ESTIMATION OF THE PARAMETERS IN AN IMPLICIT MODEL
BY MINIMIZING THE SUM OF ABSOLUTE VALUES OF ORDER P

by

Sonia Balet-Lawrence

Institute of Statistics
Mimeograph Series No. 1073
Raleigh - 1976
ESTIMATION OF THE PARAMETERS IN AN IMPLICIT MODEL BY 
MINIMIZING THE SUM OF ABSOLUTE VALUES OF ORDER P 

by 

SONIA BALET-LAWRENCE 

A thesis submitted to the Graduate Faculty of North Carolina State University at Raleigh in partial fulfillment of the requirements for the Degree of Doctor of Philosophy 

DEPARTMENT OF STATISTICS 

RALEIGH 

1975 

APPROVED BY:

Co-Chairman of Advisory Committee
LAWRENCE, SONIA BALET. Estimation of the Parameters in an Implicit Model by Minimizing the Sum of Absolute Values of Order $P$. (Under the direction of BIBHUTI B. BHATTACHARYYA and A. RONALD GALLANT.)

Consider the structural relation given by

$$Q(x_t, y_t, \theta^*) = \epsilon_t,$$

where $\{x_t, y_t\}_{t=1}^n$ is the observed sample, $\theta^*$ is the unknown vector of parameters, and $\epsilon_t$ is a random error due to the combined effect of omitted variables. The statistical problem is to estimate $\theta^*$.

If an analytic solution for the endogenous variable $y_t$ (or a function of it) does not exist in terms of exogenous variables $x_t$ and parameters plus an error term, the well-known techniques for estimating $\theta^*$ do not apply directly. An estimator $\hat{\theta}_n$ that considers this formulation of the model is proposed. The regression case is considered as a special type of implicit model.

The estimator $\hat{\theta}_n$ is defined as that value in $\Omega \subset \mathbb{R}^k$ such that

$$\min_{\theta \in \Omega} \sum_{t=1}^n |Q(x_t, y_t, \theta)|^P = \sum_{t=1}^n |Q(x_t, y_t, \hat{\theta}_n)|^P$$

for some fixed $p \geq 1$. A computing algorithm is derived and the statistical properties of $\hat{\theta}_n$ are investigated. In particular, it is shown that under certain conditions for any $p \geq 1$, $\hat{\theta}_n \to \theta_0$ with probability one as $n \to \infty$, where $\theta_0$ is the point in the interior of $\Omega$ satisfying

$$E_{\theta^*}[|Q(x, y, \theta_0)|^P] < E_{\theta^*}[|Q(x, y, \theta)|^P]$$
for any $\theta \neq \theta_0$ and $\theta$ in $\Omega$. If more than one value in $\Omega$ minimizes the above expectation, then for $n$ sufficiently large $\hat{\theta}_n$ remains in the neighborhood of any one of these points with probability one. For a choice of $p > 1$ the estimator $\hat{\theta}_n$ is shown to be asymptotically normal with mean $\theta_0$.

When the regression model is considered then under more specific conditions set on the class of distribution functions for the error term it is shown that $\theta_0 = \theta^*$ and hence the estimator is strongly consistent for $p \geq 1$ and the asymptotic distribution for $\hat{\theta}_n$ based on $p > 1$ is centered about the true parameter value $\theta^*$.

A predictor $\hat{y}_n$ for the endogenous variable $y$ is defined for the case in which an analytic or numerical solution for $y$ exists for a given value of $x$. 
The author was born on November 4, 1947, in San Juan, Puerto Rico. After receiving a Bachelor's degree in Business Administration from the University of Puerto Rico, she was granted financial support from this university to pursue studies in statistics towards a Masters at North Carolina State University.

Studies during 1970-71 were done with financial support from the National Science Foundation.

After completion of the Master's degree the author worked as an instructor at the University of Puerto Rico. In 1972, she returned to North Carolina State University to fulfill the requirements for a Doctorate's degree.
ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to the Co-Chairmen of my graduate committee, Drs. B. B. Bhattacharyya and A. R. Gallant for their guidance in the preparation of this thesis and their valuable assistance in the organization and writing of the final draft. Thanks also go to Drs. F. Gisbrecht and R. Schrimper for serving in this committee and for their cooperation given short notice during the last stages of this thesis.

Special thanks go to Dr. Thomas D. Wallace for his assistance throughout my five years of graduate work and for suggesting what eventually became part of this research.

To Linda Granger, I owe thanks for her prompt and careful typing of this work.

I am greatly indebted to the University of Puerto Rico for financial support during the first year of my graduate work and above all for being a main source of motivation in the completion of my degree.

To my parents I wish to express my gratitude for their constant encouragement.

To my husband, John, I dedicate this thesis.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. OVERVIEW OF THE PROBLEM</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Background</td>
<td>2</td>
</tr>
<tr>
<td>2. THE MODEL AND GENERAL CONCLUSIONS</td>
<td>17</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Structure and Model Specifications</td>
<td>17</td>
</tr>
<tr>
<td>2.3 The General Assumptions</td>
<td>23</td>
</tr>
<tr>
<td>3. THE ESTIMATOR</td>
<td>30</td>
</tr>
<tr>
<td>3.1 Definition of The Estimator $\hat{\theta}_n$</td>
<td>30</td>
</tr>
<tr>
<td>3.2 Computation of $\hat{\theta}_n$</td>
<td>32</td>
</tr>
<tr>
<td>3.3 Convergence of The Algorithm</td>
<td>43</td>
</tr>
<tr>
<td>4. GENERAL RESULTS</td>
<td>49</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>49</td>
</tr>
<tr>
<td>4.2 Convergence With Probability One</td>
<td>56</td>
</tr>
<tr>
<td>4.3 Asymptotic Normality of $\hat{\theta}_n$</td>
<td>63</td>
</tr>
<tr>
<td>4.4 The Predictor</td>
<td>76</td>
</tr>
<tr>
<td>5. THE CASE OF $Q(x_t, y_t, \theta^<em>) = y_t - f(x_t, \theta^</em>)$</td>
<td>81</td>
</tr>
<tr>
<td>5.1 Strong Consistency of $\hat{\theta}_n$</td>
<td>82</td>
</tr>
<tr>
<td>5.2 Asymptotic Normality $\hat{\theta}_n$</td>
<td>88</td>
</tr>
<tr>
<td>6. SUMMARY</td>
<td>113</td>
</tr>
<tr>
<td>7. NUMERICAL ILLUSTRATIONS</td>
<td>115</td>
</tr>
<tr>
<td>8. LITERATURE CITED</td>
<td>119</td>
</tr>
</tbody>
</table>
1. OVERVIEW OF THE PROBLEM

1.1 Introduction

Conventional regression models usually are of the form

\[ y_t = f(x_t, \theta^*) + \epsilon_t \]

where \( y_t \) is the observed response variable, \( f \) is the hypothesized response function in terms of exogenous variables \( x_t \), and parameters, \( \theta^* \) is the unknown vector of parameters to be estimated and \( \epsilon_t \) is a random error uncorrelated with \( x_t \). This equation attempts to satisfy the scientist's need to express a hypothetical relationship in manageable terms. Most models are derived in this form and the well-known techniques for estimating \( \theta^* \) are oriented to this specific algebraic formulation. That is, a response variable on the left hand side of the equation, and a function of exogenous variables and parameters plus an error term, on the right hand, is the eventual form of the model prior to the estimation stage.

It is possible, however, that theoretical considerations give rise to a structural relation

\[ Q(x_t, y_t, \theta^*) = \epsilon_t \]

which is an implicit function of the variable \( y \). One problem with this formulation is that an analytic solution for \( y_t \) may not exist and the well-established approaches for estimating \( \theta^* \) do not apply directly. A lesser problem is that when known numerical methods are able to generate a solution for \( y_t \), given trial values for \( \theta^* \) and
known values of $x_t$, (and equating $\epsilon_t$ to zero), the process of estimating $\theta^*$ under standard techniques may be cumbersome and costly.

Furthermore, in the case where an analytic solution for $y_t$ does exist, as in the regression model, various alternatives to the least squares approach have been explored mostly by simulation studies. However the statistical properties of the generated estimators have not been formally derived given a general response function $f(x, \theta)$, not necessarily linear in the unknown parameters.

It is to these problems that this dissertation is addressed. More specifically, we investigate the properties of a class of estimators generated under no requirement that an analytic or a unique numerical solution exists for the endogenous variable $y$. The regression model is considered as a special case, i.e., a conceptual implicit model having an analytic solution for the response $y$ which is expressed as a function of parameters, and variables uncorrelated with the additive error term.

