solution of the following minimisation problem:

$$\hat{\theta} \in \Theta : \hat{Q}(\hat{\theta}) = \inf_{\theta \in \Theta} \hat{Q}(\theta), \quad (4.1.8)$$

where

$$\hat{Q}(\theta) = (1/n) \sum_{j=1}^n [Y_j - f(X_j, \theta)]^2. \quad (4.1.9)$$

By theorem 1.6.1 such a measurable solution always exists. In the sequel of this chapter we shall give further condition for strong consistency, i.e.

$$\hat{\theta} \rightarrow \theta_0 \quad \text{a.s.} \quad \text{as } n \rightarrow \infty \quad (4.1.10)$$

or weak consistency:

$$\text{plim}_{n \rightarrow \infty} \hat{\theta} = \theta_0, \quad (4.1.11)$$

and asymptotic normality, i.e.

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N_m(0, \Omega) \text{ in distr.} \quad (4.1.12)$$

4.2 Consistency and asymptotic normality: General theory

4.2.1 Consistency

The consistency proof contains two main steps. First we set forth conditions such that

$$\hat{Q}(\theta) \rightarrow Q(\theta) \text{ a.s. (in prob.) pseudo-uniformly on } \Theta \quad (4.2.1)$$

where

$$Q(\theta) = \lim_{n \rightarrow \infty} E \hat{Q}(\theta) \text{ is continuous on } \Theta \quad (4.2.2)$$

and then we set forth conditions such that

$$\theta_* \in \Theta : Q(\theta_*) = \inf_{\theta \in \Theta} Q(\theta) \Rightarrow \theta_* = \theta_0. \quad (4.2.3)$$

Thus the latter condition says that $Q(\theta)$ has a unique infimum on Θ at θ_0 . The conditions (4.2.2) and (4.2.3) guarantee that

$$\hat{\theta} \rightarrow \theta_0 \text{ a.s. [in prob.]},$$

due to the following fundamental theorem.

Theorem 4.2.1. Let $(Q_n(\theta))$ be a sequence of random functions on a compact set $\Theta \subset \mathbb{R}^m$ such that for a continuous real function $Q(\theta)$ on Θ ,

$$Q_n(\theta) \rightarrow Q(\theta) \text{ a.s. [in prob.] pseudo-uniformly on } \Theta.$$

Let θ_n be any random vector in Θ satisfying

$$Q_n(\theta_n) = \inf_{\theta \in \Theta} Q_n(\theta)$$

and let θ_0 be a unique point in Θ such that

$$Q(\theta_0) = \inf_{\theta \in \Theta} Q(\theta).$$

Then $\theta_n \rightarrow \theta_0$ a.s. [in prob.]

Proof: We consider first the a.s. case. Let (Ω, F, P) be the probability space and let N be the null set on which the uniform convergence fails to hold. For $\omega \in \Omega \setminus N$ we then have

$$\begin{aligned} 0 &\leq Q(\theta_n(\omega)) - Q(\theta_0) \\ &= Q(\theta_n(\omega)) - Q_n(\theta_n(\omega), \omega) + Q_n(\theta_n(\omega), \omega) - Q(\theta_0) \\ &\leq Q(\theta_n(\omega)) - Q_n(\theta_n(\omega), \omega) + Q_n(\theta_0, \omega) - Q(\theta_0) \\ &\leq 2 \sup_{\theta \in \Theta} |Q_n(\theta, \omega) - Q(\theta)| \rightarrow 0 \quad \text{as } n \rightarrow \infty, \end{aligned}$$

hence

$$Q(\theta_n(\omega)) \rightarrow Q(\theta_0) \quad \text{as } n \rightarrow \infty \quad \text{and } \omega \in \Omega \setminus N. \quad (4.2.4)$$

Now let $\theta_*(\omega)$ be any limit point of the sequence $(\theta_n(\omega))$. Since this sequence lies in a compact set Θ , all its limit points lie in Θ . Thus, $\theta_*(\omega) \in \Theta$. By definition of a limit point there exists a subsequence $(\theta_{n_j}(\omega))$ such that $\lim_{j \rightarrow \infty} \theta_{n_j}(\omega) = \theta^*(\omega)$. From (4.2.4) and the continuity of $Q(\theta)$ it follows that also

$$\lim_{j \rightarrow \infty} Q(\theta_{n_j}(\omega)) = Q(\lim_{j \rightarrow \infty} \theta_{n_j}(\omega)) = Q(\theta^*(\omega)) = Q(\theta_0)$$

hence $\theta^*(\omega) = \theta_0$ by uniqueness. Thus all the limit points of $\theta_n(\omega)$ are equal to θ_0 and consequently

$$\lim_{n \rightarrow \infty} \theta_n(\omega) = \theta_0 \quad \text{for } \omega \in \Omega \setminus N.$$

By definition, this implies $\theta_n \rightarrow \theta_0$ a.s. The convergence in probability case follows easily from theorem 2.1.6 and the above argument. Q.E.D.

4.2.2 Asymptotic normality

Next we turn to asymptotic normality. We shall put our argument in a general framework, in order to use the results also for more general cases than considered in the sequel of this chapter.