1.2 Background

Data available for analysis usually consists of a set of observations $\{x_t, y_t\}_{t=1}^n$ where $\{y_t\}_{t=1}^n$ is the sequence of responses on the variable $y$, obtained for different values of the exogenous variables $x$. The functional form of the response may arise from theoretical considerations, but in many situations it is approximated by a simpler function like a polynomial of fixed degree in $x$. 
A model is said to be in reduced form if it is expressed by a system of equations (one or more) with one endogenous variable on the left hand side and a function of variables uncorrelated with the error term on the right hand side of each equation. Alternatively, a structural or implicit representation of the model is of the form

$$Q(x_t, y_t, \theta^*) = \epsilon_t.$$ 

In many applications the interest lies in constructing an equation that will predict and hopefully explain the behavior of the response variable; the model is therefore formulated accordingly. If the derived model is in implicit form, the approach is to obtain the reduced form prior to the estimation stage.

Given the data and the reduced form, the least squares technique is the most frequently used method for estimating the parameter $\theta^*$. Other criteria have been considered. In particular, Huber [9] and [11], considered the model

$$y_t = \sum_{i=1}^{k} x_{ti} \theta_i + \epsilon_t$$

and his class of estimators included $\hat{\theta}_n$, generated by minimizing

$$\sum_{t=1}^{n} \sum_{i=1}^{k} |y_t - \sum_{i=1}^{k} x_{ti} \hat{\theta}_i|^p$$

for $p \geq 1$. Under general conditions he derived the consistency and asymptotic normality of $\hat{\theta}_n$. Smith and Hall [18] and Forsythe [7] conducted simulation studies that suggested the following: Under the model

$$y_t = \sum_{i=1}^{k} x_{ti} \theta_i + \epsilon_t$$
if the probability distribution for the error \( \varepsilon_t \) is in the double exponential family or in the family of contaminated normals with density function

\[
f(\varepsilon | G, R, S) = G \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \varepsilon^2} + (1-G) \frac{1}{R\sqrt{2\pi}} e^{-\frac{1}{2} \left(\varepsilon - \frac{S}{R}\right)^2}
\]

for \( 0 \leq G \leq 1 \), \( R > 0 \), then \( \hat{\theta}_n \) satisfying

\[
\min_{\theta \in \Omega} \sum_{t=1}^{n} \left| y_t - \sum_{i=1}^{k} x_{ti} \theta_i \right|^p = \sum_{t=1}^{n} \left| y_t - \sum_{i=1}^{k} x_{ti} \hat{\theta}_i \right|^p
\]

for \( 1 < p < 2 \) in general has smaller sample mean square error than the estimator generated by the least squares approach. Note that the above density is symmetric when \( S = 0 \), and skewed when \( R = 1 \) and \( S \neq 0 \). In both cases the alternative estimator showed substantial increase in efficiency as the number of observations and the degree of contamination gets larger.

In this same context, the class of symmetric distributions with density given by

\[
f(\varepsilon | \sigma, \beta) = w \exp\left[-\frac{1}{2} \left(\frac{\varepsilon}{\sigma}\right)^{1+\beta}\right]
\]

for \( -\infty < \varepsilon < \infty \), \( 0 < \sigma < \infty \), \(-1 < \beta < 1\) and

\[
\frac{1}{w} = \Gamma\left(1 + \frac{1}{2}(1 + \beta)\right) \cdot 2^{\left(1 + \frac{1}{2}(1 + \beta)\right)} \cdot \sigma
\]

has been studied by Box and Tiao [2] as an alternative to the usual assumption of a normal distribution for the error term \( \varepsilon_t \). In particular note that if \( \beta = 0 \) we have the normal density, the double exponential corresponds to \( \beta = 1 \), and as \( \beta \to -1 \) the distribution
tends to the uniform. We remark that for this class of distributions if \( p = \left( \frac{2}{1+A} \right) \) and \( \sigma \) are known, then for the explicit model, maximizing the likelihood corresponds to the estimator \( \hat{\theta}_n \) generated by minimizing

\[
\sum_{t=1}^{n} |y_t - f(x_t, \theta)|^p.
\]

Hence, at least for \( n \) sufficiently large, minimizing the sum of \( p \)-th absolute deviations may give more efficient estimates than those generated by the least squares approach.

The implicit model may arise from considerations of the theoretical relationships underlying an observed process. Frequently, a system of differential equations, or the set of necessary conditions that must be satisfied when the process is in some optimal state, constitute the derived model. It is possible that these equations are algebraically implicit in the dependent variable. A specific model appears in Wallace and Ihnen [20]. Situations may also arise in practice in which all variables observed may be viewed as being generated from a joint probability distribution. In this case it is meaningless to distinguish between dependent and independent variables. If the variables are correlated with the error term then we might say that the usual regression model is actually in implicit form.

Usually, these models are originally stated in deterministic form. The experimenter derives a relation of the form

\[
Q(x_t, y_t, \theta^*) = 0.
\]
Yet, because of measurement error, the use of alternative measures for a given physical quantity, or due to neglected effects, an error component gets introduced into the model. We have chosen to view the data as measured without error, but to consider the combined effect of non-specified variables as manifested through the term $\epsilon_t$. The model becomes

$$Q(x_t, y_t, \theta^*) = \epsilon_t$$

and the solution to the problem of estimation relies on the prior information concerning the error term $\epsilon_t$ and the parameter space $\Omega$ that the experimenter is willing to add to the structure.

Given a probability distribution for the error $\epsilon_t$, then a maximum likelihood approach may be used and under general conditions the estimator of $\theta^*$ is consistent and asymptotically normal.

In the case of the model

$$y_t - f(x_t, \theta^*) = \epsilon_t$$

assuming $E(\epsilon_t | x_t = x) = 0$ for all $x$, also implies that

$$E_{y|x_t = x} (y_t - f(x_t, \theta^*))^2 \leq E_{y|x_t = x} (y_t - f(x_t, \theta))^2$$

for any $\theta \neq \theta^*$. Under additional restrictions on the sequence \(\{x_t\}_{t=1}^{\infty}\) the above relation becomes a strict inequality when the conditional expectation is averaged over all $x_t$ for $n$ sufficiently large. Minimizing

$$\sum_{t=1}^{n} (y_t - f(x_t, \theta))^2$$
under general conditions then leads to a consistent and asymptotically normal estimator of $\theta^*$. When the model is approached in its implicit form, it is necessary that the prior information on the model be sufficient to distinguish the true parameter values from all others. However, in many real situations knowledge of the distribution function of $\epsilon_t$ is at most highly uncertain.

As Fisher [5] pointed out, even in the case of a relation of the form

$$\beta y_t + \gamma^T x_t = \epsilon_t,$$

assuming a normal distribution with constant variance and zero expectation for all $\epsilon_t$ is not sufficient to distinguish between $(\beta, \gamma^T)$ and $(c\beta, c\gamma^T)$, for $c$ a scalar different from zero. Usually, a normalization rule is imposed on the original parameters in order to define a unique value for $(\beta, \gamma^T)$ that is in accordance with the observed data and the model specifications.

In considering a specific model, knowledge on the behavior of the errors associated with the process, may impose additional restrictions on $\theta^*$ which will define the true vector. One such restriction could relate to some measure of dispersion for the random variable $Q(x, y, \theta)$ such that at $\theta = \theta^*$ the variability of $Q$ around the value zero, as measured by the function $\rho(Q(x, y, \theta))$, satisfies

$$\rho(Q(x, y, \theta^*)) < \rho(Q(x, y, \theta))$$

for any $\theta \neq \theta^*$. 
With these comments in mind, we state the purpose of this thesis as twofold:

1. To consider the model in its implicit form

\[ Q(x_t, y_t, \theta^*) = \epsilon_t \]

where \( E(\epsilon_t) = 0 \) and no specific distribution function is assumed for \( \epsilon_t \). The standard regression model is simply a special case.

2. To explore the statistical properties of the estimator \( \hat{\theta}_n \) satisfying

\[
\min_{\theta \in \Omega} \sum_{t=1}^{n} \left| Q(x_t, y_t, \theta) \right|^p = \sum_{t=1}^{n} \left| Q(x_t, y_t, \hat{\theta}_n) \right|^p
\]

for a fixed \( p \geq 1 \).