Theorem 4.2.2. Let the conditions of theorem 4.2.1 be satisfied (for the convergence in probability case) and assume in addition that:

- (I) Θ is convex and θ_0 is an interior point of Θ ;
- (II) $(\partial/\partial\theta')Q_n(\theta)$ and $(\partial/\partial\theta)(\partial/\partial\theta')Q_n(\theta)$ are well-defined as vector and matrix, respectively, of random functions on Θ ;
- (III) $\sqrt{n}(\partial/\partial\theta')Q_n(\theta_0) \rightarrow N_m[0, A_1]$ in distr., where A_1 is a positive semi-definite $m \times m$ matrix;
- (IV) for $i_1, i_2 = 1, 2, \dots, m$,

$$(\partial/\partial\theta_{i_1})(\partial/\partial\theta_{i_2})Q_n(\theta) \rightarrow (\partial/\partial\theta_{i_1})(\partial/\partial\theta_{i_2})Q(\theta) \text{ in prob.}$$

pseudo-uniformly on a neighborhood of θ_0 , where the limit function involved is continuous in θ_0 ;

- (V) the matrix $A_2 = (\partial/\partial\theta)(\partial/\partial\theta')Q(\theta_0)$ is nonsingular.

Then $\sqrt{n}(\theta_n - \theta_0) \rightarrow N_m[0, A_2^{-1}A_1A_2^{-1}]$.

Proof: Consider the following Taylor expansion of $(\partial/\partial\theta_i)Q_n(\theta_n)$ around θ_0 :

$$\begin{aligned} (\partial/\partial\theta_i)Q_n(\theta_n) &= (\partial/\partial\theta_i)Q_n(\theta_0) \\ &+ [(\partial/\partial\theta')(\partial/\partial\theta_i)Q_n(\theta_n^{(i)})]'(\theta_n - \theta_0) \\ &\quad (i = 1, 2, \dots, m) \end{aligned} \quad (4.2.5)$$

where $\theta_n^{(i)}$ is a mean value satisfying

$$|\theta_n^{(i)} - \theta_0| \leq |\theta_n - \theta_0| \quad \text{a.s.} \quad (4.2.6)$$

In the nonrandom case the existence of such a mean value is

provided by Taylor's theorem. In the random case under review, however, we should also ask whether a measurable mean value exists. For the moment we shall ignore this problem. At the end of the proof we shall return to this issue.

Note that we have indexed the mean values by i . This indicates that these mean values may be different for different i 's. Cf. Don (1986).

Now consider the left member of equation (4.2.5). If θ_n is an interior point of Θ then $(\partial/\partial\theta_i)Q_n(\theta_n) = 0$, due to the well-known first order conditions for a minimum of $Q_n(\theta)$ at θ_n . These first order conditions may not hold if θ_n lies on the boundary of Θ . However, since $\theta_n \rightarrow \theta_0$ in prob. and θ_0 is an interior point of Θ the probability that θ_n is an interior point of Θ will converge to 1, hence

$$\lim_{n \rightarrow \infty} P[(\partial/\partial\theta_i)Q_n(\theta_n) = 0] = 1. \quad (4.2.7)$$

Next consider the $m \times m$ random matrix

$$\hat{A}_2 = [(\partial/\partial\theta_{i_1})(\partial/\partial\theta_{i_2})Q_n(\theta_n^{(i_2)})] \quad (4.2.8)$$

Then it follows from (4.2.5) and (4.2.7) that

$$\lim_{n \rightarrow \infty} P[\sqrt{n}(\theta_n - \theta_0) = -\hat{A}_2^{-1} \sqrt{n}(\partial/\partial\theta')Q_n(\theta_0)] = 1,$$

hence

$$\begin{aligned} \text{plim}_{n \rightarrow \infty} (\sqrt{n}(\theta_n - \theta_0) + (\hat{A}_2^{-1} - A_2^{-1})\sqrt{n}(\partial/\partial\theta')Q_n(\theta_0) \\ + A_2^{-1}\sqrt{n}(\partial/\partial\theta')Q_n(\theta_0)) = 0. \end{aligned} \quad (4.2.9)$$

If

$$\text{plim}_{n \rightarrow \infty} (\hat{A}_2^{-1} - A_2^{-1}) = 0 \quad (4.2.10)$$

then it follows from condition (III) and theorems 2.3.3 and 2.3.5 that

$$\text{plim}_{n \rightarrow \infty} (\hat{A}_2^{-1} - A_2^{-1})\sqrt{n}(\partial/\partial\theta')Q_n(\theta_0) = 0. \quad (4.2.11)$$

Combining (4.2.9) and (4.2.11) we then conclude

$$\text{plim}_{n \rightarrow \infty} \{ \sqrt{n}(\theta_n - \theta_0) + A_2^{-1} \sqrt{n}(\partial/\partial \theta') Q_n(\theta_0) \} = 0. \quad (4.2.12)$$

Since condition (III) implies

$$-A_2^{-1} \sqrt{n}(\partial/\partial \theta') Q_n(\theta_0) \rightarrow N_m(0, A_2^{-1} A_1 A_2^{-1}) \quad (4.2.13)$$

(note that A_2 is symmetric), it follows from (4.2.12) and theorem 2.3.5 that

$$\sqrt{n}(\theta_n - \theta_0) \rightarrow N_m(0, A_2^{-1} A_1 A_2^{-1}) \text{ in distr.}$$

Thus for proving the theorem it remains to show that (4.2.10) holds and that the mean values are indeed properly defined random vectors. For proving (4.2.10), observe from (4.2.6) that $\text{plim}_{n \rightarrow \infty} \theta_n = \theta_0$ implies

$$\text{plim}_{n \rightarrow \infty} \theta_n^{(i)} = \theta_0, \quad i=1, 2, \dots, m. \quad (4.2.14)$$

From (4.2.14), condition (IV) and theorem 2.6.1 it follows now that

$$\text{plim}_{n \rightarrow \infty} \hat{A}_2 = A_2. \quad (4.2.15)$$

Finally, since the elements of an inverse matrix are continuous functions of the elements of the inverted matrix, provided the latter is nonsingular, it follows from (4.2.15), condition (V) and theorem 2.1.7 that

$$\text{plim}_{n \rightarrow \infty} \hat{A}_2^{-1} = A_2^{-1},$$

which proves (4.2.10).