In Chapter 2 we define the implicit model in more detail and present the major assumptions underlying the results derived in later chapters. Chapter 3 develops an iterative computing scheme for finding \( \hat{\theta}_n \) when \( p > 1 \), \( n \geq k \) and \( k \) is the number of unknown parameters. Under certain conditions, the sequence generated by the algorithm is shown to converge to a stationary point of the objective function

\[ \psi_n(\theta) = \sum_{t=1}^{n} \left| Q(x_t, y_t, \theta) \right|^p. \]

The general properties of the estimator \( \hat{\theta}_n \) are derived in Chapter 4. The sequence of estimators \( \{\hat{\theta}_n\}_{n=1}^{\infty} \) is shown to be asymptotically normal and to converge with probability one to a point \( \theta_0 \) in \( \Omega \). In
Section 3 of this chapter the predictor $\hat{y}$ is defined, under the assumption that the equation

$$Q(x, \hat{y}, \hat{\theta}_n) = 0$$

has a solution in $\hat{y}$ given values of $x$ and $\hat{\theta}_n$.

The general results of Chapter 4 are applied to the regression model in Chapter 5. For the explicit regression model where

$$Q(x_t, y_t, \theta*) = y_t - f(x_t, \theta*),$$

the estimator $\hat{\theta}_n$ is strongly consistent and asymptotically normal with the asymptotic distribution centered around the true parameter value $\theta*$.

Some numerical illustrations on the performance of the algorithm are included in Chapter 7. In some cases, the results are compared to those generated by other computing schemes.

### 1.3 Review of the Literature

In the case of one dependent variable $y_t$ and an $m$-vector of independent variables $x_t$ measured without error, the usual approach to the estimation of $\theta*$ assumes a model of the form

$$Q(x_t, y_t^0, \theta*) = 0$$

where $y_t^0$ is related to the observed value $y_t$ through the equation

$$y_t = y_t^0 + \epsilon_t.$$
That is, $y_t^0$ is the value of the random variable $y$ at the time of observation. The measurement error associated with the $t$th observation is $\varepsilon_t$. If for each $x_t$, and different values of $\theta$, there exists a solution $\widehat{y}(x, \theta)$ of

$$Q(x_t, \widehat{y}(x, \theta), \theta) = 0,$$

then the estimate of $\theta^*$ is that value of $\theta$ for which

$$\sum_{t=1}^{n} (y_t - \widehat{y}(x_t, \theta))^2$$

attains its minimum.

We note that under this interpretation of the implicit model the method can be extended to consider the estimator $\hat{\theta}_n$ generated by minimizing

$$\sum_{t=1}^{n} |y_t - \widehat{y}(x_t, \theta)|^p$$

for $p \geq 1$. Then under the conditions specified in Chapter 5 the derived results apply to the estimator of $\theta^*$ where,

$$y_t = \widehat{y}(x_t, \theta^*) + \varepsilon_t.$$

Britt and Luecke [3] considered this approach. Moreover, in order to extend the results to the case where both dependent and independent variables are subject to measurement error, the authors present the model as

$$g_t(y_t^0, \theta^*) = 0$$

where $t = 1, 2, \ldots, k$. 
\[ \mathbf{y}_m = \mathbf{y}_0 + \mathbf{e} \]

and \( \mathbf{y}_m \) is a \( q \)-vector of measurements on the variables subject to error, with \( q > k \).

The unknown parameters are \( \mathbf{y}_0 \) and the \( n \times 1 \) vector \( \mathbf{\theta}^* \), defined by the \( k \) model equations

\[ g_t(\mathbf{y}_0, \mathbf{\theta}^*) = 0, \]

with \( k > n \).

The error vector is assumed to be normally distributed with zero mean and a known positive definite variance covariance matrix, \( \mathbf{R} \).

Taking a maximum likelihood approach, the relevant likelihood function is

\[
L(\mathbf{y}_0, \mathbf{\theta}) = (2\pi)^{-q/2} |\mathbf{R}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y}_m - \mathbf{y}_0)^T \mathbf{R}^{-1} (\mathbf{y}_m - \mathbf{y}_0)\right)
\]

where \( \mathbf{y}_0 \) and \( \mathbf{\theta} \) are restrained to satisfy

\[ g_t(\mathbf{y}_0, \mathbf{\theta}) = 0, \]

for \( t = 1, 2, \ldots, k \).

The objective function with the added restriction becomes

\[
\phi(\mathbf{y}, \mathbf{\theta}, \lambda) = \frac{1}{2}(\mathbf{y}_m - \mathbf{y})^T \mathbf{R}^{-1} (\mathbf{y}_m - \mathbf{y}) + \lambda^T g(\mathbf{y}, \mathbf{\theta}) .
\]

Thus, the problem becomes one of estimating \( (q + n) \) unknown parameters, \( \mathbf{\theta}^* \) and \( \mathbf{y}_0 \), subject to side constraints, given the distribution of the error terms associated with the model. An algorithm is derived to solve the system of necessary conditions satisfied by the stationary points \( (\mathbf{y}_0, \mathbf{\theta}, \lambda) \).
Bard [1] takes a similar approach and also considers another interpretation of the model. Similar to Britt and Luecke, if the data is subject to measurement error the model equations may be assumed to apply exactly to the true unknown values of the variables and parameters,

$$g(x_t, y_t^0, \theta^*) = 0,$$

for \( t = 1, 2, \ldots, n \). In this case \( g \) is an \( n \times 1 \) vector and \( y_t^0 \) depends on \( t \).

An error component enters the model through

$$y_t = y_t^0 + e_t$$

where the \( m \)-vector \( y_t \) includes all variables subject to error, the vector \( x_t \) all those measured exactly. The measurement errors are assumed to have a joint probability density function \( p((y_t - y_t^0) | \psi) \) indexed by the parameters \( \psi \).

Alternatively, the model may be viewed as applying only approximately even to the true values \( y_t^0 \). The model itself is inexact,

$$g(x_t, y_t^0, \theta^*) + e_t = 0$$

or

$$g(x_t, y_t, \theta^*) + e_t = 0$$

depending on whether the model applies to the true unknown values \( y_t^0 \) or the measured values \( y_t \). The random variable \( e_t \) is usually
assumed to have zero expectation, a symmetric distribution, and to represent the effects of non-specified variables.

If $Y_t$ is measured with error the joint probability function for the $t^{th}$ observation is

$$p((Y_t - Y_t^0)|\psi) \cdot p(\varepsilon_t|\psi')$$

where $p(\varepsilon_t|\psi')$ is the joint density function of the error vector $\varepsilon_t$ indexed by the parameters $\psi'$, and the errors $(Y_t - Y_t^0)$ are assumed to be statistically independent of $\varepsilon_t$ for each $t = 1, 2, \ldots, n$.

Under no measurement error the relevant density is simply $p(\varepsilon_t|\psi')$.

Bard also takes a maximum likelihood approach to estimation. If the model is exact and describes the relationship among $X_t$, $Y_t^0$, and $\varepsilon^*$, the likelihood is derived solely from the distribution of measurement errors

$$L((Y_t - Y_t^0)|\psi) = \prod_{t=1}^{n} p((Y_t - Y_t^0)|\psi)$$

restricted to those points $(Y_t^0, \theta)$ satisfying

$$\varepsilon(X_t, Y_t^0, \theta) = 0.$$

The relevant likelihood, under an inexact model applying to the measured values $Y_t$ is,

$$L((Y_t - Y_t^0, \theta, \psi, \psi') = \prod_{t=1}^{n} p((Y_t - Y_t^0)|\psi)$$

$$\times p(\varepsilon(X_t, Y_t^0, \theta)|\psi').$$
Alternatively, since
\[
\prod_{t=1}^{n} p(\varepsilon_t | \psi') = \prod_{t=1}^{n} p(\varepsilon_t, \mathcal{Y}_t, \theta | \psi'),
\]
if the model applies to the measured values $\mathcal{Y}_t$ and these are observed without error, the likelihood expressed in terms of the derived density for the observed data is
\[
L(\theta, \psi') = \prod_{t=1}^{n} p(\varepsilon_t, \mathcal{Y}_t, \theta | \psi') |\det \left( \frac{\partial \mathcal{E}_t}{\partial \psi_t} \right)|.
\]
Usually the distribution function assumed for the errors in the model is the multivariate normal.

A second approach to the implicit model situation is presented by Salmond [17]. The problem arose in which a mathematical model proposed did not supply enough prior information to define uniquely the parameters indexing the process.

The model postulated is a single equation model given by
\[
l = X_j ' \beta + \epsilon_j
\]
where
- $X_j'$ is a multivariate observation associated with sample unit $j$,
- $\beta$ is the unknown vector of weighting coefficients to be estimated,
- $X_j ' \beta$ is the index associated with sample unit $j$,
- $\epsilon_j$ is a random error.
The vector $X_j$ was considered a multivariate random variable, correlated with the error term, and additional information was not sufficient to postulate a probability distribution for $e_t$. In this case the vector of parameters is not defined by the model; additional restrictions on $\beta$ were added to accomplish this. For each unit $j$ it was "sensibly" implied from the process that at the desired parameter vector $\beta$, $X_j\beta$ was close to 1 in some sense. Hence, $\beta$ was defined as the value $\beta = (b_1, b_2, \ldots, b_n)$ which minimizes

$$E(1 - X_j'\beta)^2$$

subject to

$$E(1 - X_j'\beta) = 0.$$  

Under the assumption that $X$ had an $n$-multivariate normal distribution with mean $\mu$ and positive definite matrix of dispersion $\Gamma$, the constrained minimization problem led to

$$\beta = \Gamma^{-1}\mu'\mu^{-1}u.$$  

Hence, $\beta = (b_1, b_2, \ldots, b_n)$ was defined by the above expression.

The derived maximum likelihood estimates of $\mu$ and $\Gamma$ were

$$\hat{\mu} = \frac{1}{N} \sum_{j=1}^{n} X_{i,j} = \bar{X}_j$$

$$\hat{\gamma}_{kl} = \frac{1}{N} \sum_{j=1}^{n} (X_{kj} - \bar{X}_k)(X_{lj} - \bar{X}_l)$$

for $1 \leq k \leq n$; $1 \leq k \leq n$.  

Hence,

\[ \hat{\beta} = \hat{F}^{-1}u \hat{F}^{-1}u . \]

Under more general conditions for \( X \), that is, a non-specified continuous distribution, and assuming that the first four cummulants of the distribution existed, the author concluded that the estimators of \( u \) and \( \Gamma \) correspond to those derived with the multivariate normal distribution assumption. In either case, the estimator \( \hat{\beta} \) was shown to be asymptotically unbiased, given a degree approximation of \( \hat{\beta} \), and an estimator of the asymptotic variance covariance matrix of \( \hat{\beta} \) was derived.
2. THE MODEL AND GENERAL ASSUMPTIONS

2.1 Introduction

The implicit model was introduced in Chapter 1 as a mathematical representation of a process, given by

\[ Q(x_t, y_t, \theta^*) = \varepsilon_t. \]

Given the model formulation and a sample of points \( \{x_t, y_t\}_{t=1}^n \), the statistical problem is to estimate \( \theta^* \), and to derive the properties of the estimator. To this end, specific restrictions are imposed on the function \( Q \), its arguments, and the error term \( \varepsilon_t \). In the present discussion the implicit model will consist of the structural relation \( Q \) and the conditions imposed on it and on the error term \( \varepsilon_t \).

The following sections introduce the model formally, establish the notation, and present the major assumptions.

2.2 Structure and Model Specifications

The structural model is one equation of the form

\[ Q(x_t, y_t, \theta^*) = \varepsilon_t \]

where \( \varepsilon_t \) is an unobservable random variable satisfying

\[ E(\varepsilon_t) = 0 \quad 0 < E(\varepsilon_t^2) < \infty \]

for \( t = 1, 2, \ldots \). Formally, \( \varepsilon_t = \varepsilon_t(\lambda) \), is a random variable on \((\Lambda, \mathcal{G}, \mathbb{P})\) into \( \mathbb{R} \). The vector \( \{\varepsilon_1(\lambda), \varepsilon_2(\lambda), \ldots\} \) denotes a
realization of independent and identically distributed error terms
where each $\varepsilon_t$ has probability function $\rho$. The unknown parameter
to be estimated is $\theta^*$, a $k \times 1$ vector in $\Omega \subset \mathbb{R}^k$.

As in the conventional linear regression model the choice must be
made whether to view $\{x_t\}_{t=1}^{\infty}$ as a sequence of known constants or as
a realization of random variables. The theoretical results derived in
later chapters do not depend on which of the two interpretations one
takes, as long as limits of functions of the form

$$\frac{1}{n} \sum_{t=1}^{n} g(x_t, y_t, \theta)$$

and

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} g(x_t, y_t, \theta)$$

behave properly, (when $\theta$ is a point in $\Omega$). Assumptions 14 and 15
in Section 2 of this chapter explain what we mean by the above comment.

Hence, we will assume that $\{x_t\}_{t=1}^{\infty}$ is generated in one of two
ways:

1. The vectors $x_1, x_2, \ldots$ are a sequence of random
variables mapping $(\Lambda, \mathcal{G})$ into $\mathbb{R}^m$.

2. The $m$-dimensional vectors $x_1, x_2, \ldots$ are constants
specified by the experimenter.

Often, there exists a real valued function $h$, such that $y_t$ is
a random variable on $(\Lambda, \mathcal{G})$ satisfying,

$$y_t(\lambda) = h(x_t, \varepsilon_t(\lambda), \theta^*)$$
However, by specifying the model in its implicit form

\[ Q(x_t, y_t, \theta^*) = \epsilon_t \]

we include the case when \( h \) cannot be formulated explicitly or does not exist uniquely. Standard regression methods require data on the response \( y_t \) and a predictor \( \hat{y}_t \) for different values of \( \theta \) and each \( t = 1, 2, \ldots, n \). Hence, the derivation of the model is most often approached formulating a response function in terms of independent variables and parameters. For the purpose of estimation, the function \( h \) will not be assumed to exist analytically or numerically. When considering the prediction of \( y \) it will be required that if the true value \( \theta^* \) and the error component \( \epsilon_t \) are known, then for any value of \( x \) in \( x \in R^m \) a solution for \( y \) can be found using standard methods of numerical analysis. Note that if \( h \) is known and is of the form

\[ h(x_t, \epsilon_t, \theta^*) = f(x_t, \theta^*) + \epsilon_t \]

where \( f: X \times \Omega \rightarrow R \) is a linear or nonlinear function of the parameters in \( \theta^* \), and the \( x_t \) vector is uncorrelated with \( \epsilon_t \), then we have the regression case,

\[ y_t - f(x_t, \theta^*) = \epsilon_t\]

Furthermore, as indicated in Chapter 1, it is assumed

(a) that the structural model \( Q(x_t, y_t, \theta^*) = \epsilon_t \) is satisfied by the measured values \( \{x_t, y_t\}_{t=1}^{n} \), and
(b) that the error term $e_t$ reflects the influence of unknown variables not individually identified by the structure.

Moreover, Assumptions 2, 4, and 5 in Section 2 of this chapter impose regularity conditions on the function $Q$ mapping $S$ into $R$ where $S$ is a Borel subset of $\mathbb{R}^{m+1}$.

As noted previously the inputs $\{x_t\}_{t=1}^{\infty}$ are either stochastic or a sequence of constants. For each case, additional requirements are as follows:

1. The vectors $x_1, x_2, \ldots$ are random variables and $\{(x_1, y_1), (x_2, y_2), \ldots\}$ denotes independent and identically distributed random vectors $(x_t, y_t)$ each with probability function $P_{e^*}$ on $(S, \mathcal{B})$, where $\mathcal{B}$ is the collection of Borel subsets of $S$. As we will see, in this case, the Strong Law of Large Numbers and the Multivariate Central Limit Theorem hold and imply Assumption 14 and 15. We denote the conditional distribution function of $y_t$ given $x_t = x$ by $F_{y, \theta^*|x}$ for $t = 1, 2, \ldots$ and the conditional expectation of a $\mathcal{B}$ measurable function $g$ on $S \times \Omega$ by

$$E(g|x) = \int_{\mathbb{R}} g(x, y, \theta) dF_{y, \theta^*|x}.$$
Expectations of functions on $S \times \Omega$ will be taken with respect to the joint measure $P_{\theta^*}$. A formal definition of the function $E(g/x)$ appears in Loeve [13, p. 341].