We now conclude our proof with the following lemma of Jennrich:

Lemma 4.2.1. Let $f(x, \theta)$ be a real-valued function on $R^k \times \Theta$, where Θ is a convex compact subset of R^m . For each θ in Θ let $f(x, \theta)$ be a Borel measurable function on R^k and for each $x \in R^k$ let $f(x, \theta)$ be a continuously differentiable function on Θ . Let $\theta_1(x)$ and $\theta_2(x)$ be Borel measurable functions from R^k into Θ . Then there exists a Borel measurable function $\theta_*(x)$ from R^k into Θ such that

$$(I) \quad f(x, \theta_1(x)) - f(x, \theta_2(x)) = (\partial/\partial\theta)f(x, \theta_*(x))(\theta_1(x) - \theta_2(x)),$$

(II) $\theta_*(x)$ lies on the segment joining $\theta_1(x)$ and $\theta_2(x)$.

Proof: Jennrich (1969; lemma 3).

Thus, it follows from condition (I) and Lemma 4.2.1 that the mean values $\theta_n^{(i)}$ are measurable and that

$$|\theta_n^{(i)} - \theta_0| \leq |\theta_n - \theta_0| \text{ a.s.}$$

Q.E.D.

Exercises:

1. Let $Y_1, \dots, Y_n$ be independent standard normally distributed random variables. Let

$$Q_n(\theta) = (1/n) \sum_{j=1}^n |Y_j - \theta|, \quad \Theta = [a, b], \quad -\infty < a < 0 < b < \infty.$$

Let θ_n be defined in theorem 4.2.1. Prove $\theta_n \rightarrow 0$ a.s. Hint: use theorems 2.7.4 and 4.2.1.

2. Let $Y_1, \dots, Y_n$ be i.i.d. random variables satisfying

$$P(Y_j=1)=P(Y_j=-1)=\frac{1}{2}.$$

Let

$$Q_n(\theta) = (1/n) \sum_{j=1}^n (Y_j - \theta)^4$$

and let θ be as in exercise 1.

- Prove $\theta_n \rightarrow 0$ a.s.
- Prove $\sqrt{n} \theta_n \rightarrow N(0, \sigma^2)$ in distr.
- Determine σ^2 .

4.3 Consistency and asymptotic normality of nonlinear least squares estimators in the i.i.d. case

4.3.1 Consistency

We now consider the case where $(Y_1, X_1), \dots, (Y_n, X_n), \dots$ are independent identically distributed random vectors in $\mathbb{R} \times \mathbb{R}^k$, and we verify the conditions for consistency and asymptotic normality in section 4.2 for the nonlinear least squares estimator. Thus, we augment assumption 4.1.1 with:

Assumption 4.3.1. The random vectors (Y_j, X_j) , $j=1, 2, \dots, n, \dots$ are independent random drawings from a $k+1$ -variate distribution. Moreover, $E Y_j^2 < \infty$.

For proving strong consistency we use theorem 2.7.5 in order to show (4.2.1). First,

Assumption 4.3.2. Let $E \sup_{\theta \in \Theta} f(X_1, \theta)^2 < \infty$.

Then

$$E \sup_{\theta \in \Theta} [Y_1 - f(X_1, \theta)]^2 \leq 2E Y_1^2 + 2 E \sup_{\theta \in \Theta} f(X_1, \theta)^2 < \infty, \quad (4.3.1)$$

hence by theorem 2.7.5

$$\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \rightarrow 0 \text{ a.s.}, \quad (4.3.2)$$

where

$$Q(\theta) = E[Y_1 - f(X_1, \theta)]^2 \quad (4.3.3)$$

From assumption 4.1.1 it follows that

$$Y_1 = f(X_1, \theta_0) + U_1, \text{ with } E(U_1 | X_1) = 0 \text{ a.s.}, \quad (4.3.4)$$

hence

$$\begin{aligned}
Q(\theta) &= EU_1^2 + 2E U_1 [f(X_1, \theta_0) - f(X_1, \theta)] \\
&+ E[f(X_1, \theta_0) - f(X_1, \theta)]^2 = E U_1^2 + E[f(X_1, \theta_0) - f(X_1, \theta)]^2.
\end{aligned}
\tag{4.3.5}$$

Note that $E Y_1^2 < \infty$ implies $E U_1^2 < \infty$. Since $f(x, \theta)$ is continuous in θ for each $x \in R^k$ (cf. assumption 4.1.1), Q is continuous on Θ . Now assume:

Assumption 4.3.3. $Q(\theta)$ takes a unique global minimum on Θ at $\theta = \theta_0$.

Referring to theorem 4.2.1 we then conclude:

Theorem 4.3.1. Under assumptions 4.1.1 and 4.3.1-4.3.3., $\hat{\theta} \rightarrow \theta_0$ a.s.

4.3.2 Asymptotic normality

Next, we set forth the conditions for asymptotic normality by specializing the conditions of theorem 4.2.2. First,

Assumption 4.3.4. Let Θ be convex, let θ_0 be an interior point of Θ and let $f(x, \theta)$ be twice continuously differentiable in θ for each $x \in R^k$.