2. For each $x_t$ in the sequence of constants
   \[ \{x_t\}_{t=1}^{\infty}, \]
   $y_t$ is a random variable on $(\Lambda, \mathcal{G})$, and there may exist a real valued function $h$ such that,
   \[ y_t(\lambda) = h(x_t, \epsilon_t(\lambda), \theta^*). \]
   If $h$ exists, its explicit form may or may not be known. The distribution function of $y_t$ for
   $x_t = x$ is $F_y(x, \theta^*)$ for $t = 1, 2, \ldots, n$. Furthermore, the sequence $\{x_t\}_{t=1}^{\infty}$ is generated such that for each $n$ there is a probability measure $u_n(A)$ defined on $(X, \mathcal{G}_X)$, where $X$ is a Borel subset of $\mathbb{R}^m$ and $\Lambda \in \mathcal{G}_X$, the collection of Borel subsets of $X$. In addition, the measure $u_n(A)$ converges weakly to a probability measure $u_X(A)$ for all $A$ in $\mathcal{G}_X$.

Definition 2.1

Let $\{u_n(A)\}_{n=1}^\infty$ be a sequence of probability measures on $(X, \mathcal{G}_X)$ and let $u_X(A)$ be a probability measure on $(X, \mathcal{G}_X)$, then

\[ u_n(A) \xrightarrow{\text{weakly}} u_X(A) \]
if for every bounded continuous function \( f(x) \)

\[
\int_X f(x) d\mu_n(x) \to \int_X f(x) d\mu(x)
\]

as \( n \to \infty \).

Let the probability measure on \((S, \mathcal{B})\) be

\[
P_{\theta^*}(E) = \int_X \left( \int_{E_x} dF_y(x, \theta^*) \right) d\mu_X,
\]

where \( E \in \mathcal{B} \) and \( E_x = \{ y \in R \mid (x, y) \in E \} \) is a measurable set. Then, for \( g(x, y, \theta) \) a \( \mathcal{B} \)-measurable function on \( S \times \Omega \),

\[
E_{\theta^*} g(x, y, \theta) = \int_S g(x, y, \theta) dP_{\theta^*}
\]

\[
= \int_R \left( \int_R g(x, y, \theta) dF_y(x, \theta^*) \right) d\mu_X
\]

whenever the expectation exists with respect to \( P_{\theta^*} \).

For both interpretations of the sequence \( \{x_t\}_{t=1}^\infty \) we will use the following notation throughout:

\[
E_{\theta^*} g(x, y, \theta) = \int_R E(g(x, y, \theta) \mid x) dP_x, \quad \theta \in \Omega
\]

and

\[
E(g(x, y, \theta) \mid x) = \begin{cases} 
\int_R g(x, y, \theta) dF_y(x, \theta^* \mid x) \ x_t \text{ stochastic} \\
\int_R g(x, y, \theta) dF_y(x, \theta^*) \ x_t \text{ non-stochastic}
\end{cases}
\]
2.3 The General Assumptions

For convenience, all assumptions are listed in this section. First, the basic notation used throughout is established. This is,

\[ \theta = (\theta_1, \theta_2, \ldots, \theta_k) \in \Omega \subset \mathbb{R}^k, \]

\[ x^T = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m, \]

\[ \Omega^T(\theta) = (Q(x_1, y_1, \theta), Q(x_2, y_2, \theta), \ldots, Q(x_n, y_n, \theta)), \]

and

\[ \Omega^T(\theta) = (\rho(x_1, y_1, \theta), \rho(x_2, y_2, \theta), \ldots, \rho(x_n, y_n, \theta)). \]

Let \( Q_t(\theta) \) and \( \rho_t(\theta) \) denote the \( t^{th} \) term in the vectors \( \Omega^T(\theta) \) and \( \Omega^T(\theta) \) respectively and \( \rho_t(\theta) = |Q_t(\theta)|^p \).

\[ \Psi_n(\theta) = \sum_{t=1}^{n} |Q(x_t, y_t, \theta)|^p = \sum_{t=1}^{n} |Q_t(\theta)|^p. \]

\( D_n(\theta) \) is an \( n \times k \) matrix with \( i^{th} \) row given by

\[ \left( \frac{\partial Q_1(\theta)}{\partial \theta_1}, \frac{\partial Q_1(\theta)}{\partial \theta_2}, \ldots, \frac{\partial Q_k(\theta)}{\partial \theta_k} \right). \]

The rank of \( D_n(\theta) \) is \( k \).

\[ W_n(\theta) = \text{diag}(|Q_1(\theta)|^{p-2}, |Q_2(\theta)|^{p-2}, \ldots, |Q_n(\theta)|^{p-2}). \]

\( G_n(\theta) \) is a \( k \times k \) matrix with \( i^{th} \) row given by

\[ \left( \sum_{t=1}^{n} \frac{\partial^2 \rho_t(\theta)}{\partial \theta_1 \partial \theta_1}, \sum_{t=1}^{n} \frac{\partial^2 \rho_t(\theta)}{\partial \theta_1 \partial \theta_2}, \ldots, \sum_{t=1}^{n} \frac{\partial^2 \rho_t(\theta)}{\partial \theta_1 \partial \theta_k} \right). \]
\[ G(\theta) = a \ k \times \ k \ matrix \ with \ \text{ith} \ row \ given \ by \]
\[
\begin{pmatrix}
\frac{\partial^2 \rho(x,y,\theta)}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 \rho(x,y,\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \rho(x,y,\theta)}{\partial \theta_1 \partial \theta_k}
\end{pmatrix}.
\]

\[ \Lambda_n(\theta) = a \ k \times \ k \ matrix \ with \ \text{ith} \ row \ given \ by \]
\[
\begin{pmatrix}
\Sigma_{t=1}^{n} \frac{\partial \rho_t(\theta)}{\partial \theta_1} & \Sigma_{t=1}^{n} \frac{\partial \rho_t(\theta)}{\partial \theta_2} & \cdots & \Sigma_{t=1}^{n} \frac{\partial \rho_t(\theta)}{\partial \theta_k}
\end{pmatrix}.
\]

\[ \Lambda(\theta) = a \ k \times \ k \ matrix \ with \ \text{ith} \ row \ given \ by \]
\[
\begin{pmatrix}
\frac{\partial \rho(x,y,\theta)}{\partial \theta_1} & \frac{\partial \rho(x,y,\theta)}{\partial \theta_2} & \cdots & \frac{\partial \rho(x,y,\theta)}{\partial \theta_k}
\end{pmatrix}.
\]

\[ (D^T WD)_n(\theta) = \text{the product} \]
\[ D^T_n(\theta) W_n(\theta) D_n(\theta) \]

\[ (D^T W Q)_n(\theta) = \text{the product} \]
\[ D^T_n(\theta) W_n(\theta) Q(\theta) \]

\[ A(\theta, \delta) = \text{a spherical neighborhood with radius} \ \delta > 0 \ \text{and centered at} \]

\[ \theta \text{ and } \delta \] is the termination of a lemma, proposition or theorem.

The Assumptions on which this thesis is based appear below.

Included in their description are explanatory comments to identify the
major results associated with a particular assumption or a group of
them. These are:
1. \( \Omega \) is a compact subset in \( \mathbb{R}^k \).

2. \( Q(x,y,\theta) \) is a real valued function on \( S \times \Omega \), continuous in \( \theta \) for \( (x,y) \) in \( S \), \( \mathcal{B} \) measurable for each \( \theta \) in \( \Omega \).

3. For \( \theta_0 \) as defined in Assumption 7

\[
P_{\theta_0}(\{x,y\} | Q(x,y,\theta_0) = 0) = 0.
\]

4. \( \partial Q(x,y,\theta)/\partial \theta_i \) exists on \( S \times \Omega \), and is real valued and continuous in \( \theta \) for \( (x,y) \) in \( S \), and \( \mathcal{B} \) measurable for each \( \theta \) in \( \Omega \) for all \( i = 1, 2, \ldots, k \).

5. \( \partial^2 Q(x,y,\theta)/\partial \theta_i \partial \theta_j \) exists on \( S \times \Omega \), and is real valued continuous in \( \theta \) for \( (x,y) \) in \( S \) and \( \mathcal{B} \) measurable for each \( \theta \) in \( \Omega \) for all \( i,j = 1, 2, \ldots, k \).

6. There exists an open set \( V \) in the interior of \( \Omega \) such that

\[
\frac{\partial \psi_n(\theta')}{\partial \theta} = \frac{\partial \psi_n(\theta'')}{\partial \theta} = 0
\]

implies \( \theta' = \theta'' \).

6. (a) There exists a \( \theta_* \) in \( V \) of Assumption 6 such that
\[ \psi_n(\theta^*_x) < \inf \psi_n(\theta), \]

for \( \theta \) any boundary point in \( V \).