Then the conditions (I) and (II) of theorem 4.2.2 hold. Now consider the first and second derivatives of $Q(\theta)$:

$$(\partial/\partial\theta') \hat{Q}(\theta) = (-2)(1/n) \sum_{j=1}^n [Y_j - f(X_j, \theta)] (\partial/\partial\theta') f(X_j, \theta),
\tag{4.3.6}$$

$$\begin{aligned}
&(\partial/\partial\theta) (\partial\hat{Q}/\partial\theta') \\
&= (-2)(1/n) \sum_{j=1}^n [Y_j - f(X_j, \theta)] (\partial/\partial\theta) (\partial/\partial\theta') f(X_j, \theta) \\
&+ 2(1/n) \sum_{j=1}^n [(\partial/\partial\theta') f(X_j, \theta)] [(\partial/\partial\theta) f(X_j, \theta)]
\end{aligned}
\tag{4.3.7}$$

Assumption 4.3.5. Let for $i, i_1, i_2 = 1, 2, \dots, m$,

$$E \sup_{\theta \in \Theta} [(\partial/\partial \theta_{i_1}) f(X_1, \theta)]^2 < \infty;$$

$$E \sup_{\theta \in \Theta} [(\partial/\partial \theta_{i_1})(\partial/\partial \theta_{i_2}) f(X_1, \theta)]^2 < \infty.$$

Then by Cauchy-Schwarz inequality,

$$\begin{aligned} E \sup_{\theta \in \Theta} |[(\partial/\partial \theta_{i_1}) f(X_1, \theta)][(\partial/\partial \theta_{i_2}) f(X_1, \theta)]| \\ \leq (E \sup_{\theta \in \Theta} [(\partial/\partial \theta_{i_1}) f(X_1, \theta)]^2)^{1/2} \\ \times (E \sup_{\theta \in \Theta} [(\partial/\partial \theta_{i_2}) f(X_1, \theta)]^2)^{1/2} < \infty \end{aligned} \quad (4.3.8)$$

and

$$\begin{aligned} E \sup_{\theta \in \Theta} |[(Y_1 - f(X_1, \theta))(\partial/\partial \theta_{i_1})(\partial/\partial \theta_{i_2}) f(X_1, \theta)]| \\ \leq (E \sup_{\theta \in \Theta} [Y_1 - f(X_1, \theta)]^2)^{1/2} \\ \times (E \sup_{\theta \in \Theta} [(\partial/\partial \theta_{i_1})(\partial/\partial \theta_{i_2}) f(X_1, \theta)]^2)^{1/2} < \infty \end{aligned} \quad (4.3.9)$$

where the last inequality follows partly from (4.3.1). Using theorem 2.7.5 we now conclude from (4.3.8) and (4.3.9) that for $i_1, i_2 = 1, 2, \dots, m$.

$$\begin{aligned} \sup_{\theta \in \Theta} |(\partial/\partial \theta_{i_1})(\partial/\partial \theta_{i_2}) \hat{Q}(\theta) \\ - E(\partial/\partial \theta_{i_1})(\partial/\partial \theta_{i_2}) \hat{Q}(\theta)| \rightarrow 0 \text{ a.s.} \end{aligned} \quad (4.3.10)$$

which proves condition (IV) of theorem 4.2.2. Now consider the matrix A_2 :

$$\begin{aligned} A_2 &= (\partial/\partial \theta)(\partial/\partial \theta') Q(\theta_0) = E(\partial/\partial \theta)(\partial/\partial \theta') \hat{Q}(\theta_0) \\ &= -2 E[Y_1 - f(X_1, \theta_0)](\partial/\partial \theta)(\partial/\partial \theta') f(X_1, \theta_0) \\ &\quad + 2 E[(\partial/\partial \theta') f(X_1, \theta_0)][(\partial/\partial \theta) f(X_1, \theta_0)] \end{aligned} \quad (4.3.11)$$

Since $E[Y_1 - f(X_1, \theta_0) | X_1] = E(U_1 | X_1) = 0$ a.s., the first term

vanishes. Thus:

$$A_2 = 2\Omega_2, \quad (4.3.12)$$

where

$$\Omega_2 = E[(\partial/\partial\theta')f(X_1, \theta_0)][(\partial/\partial\theta)f(X_1, \theta_0)]. \quad (4.3.13)$$

Assumption 4.3.6. Let Ω_2 be nonsingular.

Then condition (V) of theorem 4.3.2 is satisfied. So it remains to show that condition (III) of theorem 4.2.2 holds. Observe from (4.3.4) and (4.3.6) that

$$\begin{aligned} (\partial/\partial\theta')\hat{Q}(\theta_0) &= -2(1/n)\sum_{j=1}^n U_j (\partial/\partial\theta')f(X_j, \theta) \\ &= -2(1/n)\sum_{j=1}^n Z_j, \end{aligned} \quad (4.3.14)$$

say, where (Z_j) , $j=1,2,\dots$, is a sequence of i.i.d. random vectors in R^k with

$$E Z_1 = E U_1 (\partial/\partial\theta')f(X_1, \theta_0) = E(E(U_1 | X_1)(\partial/\partial\theta')f(X_1, \theta_0)) = 0 \quad (4.3.15)$$

(for $E(U_1 | X_1) = 0$ a.s.). Moreover,

$$E Z_1 Z_1' = E U_1^2 \{(\partial/\partial\theta')f(X_1, \theta_0)\} \{(\partial/\partial\theta)f(X_1, \theta_0)\} = \Omega_1 \quad (4.3.16)$$

say. If

Assumption 4.3.7. For $i_1, i_2=1,2,\dots,m$,

$$E \sup_{\theta \in \Theta} [Y_1 - f(X_1, \theta)]^2 |(\partial/\partial\theta_{i_1})f(X_1, \theta)| |(\partial/\partial\theta_{i_2})f(X_1, \theta)| < \infty,$$

then the elements of the matrix Ω_1 are finite, hence we may apply theorem 2.4.1 to arbitrary linear combinations $\xi'Z_j$, ($\xi \in R^m$):

$$(1/\sqrt{n})\sum_{j=1}^n \xi' Z_j \rightarrow N(0; \xi' \Omega_1 \xi) \text{ in distr.} \quad (4.3.17)$$

From theorem 2.3.7 we then conclude:

$$(1/\sqrt{n})\sum_{j=1}^n Z_j \rightarrow N_m[0, \Omega_1] \text{ in distr.} \quad (4.3.18)$$

Thus

$$\sqrt{n}(\partial/\partial \theta') \hat{Q}(\theta_0) \rightarrow N_m(0, 4\Omega_1) \text{ in distr.}$$

This proves condition (III) of theorem 4.2.2 with $A_1 = 4\Omega_1$. Summarizing, we have shown:

Theorem 4.3.2. Under the conditions of theorem 4.3.1 and the additional assumptions 4.3.4-4.3.7 we have:

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N_m[0, \Omega_2^{-1} \Omega_1 \Omega_2^{-1}] \text{ in distr.,}$$

where Ω_1 and Ω_2 are defined by (4.3.16) and (4.3.13), respectively.

Cf. White (1980b). Quite often it is assumed that U_j is independent of X_j or that

Assumption 4.3.8. $E(U_1^2 | X_1) = E U_1^2 = \sigma^2$ a.s.

Then $\Omega_1 = \sigma^2 \Omega_2$, and moreover assumption 4.3.7 follows then from assumption 4.3.5 and the condition $E Y_1^2 < \infty$. Thus:

Theorem 4.3.3. Under the conditions of theorem 4.3.2 with, assumption 4.3.7 replaced by assumption 4.3.8, we have:

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N_m[0, \sigma^2 \Omega_2^{-1}] \text{ in distr.}$$

4.3.3 Consistent estimation of the asymptotic variance matrix

Finally, we turn to the problem how to consistently estimate the asymptotic variance matrices in theorems 4.3.2 and 4.3.3. Let

$$\hat{\Omega}_1(\theta) = (1/n) \sum_{j=1}^n [Y_j - f(X_j, \theta)]^2 [(\partial/\partial \theta') f(X_j, \theta)] [(\partial/\partial \theta) f(X_j, \theta)] \quad (4.3.19)$$

$$\hat{\Omega}_2(\theta) = (1/n) \sum_{j=1}^n [(\partial/\partial \theta') f(X_j, \theta)] [(\partial/\partial \theta) f(X_j, \theta)] \quad (4.3.20)$$

$$\hat{\Omega}_1 = \hat{\Omega}_1(\hat{\theta}), \quad \hat{\Omega}_2 = \hat{\Omega}_2(\hat{\theta}), \quad \hat{\sigma}^2 = \hat{Q}(\hat{\theta}) \quad (4.3.21)$$

We have shown that under the conditions of theorem 4.3.2:

$$\hat{\Omega}_1(\theta) \rightarrow E \hat{\Omega}_1(\theta) \text{ a.s. uniformly on } \theta, \quad (4.3.22)$$

$$\hat{\Omega}_2(\theta) \rightarrow E \hat{\Omega}_2(\theta) \text{ a.s. uniformly on } \theta, \quad (4.3.23)$$

$$\hat{Q}(\theta) \rightarrow E \hat{Q}(\theta) \text{ a.s. uniformly on } \theta \quad (4.3.24)$$

and

$$\Omega_1 = E \hat{\Omega}_1(\theta_0), \quad \Omega_2 = E \hat{\Omega}_2(\theta_0), \quad \sigma^2 = E U_1^2 = E \hat{Q}(\theta_0).$$

From theorems 2.6.1 and 4.3.1 it therefore follows:

$$\hat{\Omega}_1 \rightarrow \Omega_1 \text{ a.s.}; \quad \hat{\Omega}_2 \rightarrow \Omega_2 \text{ a.s.}, \quad \hat{\sigma}^2 \rightarrow \sigma^2 \text{ a.s.}, \quad (4.3.25)$$

hence, using theorem 2.1.7, we have:

Theorem 4.3.4. Under the conditions of theorem 4.3.2,

$$\hat{\Omega}_2^{-1} \hat{\Omega}_1 \hat{\Omega}_2^{-1} \rightarrow \Omega_2^{-1} \Omega_1 \Omega_2^{-1} \text{ a.s.},$$

and under the conditions of theorem 4.3.3 we have:

$$\hat{\sigma}^2 \hat{\Omega}_2^{-1} \rightarrow \sigma^2 \Omega_2^{-1} \text{ a.s.}$$

Exercises:

1. Verify the conditions for strong consistency and asymptotic normality of $\hat{\theta}$ for the standard linear regression model $Y_j = \theta_0' X_j + U_j$, where U_j and X_j are mutually independent.
2. Suppose that $f(x, \theta)$ is incorrectly specified as $\theta'x$, with $\theta \in \Theta \subset \mathbb{R}^k$, $x \in \mathbb{R}^k$. Set forth conditions such that $\hat{\theta} \rightarrow \theta_0$ a.s. and

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N_k[0, (E X_1 X_1')^{-1} (E(Y_1 - \theta_0' X_1) X_1 X_1') (E X_1 X_1')^{-1}]$$

in distr., where $\theta_0 = (E X_1 X_1')^{-1} E X_1 Y_1$. Cf. White (1980a)