6. (b) For each natural number \( n \), the value \( \theta^*_n \) in \( \Omega \) minimizing \( \psi_n(\theta) \) is in the interior of \( \Omega \).

Assumptions 1, 2, 4, 6, and 6(a) are fundamental in proving convergence of the computing algorithm generating the estimator \( \hat{\theta}_n \).

The derivation of this algorithm is based on Assumptions 4 and 6(b).

7. There exists a unique value \( \theta_0 \) in \( \text{int}(\Omega) \) such that

\[ E_{\theta^*_0} p(x,y,\theta_0) < E_{\theta^*_0} p(x,y,\theta) \]

for all \( \theta \neq \theta_0 \) and \( \theta \) in \( \Omega \).

8. \( E_{\theta^*_0} p(x,y,\theta) < \infty \) for all \( \theta \) in \( \Omega \).

Assumptions 1, 2, 7, 8 and 14 are sufficient for showing that the estimator \( \hat{\theta}_n \) converges to \( \theta_0 \) with probability one as \( n \to \infty \).

9. \( E_{\theta^*_0} \left( \frac{\partial p(x,y,\theta)}{\partial \theta} \right) \) is well defined for \( \theta \) in \( \Omega \).

10. \( E_{\theta^*_0} \left| \frac{\partial^2 p(x,y,\theta)}{\partial \theta_j \partial \theta_i} \right|_{\theta=\theta_0} < \infty \) for \( i,j = 1, 2, \ldots, k \) and \( \theta_0 \) as in Assumption 7.

11. \( E_{\theta^*_0} \left| \frac{\partial p(x,y,\theta)}{\partial \theta_i} \frac{\partial p(x,y,\theta)}{\partial \theta_j} \right| < \infty \) and \( E_{\theta^*_0} \left( \frac{\partial p(x,y,\theta)}{\partial \theta_i} \right)^2 > 0 \)

for \( i,j = 1, 2, \ldots, k \) and \( \theta \) in \( \Omega \).
12. Det(G(θ₀)) ≠ 0 where G(θ₀) is the k x k matrix with (i,j) element

\[ \frac{\partial^2 \rho(x,y,θ₀)}{\partial \theta_i \partial \theta_j} \]

13. The matrix (D'WD)(θ) is positive definite.

14. Let g(x,y,θ) be a real valued measurable function for each θ in Ω such that

\[ \mathbb{E}_{θ^*}|g(x,y,θ)| < \infty, \]

then

\[ \frac{1}{n} \sum_{t=1}^{n} g(x_t,y_t,θ) \rightarrow \mathbb{E}_{θ^*}g(x,y,θ) \]

with probability one as n → ∞.

15. Let g(x,y,θ) be a k x 1 vector such that g_i(x,y,θ) is real valued and measurable for each θ in Ω and

\[ 0 < \mathbb{E}_{θ^*}g_i^2(x,y,θ) < \infty, \]

for all i = 1, 2, ..., k and each θ in Ω then

\[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (g(x_t,y_t,θ) - \mathbb{E}_{θ^*}g(x,y,θ)) \xrightarrow{L} N_k(0,Σ) \]

where Σ is the (k x k) matrix with (i,j) element
The derivation of Theorems 4.3 and 4.4 concerning the asymptotic normality of the sequence of estimators \( \hat{\theta}_n \) is based on Assumptions 2, 4, 5, 7, 10, 11, 12, 14, 15 and relations 4.3 and 4.4 respectively.

The latter are presented with the theorems.

The following assumptions relate specifically to the special case when \( Q(x, y, \theta^*) = y - f(x, \theta^*) = t \).

For each \( t = 1, 2, \ldots \), the distribution of \( \varepsilon_t \) for a given vector \( x = x \) is symmetric about zero with a unique median.

Let \( C(\theta) = \{ x | (f(x, \theta) - f(x, \theta^*)) \neq 0 \} \) for \( \theta \) in \( \Omega \), \( \theta \neq \theta^* \), then \( u(x, C(\theta)) > 0 \).

\( \Omega \) is a compact connected subset of \( \mathbb{R}^k \) and \( \theta^* \) is an interior point of \( \Omega \).

The error \( \varepsilon_t \) has a bounded density function with respect to the Lebesque Measure.

There exists a \( P_x \) integrable function \( r_{i,j}(x) \) for each \( i, j = 1, 2, \ldots, k \) such that for every \( \theta \) in \( \Omega \)

\[
| \frac{\partial f(x, \theta)}{\partial \theta_i} \cdot \frac{\partial f(x, \theta)}{\partial \theta_j} | \leq r_{i,j}(x).
\]
20. (a) There exists a constant $T$ such that

$$\sup_{X \in A(\theta^*, \delta)} \left( \sup_{X} \left| f(x, \theta) - f(x, \theta^*) \right| \right) \leq T$$

for all $0 < \delta < \varepsilon$ and some $\varepsilon > 0$.

20. (b) There exists a $P_X$ integrable function $S(x)$ such that

$$\left| \frac{\partial^2 f(x, \theta)}{\partial \theta_i \partial \theta_j} \right| \leq S(x)$$

for $i, j = 1, 2, \ldots, k$.

20. (c) There exists a $P_X$ integrable function $r_{i,j,r}^{'}(x)$ for each $i, j, r = 1, 2, \ldots, k$ such that for every $\theta$ in $\Omega$

$$\left| \frac{\partial f(x, \theta)}{\partial \theta_i} \frac{\partial f(x, \theta)}{\partial \theta_j} \frac{\partial f(x, \theta)}{\partial \theta_r} \right| \leq r_{i,j,r}^{'}(x).$$

Note that if the set $X$ in $R^m$ is also compact, Assumptions 20, 20(a), 20(b), and 20(c) are satisfied.
3. THE ESTIMATOR

3.1 Definition of The Estimator \( \hat{\theta}_n \)

This chapter introduces the estimator \( \hat{\theta}_n \), and develops an iterative computing scheme to generate \( \hat{\theta}_n \) given the observed sample \( \{(x_t^t, y_t^t)\}_{t=1}^n \) for each \( n \geq k \), where \( k \) is the number of unknown parameters.

Consider the implicit model

\[
Q(x_t^t, y_t^t, \theta^*) = \epsilon_t, 
\]

where the structure \( Q \), the random error \( \epsilon_t \), and the arguments \( (x_t^t, y_t^t, \theta^*) \) follow the model specifications presented in Chapter 2.

Given a choice of \( p \), \( \hat{\theta}_n \) will be defined as a measurable function of \( \{(x_t^t, y_t^t)\}_{t=1}^n \) into \( \Omega \) such that,

\[
\min_{\theta \in \Omega} \psi_n(\theta) = \min_{\theta \in \Omega} \sum_{t=1}^n |Q(x_t^t, y_t^t, \theta)|^p = \sum_{t=1}^n |Q(x_t^t, y_t^t, \hat{\theta}_n)|^p.
\]

Since \( \psi_n(\theta) \) is bounded below by zero and \( \Omega \) is compact, there exists at least one value \( \theta' \) in \( \Omega \) such that

\[
\psi_n(\theta') = \min_{\theta \in \Omega} \psi_n(\theta).
\]

Moreover, we will show that there exists a measurable version \( \theta'_n \) of \( \theta' \) satisfying

\[
\psi_n(\theta'_n) = \psi_n(\theta').
\]
Lemma 3.1

Let $Q$ be a real valued function on $\Theta \times \Psi$ where $\Theta$ is a compact subset of a Euclidian Space and $\Psi$ is a measurable space. For $\theta$ in $\Theta$, let $Q(\theta, y)$ be a measurable function of $y$, and for each $y$ in $\Psi$, a continuous function of $\theta$. Then, there exists a measurable function $\hat{\theta}$ from $\Psi$ into $\Theta$ such that for all $y$ in $\Psi$

$$Q(\hat{\theta}(y), y) = \inf_{\theta} Q(\theta, y).$$

Proof

Lemma 3.1 and its proof appear in Jennrich [12].

For each $n = 1, 2, \ldots$, let $S^n = S \times S \times \ldots \times S$, where $S$ is the domain of $Q$, $S \subseteq \mathbb{R}^{m+1}$, and let $\mathcal{B}_n$ be the collection of Borel subsets of $S^n$, then the space $(S^n, \mathcal{B}_n)$ is a measurable space.

The following theorem follows immediately from Lemma 3.1.

Theorem 3.1

Let Assumptions 1 and 2 hold, and let the implicit model be

$$Q(x_t, y_t, \theta^*) = \epsilon_t$$

then, for any $n = 1, 2, \ldots$, there exists a $\mathcal{B}_n$ measurable function $\hat{\theta}_n$ with arguments $\{(x_t, y_t)\}_{t=1}^n$ and range in $\Omega$ such that,

$$\sum_{t=1}^n |Q(x_t, y_t, \hat{\theta}_n)|^p = \min_{\theta \in \Omega} \sum_{t=1}^n |Q(x_t, y_t, \theta)|^p.$$
Proof

$\Omega$ is a compact subset of $\mathbb{R}^k$ by Assumption 1. Assumption 2 implies $\Psi_n(\theta)$ is continuous in $\theta$ for each $\{(x_t, y_t)\}_{t=1}^n$ in $\mathbb{S}^n$ and $\Theta_n$ measurable for each $\theta$ in $\Omega$. Hence, by Lemma 3.1 the conclusion of the theorem follows.

Given the sample values $\{(x_t, y_t)\}_{t=1}^n$ from the implicit model

$$Q(x_t, y_t, \theta^*) = \epsilon_t,$$

the estimator $\hat{\theta}_n$ of $\theta^*$ is defined as a $\Theta_n$ measurable function on $\mathbb{S}^n$ into $\Omega$ such that

$$\min_{\theta \in \Omega} \sum_{t=1}^n |Q(x_t, y_t, \theta)|^p = \sum_{t=1}^n |Q(x_t, y_t, \hat{\theta}_n)|^p = \Psi_n(\hat{\theta}_n).$$

3.2 Computation of $\hat{\theta}_n$

This section proposes an iterative method which generates a sequence

$$\{\theta_n\}_{r=1}^\infty,$$

for a given initial guess of the true parameter value $\theta^*$ and a given observed sample $\{(x_t, y_t)\}_{t=1}^n$, such that under general conditions $\{\theta_n\}_{r=1}^\infty$ converges to $\theta'$, where $\theta'$ is a stationary point of $\Psi_n(\theta)$. If $\Psi_n(\theta)$ has a unique minimum then $\theta'$ generated by the algorithm corresponds to the measurable version $\hat{\theta}_n$. This is not necessarily so if there is more than one value in $\Omega$ for which $\Psi_n(\theta)$ attains its minimum. For practical purposes the
estimate is that value $\theta^*$ given by the computing method. However, when we speak of the estimator or of $\hat{\theta}_n$ we shall assume it to be a $\theta_n$ measurable function.

Under Assumptions 4 and 6b, $\theta^*$ is also a solution to

$$\frac{\partial \Psi_n(\theta)}{\partial \theta} = 0$$

where $(\partial \Psi_n(\theta)/\partial \theta)$ is the $k \times 1$ vector of partial derivatives with respect to $\theta$, and

$$\frac{\partial \Psi_n(\theta)}{\partial \theta_i} = \sum_{t=1}^n \frac{\partial |Q(x_t, y_t, \theta)|^p}{\partial \theta_i}$$

for $i = 1, 2, \ldots, k$. For a choice of $p > 1$ then,

$$\frac{\partial |Q_t(\theta)|^p}{\partial \theta_i} = \begin{cases} 
  p |Q_t(\theta)|^{p-1} \frac{\partial Q_t(\theta)}{\partial \theta_i} & Q_t(\theta) > 0 \\
  -p (-Q_t)^{p-1} \frac{\partial Q_t(\theta)}{\partial \theta_i} & Q_t(\theta) < 0 \\
  0 & Q_t(\theta) = 0 
\end{cases}$$

for $t = 1, 2, \ldots, n$, $i = 1, 2, \ldots, k$.

The above is equivalent to

$$\frac{\partial |Q_t(\theta)|^p}{\partial \theta_i} = \begin{cases} 
  p |Q_t(\theta)|^{p-1} Q_t(\theta) \frac{\partial Q_t(\theta)}{\partial \theta_i} & Q_t(\theta) \neq 0 \\
  0 & Q_t(\theta) = 0 
\end{cases}$$

for $t = 1, 2, \ldots, n$, $i = 1, 2, \ldots, k$. The estimate $\theta^*$ must be a solution to
Underlying the derivation of the algorithm is the existence of the first partials of \( \Psi_n(\theta) \) with respect to \( \theta \). For this reason, the resulting system of equations is not appropriate when \( p = 1 \) since in this case the first partials of \( \Psi_n(\theta) \) may not exist at those points \((x, y, \theta)\) for which \( Q(x, y, \theta) = 0 \). However, note that the estimator is defined for any finite \( p \geq 1 \) and any of the known methods for minimizing the sum of absolute deviations can be utilized in generating \( \theta' \). Parenthetically, it should be noted that we applied the algorithm to several examples for which \( p = 1 \) and the value of \( \theta' \) generated coincided with known values obtained from other computing techniques. See [19].

In matrix notation, the system of equations generating \( \theta_n^r \) is

\[
\frac{\partial \Psi_n(\theta)}{\partial \theta} = p (D'WQ)_n(\theta_n^r) = 0
\]

where as specified in Section 2, Chapter 2, the \( i^{th} \) element of the \((k \times 1)\) vector \((D'WQ)_n(\theta)\) is

\[
((D'WQ)_n(\theta))_i = \sum_{t=1}^{n} \frac{\partial Q_t(\theta)}{\partial \theta_i} |Q_t(\theta)|^{p-2} Q_t(\theta).
\]

Evaluating the product \( \frac{\partial Q_t(\theta)}{\partial \theta_i} \cdot |Q_t(\theta)|^{p-2} \) in the above expression at the current value of \( \theta = \theta_n^r \), and substituting the last factor \( Q_t(\theta) \) by a first order Taylor's approximation around \( \theta_n^r \), the expression for \((D'WQ)_n(\theta))_i\) becomes
\[
\frac{\partial \Psi_n(\theta)}{\partial \theta_i} = p \sum_{t=1}^{n} \frac{\partial Q_t(\theta_r)}{\partial \theta_i} |Q_t(\theta_r)|^{p-2} Q_t(\theta_r) + \sum_{j=1}^{k} \frac{\partial Q_t(\theta_r)}{\partial \theta_j} (\theta_j - \theta_r)
\]

\[
= p \sum_{t=1}^{n} \frac{\partial Q_t(\theta_r)}{\partial \theta_i} |Q_t(\theta_r)|^{p-2} Q_t(\theta_r) + p \sum_{t=1}^{n} \frac{\partial Q_t(\theta_r)}{\partial \theta_i} |Q_t(\theta_r)|^{p-2} \left( \sum_{j=1}^{k} \frac{\partial Q_t(\theta_r)}{\partial \theta_j} (\theta_j - \theta_r) \right)
\]

for \( i = 1, 2, \ldots, k \).

In matrix notation,

\[
\left( \frac{\partial \Psi_n(\theta)}{\partial \theta} \right) = p(D'WQ)_n(\theta_r) + p(D'WD)_n(\theta_r)(\theta - \theta_r)).
\]

The matrix \((D'WD)_n(\theta)\) has \((i,j)\) element

\[
[(D'WD)_n(\theta)]_{i,j} = \sum_{t=1}^{n} \frac{\partial Q_t(\theta)}{\partial \theta_i} |Q_t(\theta)|^{p-2} \frac{Q_t(\theta)}{\partial \theta_j}
\]

for \( i,j = 1, 2, \ldots, k \). The matrices \( D_n \) and \( W_n \) and the vector \( Q_n \) are as described in Section 2 of Chapter 2.