4.4 Consistency and asymptotic normality of the nonlinear least squares estimator under data heterogeneity

4.4.1 Data heterogeneity

In medicine, biology, psychology and physics the source of the data set $\{(Y_j, X_j)\}$, $j=1, \dots, n$, is quite often a statistical experiment, where Y_j is the outcome of the experiment and X_j is a vector of control variables set by the analyst. In this case the vector X_j is in fact nonrandom. However, nonrandom variables or vectors may be considered as random variables or vectors, respectively, taking a value with probability 1. Thus, if (x_j) , $j=1, 2, \dots$ is a sequence of control vectors then we define:

$$X_j = x_j \text{ a.s.}$$

The distribution function of X_j is then

$$H_j(x) = \prod_{\ell=1}^k I(X_{\ell j} \leq x^{(\ell)}),$$

where $I(\cdot)$ is the indicator function and $X_{\ell j}$ and $x^{(\ell)}$ are the ℓ -th components of X_j and x , respectively. Moreover, suppose

$$E Y_j = g(x_j), \quad U_j = Y_j - g(x_j).$$

Then

$$Y_j = g(X_j) + U_j \text{ a.s.,}$$

where $E U_j = 0$ and U_j is independent of X_j . The joint distribu-

tion function $F_j(y, x)$ of (Y_j, X_j) is now:

$$F_j(y, x) = P[U_j \leq y - g(x_j)]H_j(x).$$

These kind of data-generating processes will be called *heterogeneous*. Data heterogeneity also occurs when the sample is stratified. For example, let Y_j be the expenditures of household j on a certain commodity and let X_j be the disposable income of household j . We wish to estimate the Engel curve $E(Y_j | X_j)$ (cf. Cramer (1969)) on the basis of a stratified sample of size n with q strata of different sizes n_i ($i=1, \dots, q$, $n=n_1+\dots+n_q$), corresponding to q income classes. Then the distribution of (Y_j, X_j) depends on the stratum to which j belongs. See further White (1980b).

In order to cover data heterogeneity as well, we now augment assumption 4.1.1 with:

Assumption 4.4.1. The random vectors (Y_j, X_j) , $j=1, 2, \dots, n, \dots$ are independent with joint distribution functions $F_j(y, x)$, $j=1, 2, \dots, n, \dots$, respectively. Moreover, denoting

$$F^{(n)}(y, x) = (1/n) \sum_{j=1}^n F_j(y, x),$$

one of the following alternative conditions hold:

- (a) $F^{(n)} \rightarrow F$ properly, setwise,
- (b) $F^{(n)} \rightarrow F$ properly, pointwise, and $f(x, \theta)$ is continuous in both of its arguments.

The distinction between the alternatives (a) and (b) is due to the distinction between the conditions of theorems 2.7.1 through 2.7.4, i.e., under assumption 4.4.1 (a) we shall apply theorems 2.7.3 and 2.7.4 and under assumption 4.4.1 (b) we shall apply theorems 2.7.1 and 2.7.2.

4.4.2 Strong and weak consistency

Now let us modify the assumptions of section 4.3 to the case under review. Depending on whether we wish to prove weak consistency, i.e.

$$\text{plim}_{n \rightarrow \infty} \hat{\theta} = \theta_0,$$

or strong consistency, i.e.

$$\hat{\theta} \rightarrow \theta_0 \text{ a.s.}$$

we assume either

Assumption 4.4.2.(I). Let for some $\delta > 0$,

$$\sup_n (1/n) \sum_{j=1}^n E \sup_{\theta \in \Theta} |Y_j - f(X_j, \theta)|^{2+\delta} < \infty,$$

or

Assumption 4.4.2.(II). Let for some $\delta > 0$,

$$\sup_n (1/n) \sum_{j=1}^n E \sup_{\theta \in \Theta} |Y_j - f(X_j, \theta)|^{4+\delta} < \infty.$$

Denoting

$$Q(\theta) = \int [y - f(x, \theta)]^2 dF(y, x) \quad (4.4.1)$$

it follows from assumptions 4.1.1, 4.4.1 and 4.4.2(I) and theorems 2.7.2 and 2.7.4 that

$$\text{plim}_{n \rightarrow \infty} \sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| = 0, \quad (4.4.2)$$

whereas from assumptions 4.1.1., 4.4.1, 4.4.2.II and theorems 2.7.1 and 2.7.3 it follows:

$$\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \rightarrow 0 \text{ a.s.} \quad (4.4.3)$$

Moreover, in both cases the limit function $Q(\theta)$ is continuous. Furthermore, it follows from theorems 2.5.1 and 2.5.4,

$$\begin{aligned} Q(\theta) &= \lim_{n \rightarrow \infty} E \hat{Q}(\theta) \\ &= \lim_{n \rightarrow \infty} (1/n) \sum_{j=1}^n E [U_j + f(X_j, \theta_0) - f(X_j, \theta)]^2 \\ &= \lim_{n \rightarrow \infty} (1/n) \sum_{j=1}^n E U_j^2 + \lim_{n \rightarrow \infty} (1/n) \sum_{j=1}^n E [f(X_j, \theta_0) - f(X_j, \theta)]^2 \end{aligned} \quad (4.4.4)$$

hence

$$Q(\theta_0) = \inf_{\theta \in \Theta} Q(\theta). \quad (4.4.5)$$

Thus, if

Assumption 4.4.3. $Q(\theta)$ takes a unique minimum on Θ ,

then by theorem 4.2.1:

Theorem 4.4.1.

- (I) Under assumptions 4.1.1, 4.4.1, 4.4.2(I) and 4.4.3,
 $\text{plim}_{n \rightarrow \infty} \hat{\theta} = \theta_0$.
- (II) Under assumptions 4.1.1, 4.4.1, 4.4.2(II) and 4.4.3,
 $\hat{\theta} \rightarrow \theta_0$ a.s.

4.4.3 Asymptotic normality

Next we shall modify the assumptions 4.3.4-4.3.8 such that we may replace the reference to theorem 2.7.5 by references to theorems 2.7.2 and 2.7.4, and the reference to theorem 2.4.1 by a reference to theorem 2.4.3.

Assumption 4.4.4. Let assumption 4.3.4 hold. If part (b) of assumption 4.4.1 hold then the first and second partial derivatives of $f(x, \theta)$ to θ are continuous in both arguments.

Assumption 4.4.5. Let for some $\delta > 0$ and $i, i_1, i_2 = 1, 2, \dots, m$,

$$\sup_n (1/n) \sum_{j=1}^n E \sup_{\theta \in \Theta} |(\partial/\partial \theta_{i_1}) f(X_j, \theta)|^{2+\delta} < \infty;$$

$$\sup_n (1/n) \sum_{j=1}^n E \sup_{\theta \in \Theta} |(\partial/\partial \theta_{i_1})(\partial/\partial \theta_{i_2}) f(X_j, \theta)|^{2+\delta} < \infty.$$

Assumption 4.4.6. The matrix

$$\Omega_2 = \int [(\partial/\partial \theta') f(x, \theta_0)] [(\partial/\partial \theta) f(x, \theta_0)] dF(y, x)$$

is nonsingular.

Assumption 4.4.7. Let for some $\delta > 0$ and $i_1, i_2 = 1, 2, \dots, m$,

$$\sup_n (1/n) \sum_{j=1}^n E \sup_{\theta \in \Theta} |Y_j - f(X_j, \theta)|^{2+\delta} |(\partial/\partial \theta_{i_1}) f(X_j, \theta)|^{1+\delta} \\ \times |(\partial/\partial \theta_{i_2}) f(X_j, \theta)|^{1+\delta} < \infty.$$

Assumption 4.4.8. For $j=1, 2, 3, \dots$, $E(U_j^2 | X_j) = E U_j^2 = \sigma^2$ a.s.

Moreover, let

$$\Omega_1 = \int [y - f(x, \theta_0)]^2 [(\partial/\partial \theta') f(x, \theta_0)] [(\partial/\partial \theta) f(x, \theta_0)] dF(y, x).$$

Then similarly to theorems 4.3.2, 4.3.3 and 4.3.4 we have:

Theorem 4.4.2. Let the conditions of theorem 4.4.1 (part I) be satisfied.

(I) Under the additional conditions 4.4.4-4.4.7 we have:

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N_m[0, \Omega_2^{-1} \Omega_1 \Omega_2^{-1}] \text{ in distr.},$$

$$\text{plim}_{n \rightarrow \infty} \hat{\Omega}_2^{-1} \hat{\Omega}_1 \hat{\Omega}_2^{-1} = \Omega_2^{-1} \Omega_1 \Omega_2^{-1}.$$

(II) Under the additional conditions 4.4.4-4.4.6, 4.4.8 we have:

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N_m[0, \sigma^2 \Omega_2^{-1}] \text{ in distr. and}$$

$$\text{plim}_{n \rightarrow \infty} \hat{\sigma}^2 \hat{\Omega}_2^{-1} = \sigma^2 \Omega_2^{-1}.$$

4.5 Testing parameter restrictions: The Wald test

Estimation of a regression model is usually not the final stage of regression analysis. What we actually want to know is the properties of the true model rather than of the estimated model. Given the truth of the functional form, we thus want to make inference about the true parameter vector θ_0 . Various restrictions on θ_0 may correspond to various theories about the phenomenon under review. In particular the theory of interest may correspond to a set of q (non)linear restrictions in θ_0 , i.e.,

$$H_0 : \eta_i(\theta_0) = 0 \text{ for } i=1,2,\dots,q \leq m \quad (4.5.1)$$

Throughout we assume that the functions η_i are continuously differentiable in a neighborhood of θ_0 and that the $q \times m$ matrix

$$\Gamma = [(\partial/\partial\theta_{i_2})\eta_{i_1}(\theta_0)] \quad (4.5.2)$$

has rank q . We now derive the limiting distribution of

$$(\eta_1(\hat{\theta}), \dots, \eta_q(\hat{\theta}))'$$

under H_0 , given that

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N_m[0, \Omega], \quad (4.5.3)$$

where Ω is nonsingular, and that $\hat{\Omega}$ is a consistent estimator of Ω :

$$\text{plim}_{n \rightarrow \infty} \hat{\Omega} = \Omega. \quad (4.5.4)$$

Observe from the mean value theorem that

$$\sqrt{n}(\eta_i(\hat{\theta}) - \eta_i(\theta_0)) = [(\partial/\partial\theta)\eta_i(\hat{\theta}^{(i)})]\sqrt{n}(\hat{\theta} - \theta_0) \quad (4.5.5)$$

where $\hat{\theta}^{(i)}$ satisfies

$$|\hat{\theta}^{(i)} - \theta_0| \leq |\hat{\theta} - \theta_0| \text{ a.s.} \quad (4.5.6)$$

Since (4.5.3) implies

$$\text{plim}_{n \rightarrow \infty} \hat{\theta}^{(i)} = \theta_0,$$

it follows from theorem 2.6.1 that

$$\text{plim}_{n \rightarrow \infty} (\partial/\partial\theta)\eta_i(\hat{\theta}^{(i)}) = (\partial/\partial\theta)\eta_i(\theta_0) \quad (4.5.7)$$

and consequently by (4.5.3) and theorems 2.3.3 and 2.3.5,

$$\text{plim}_{n \rightarrow \infty} [(\partial/\partial\theta)\eta_i(\hat{\theta}^{(i)}) - (\partial/\partial\theta)\eta_i(\theta_0)]\sqrt{n}(\hat{\theta} - \theta_0) = 0 \quad (4.5.8)$$

Combining (4.5.5) and (4.5.8) it follows

$$\text{plim}_{n \rightarrow \infty} (\sqrt{n}(\eta_1(\hat{\theta}) - \eta_1(\theta_0)) - (\partial/\partial \theta) \eta_1(\theta_0) \sqrt{n}(\hat{\theta} - \theta_0)) = 0 \quad (4.5.9)$$

and consequently

$$\text{plim}_{n \rightarrow \infty} (\sqrt{n}(\hat{\eta} - \eta) - \Gamma \sqrt{n}(\hat{\theta} - \theta_0)) = 0, \quad (4.5.10)$$

where

$$\eta = (\eta_1(\theta_0), \dots, \eta_i(\theta_0))' \quad (4.5.11)$$

Since (4.5.3) implies that

$$\Gamma \sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N_q(0, \Gamma \Omega \Gamma') \text{ in dist.},$$

it follows now from (4.5.10),

$$\sqrt{n}(\hat{\eta} - \eta) \rightarrow N_q[0, \Gamma \Omega \Gamma'] \text{ in distr.} \quad (4.5.12)$$

Note that $\Gamma \Omega \Gamma'$ has rank q and is therefore nonsingular. Now let

$$\hat{\Gamma} = [(\partial/\partial \theta_{i_2}) \eta_{i_1}(\hat{\theta})] \quad (4.5.13)$$

Then $\text{plim}_{n \rightarrow \infty} \hat{\theta} = \theta_0$ implies

$$\text{plim}_{n \rightarrow \infty} \hat{\Gamma} = \Gamma \quad (4.5.14)$$

and thus

$$\text{plim}_{n \rightarrow \infty} \hat{\Gamma} \hat{\Omega} \hat{\Gamma}' = \Gamma \Omega \Gamma'. \quad (4.5.15)$$

Using theorem 2.3.5 we conclude from (4.5.12) and (4.5.15) that

$$(\hat{\Gamma} \hat{\Omega} \hat{\Gamma}')^{-1/2} \sqrt{n}(\hat{\eta} - \eta) \rightarrow N_q(0, I) \text{ in distr.} \quad (4.5.16)$$

and consequently by theorem 2.3.4,

$$n(\hat{\eta} - \eta)' (\hat{\Gamma} \hat{\Omega} \hat{\Gamma}')^{-1} (\hat{\eta} - \eta) \rightarrow \chi_q^2 \text{ in distr.} \quad (4.5.17)$$

The Wald statistic for testing $H_0 : \eta = 0$ is now

$$W_n = n\hat{\eta}'(\hat{\Gamma}'\hat{\Omega}\hat{\Gamma}')^{-1}\hat{\eta}.$$

Clearly,

$$W_n \rightarrow \chi_q^2 \text{ in distr. if } H_0 \text{ is true.}$$

If H_0 is false, i.e., $\eta \neq 0$, then

$$\text{plim}_{n \rightarrow \infty} W_n/n = \eta'(\Gamma\Omega\Gamma')^{-1}\eta > 0,$$

hence

$$\text{plim}_{n \rightarrow \infty} W_n = \infty.$$

Remark. A drawback of the Wald test is that changing the form of the nonlinear restrictions $\eta_i(\theta_0) = 0$ to a form which is algebraically equivalent under the null hypothesis will change the Wald statistic W_n . Recently, Gregory and Veall (1985) have shown on the basis of Monte Carlo experiments that differences in functional form of the restrictions are likely to be important in small samples.

Note that the result in this section is in fact a straightforward corollary of the following general theorem.

Theorem 4.5.1. Let (θ_n) be a sequence of random vectors in R^k satisfying $r_n(\theta_n - \theta_0) \rightarrow N_k(0, \Omega)$, where (r_n) is a sequence of real numbers converging to infinity and $\theta_0 \in R^k$ is nonrandom. Let $\eta_1(\theta), \dots, \eta_m(\theta)$ be Borel measurable real functions on R^k such that for $j=1, \dots, m$, $\eta_j(\theta)$ is continuously differentiable in θ_0 . Let $\eta(\theta) = (\eta_1(\theta), \dots, \eta_m(\theta))'$ and let Γ be the $q \times m$ matrix

$$\Gamma = [(\partial/\partial\theta_{i_2})\eta_{i_1}(\theta_0)]$$

Then $r_n[\eta(\theta_n) - \eta(\theta_0)] \rightarrow N_m[0, \Gamma\Omega\Gamma']$ in distr. (N.B. m may be larger than k)

References

- Arrow, K.J., H.B.Chenery, B.S.Minhas and R.M.Solow (1961), "Capital-Labor Substitution and Economic Efficiency", *Review of Economics and Statistics* 43, 225-250.
- Box, G.E.P. and D.R.Cox (1964), "An Analysis of Transformations", *Journal of the Royal Statistical Society, Series B* 26, 211-243.
- Cramer, J.S. (1969), *Empirical Econometrics*, Amsterdam: North-Holland.
- Don, F.J.H. (1986), "Constrained Non-Linear Least Squares", *Statistica Neerlandica* 40, 109-115.
- Gregory, A.W. and M.R.Veall (1985), "Formulating Wald Tests of Non-linear Restrictions", *Econometrica* 53, 1465-1468.
- Jennrich, R.I. (1969), "Asymptotic Properties of Nonlinear Least Squares Estimators", *Annals of Mathematical Statistics* 40, 633-643.
- Malinvaud, E. (1970a), *Statistical Methods of Econometrics*, Amsterdam: North Holland
- Malinvaud, E. (1970b), "The Consistency of Nonlinear Regression", *Annals of Mathematical Statistics* 41, 956-969
- White, H. (1980a), "Using Least Squares to Approximate Unknown Regression Functions", *International Economic Review* 21, 149-170.
- White, H. (1980b), "Nonlinear Regression on Cross-Section Data", *Econometrica* 48, 721-746.
- White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models", *Econometrica* 50, 1-25.

P56970-E