Thus, the value of \( \theta \) at the \((r+1)\) iteration is \( \theta_{r+1} \), where

\[
\left( \frac{\partial \Psi_n(\theta_{r+1})}{\partial \theta} \right) = p(D'WQ)_n(\theta_r) + (D'WD)_n(\theta)[(\theta_{r+1} - \theta_r)] = 0.
\]

Equivalently,

\[
(\theta_{r+1} - \theta_r) = -(D'WD)_n^{-1}(\theta_r)(D'WQ)_n(\theta_r).
\]

In the language of iterative techniques, the algorithm is in the class of gradient methods with a positive definite direction matrix \((D'WD)_n^{-1}\) and direction of search given by \(-(D'WD)_n^{-1}(\theta_r)(D'WQ)_n(\theta_r)\).
However, to insure that the value of the objective function $\Psi_n(\theta)$ is improved at each step we will choose $\theta_{r+1}$ to satisfy

$$
\theta_{r+1} = \theta_r - \lambda_r (D'WD)^{-1}_n(\theta_r)D'WQ(\theta_r)
$$

where $\lambda_r$ is in $[0,1+\varepsilon]$ such that

$$
\Psi_n(\theta_{r+1}) \leq \Psi_n(\theta).
$$

Under the conditions of Lemma 3.2 such a $\lambda_r$ exists for each $r = 1, 2, \ldots$ whenever $(D'WD)^{-1}_n(\theta_r)$ exists. Moreover, if

$$
\frac{\partial \Psi_n(\theta_r)}{\partial \theta} \neq 0
$$

and the direction matrix $(D'WD)^{-1}_n(\theta_r)$ is positive definite then Lemma 3.2 concludes that for some $\lambda_r > 0$, $\theta_{r+1}$ as defined above exists in the int($\Omega$) such that

$$
\Psi_n(\theta_{r+1}) < \Psi_n(\theta_r).
$$

**Lemma 3.2**

Let $\bar{\theta}$ be an interior point in $\Omega \subseteq R^k$, let Assumptions 4 and 13 be satisfied at $\theta = \bar{\theta}$, and let $(Q'WD)(\bar{\theta}) \neq \mathbf{0}$, then there exists a value $\lambda_0$ such that

$$
\Psi_n(\bar{\theta} - \lambda(D'WD)^{-1}_nD'WQ(\bar{\theta})) < \Psi_n(\bar{\theta})
$$

for all $\lambda$ in $(0,\lambda_0)$.

**Proof:**

By Assumption 4 $\Psi_n(\theta)$ is differentiable for all $\theta$ in the interior of $\Omega$ with derivative at $\theta = \bar{\theta}$ given by $p(Q'WD)_n(\bar{\theta})$. 

From the definition of derivative of a function on $\Omega \subset \mathbb{R}^k \to \mathbb{R}^r$, and by Assumption 13, there exists a value $\lambda_0(\varepsilon)$ for any $\varepsilon > 0$, such that

$$\frac{1}{\lambda} [\Psi_n(\tilde{\theta}) - \lambda (D'WD)^{-1}_n (D'WQ)_n(\tilde{\theta})) - \Psi_n(\tilde{\theta})]$$

$$+ p(Q'WD)_n (D'WD)^{-1}_n (D'WQ)_n(\tilde{\theta})$$

$$< \varepsilon \| (D'WD)^{-1}_n (D'WQ)_n(\tilde{\theta}) \|$$

for all $\lambda \in (0, \lambda_0(\varepsilon))$ and where $\| \|$ denotes the Euclidian norm.

Note that the matrix $(D'WD)_n(\theta)$ is at least positive semi-definite since $W_n(\theta)$ is a diagonal matrix with elements $[W_n(\theta)]_{t,t} = |Q_t(\theta)|^{p-2}$, for $t = 1, 2, \ldots, n$.

Let

$$C(\tilde{\theta}) = \frac{p(Q'WD)_n (\tilde{\theta}) (D'WD)^{-1}_n (\tilde{\theta}) (D'WQ)_n(\tilde{\theta})}{\| (D'WD)^{-1}_n (D'WQ)_n(\tilde{\theta}) \|} > 0$$

by Assumption 13, and let $\varepsilon < C(\tilde{\theta})$, then

$$\Psi_n(\tilde{\theta}) - \lambda (D'WD)_n(\tilde{\theta}) (D'WQ)_n(\tilde{\theta}) < \Psi_n(\tilde{\theta})$$

where

$$(\tilde{\theta} - \lambda (D'WD)^{-1}_n(\tilde{\theta}) (D'WQ)(\tilde{\theta})$$

is in the interior of $\Omega$ for all $\lambda$ in $(0, \lambda_0(\varepsilon))$. \[\square\]
Given the underlying assumptions up to this point, any choice of $p > 1$ is valid. We also investigated a method utilizing the knowledge on the second partials of $\psi_n(\theta)$, like the Newton-Raphson, bearing in mind that the existence and finiteness of these partials are required and hence the method will not apply always for a choice of $p$ in $(1,2)$.

Under this approach the approximation to $\psi_n(\theta)$ around $\theta_r$ is

$$\psi_n(\theta) \approx \psi_n(\theta_r) + \sum_{j=1}^{k} \frac{\partial \psi_n(\theta_r)}{\partial \theta_j} (\theta_j - \theta_{r,j}) + \frac{1}{2} (\theta - \theta_r)^T \psi''_n(\theta_r)(\theta - \theta_r).$$

Taking partial derivatives with respect to $\theta$ we obtain

$$\frac{\partial \psi_n(\theta)}{\partial \theta} = \frac{\partial \psi_n(\theta_r)}{\partial \theta} + \psi''(\theta_r)[(\theta - \theta_r)] = 0$$

where $\psi''(\theta_r)$ is a $k \times k$ matrix evaluated at $\theta = \theta_r$, with $(i,j)$ element

$$[\psi''(\theta)]_{i,j} = \frac{\partial^2 \psi_n(\theta)}{\partial \theta_i \partial \theta_j} = \sum_{t=1}^{n} \frac{\partial^2 |Q_t(\theta)|^p}{\partial \theta_i \partial \theta_j}$$

for $i,j = 1, 2, \ldots, k$.

The above expression is equivalent to
\[
\frac{\partial^2 Q_T(\theta)}{\partial \theta_i \partial \theta_j} \bigg|^{\theta = \bar{\theta}} = \begin{cases} 
(p-1)Q_T^{p-2} \frac{\partial Q_T(\theta)}{\partial \theta_i} \frac{\partial Q_T(\theta)}{\partial \theta_j} + pQ_T(\theta)\bigg|_{\theta = \bar{\theta}}^{p-1} \frac{\partial^2 Q_T(\theta)}{\partial \theta_i \partial \theta_j} \\
Q_T(\theta) > 0, \quad p > 1 \\
(p-1)|Q_T|^{p-2} \frac{\partial Q_T(\theta)}{\partial \theta_i} \frac{\partial Q_T(\theta)}{\partial \theta_j} - p|Q_T(\theta)|^{p-1} \frac{\partial^2 Q_T(\theta)}{\partial \theta_i \partial \theta_j} \\
Q_T(\theta) < 0, \quad p > 1 \\
0 \quad Q_T(\theta) = 0, \quad p \geq 2 \\
\infty \quad Q_T(\theta) = 0, \quad p \in (1,2) \\
\end{cases}
\]

Note that each of these terms becomes infinity when \( p \) is a value in \((1,2)\) and \( Q_T(\theta) = 0 \), unless \( \frac{\partial Q_T(\theta)}{\partial \theta_i} = 0 \).

Equivalently,
\[ \frac{\partial^2 Q_T(\theta)}{\partial \theta_i \partial \theta_j} = \begin{cases} 
\frac{(\partial Q_T(\theta))}{\partial \theta_1} & Q_T(\theta) \neq 0, \quad p > 1 \\
0 & Q_T(\theta) = 0, \quad p \geq 2 \\
\infty & Q_T(\theta) = 0, \quad p \in (1,2) \\
\frac{\partial^2 Q_T(\theta)}{\partial \theta_1 \partial \theta_j} & \left(\frac{\partial Q_T(\theta)}{\partial \theta_1} \right) \neq 0 \end{cases} \]

for \( i, j = 1, 2, \ldots, k \).

The matrix \( \psi''(\theta) \) is the sum of two matrices,

\[ \psi''(\theta) = p[(p-1)(D'WD)_n(\theta) + F_n(\theta)] \]

where

\[ F_n(\theta) = \begin{bmatrix} 
\sum_{t=1}^{n} |Q_T(\theta)|^{p-2} Q_T(\theta) \frac{\partial^2 Q_T(\theta)}{\partial \theta_1 \partial \theta_j} & \ldots & \sum_{t=1}^{n} |Q_T(\theta)|^{p-2} Q_T(\theta) \frac{\partial^2 Q_T(\theta)}{\partial \theta_1 \partial \theta_k} \\
\vdots & \ddots & \vdots \\
\sum_{t=1}^{n} |Q_T(\theta)|^{p-2} Q_T(\theta) \frac{\partial^2 Q_T(\theta)}{\partial \theta_1 \partial \theta_j} & \ldots & \sum_{t=1}^{n} |Q_T(\theta)|^{p-2} Q_T(\theta) \frac{\partial^2 Q_T(\theta)}{\partial \theta_1 \partial \theta_k} \\
\vdots & \ddots & \vdots \\
\sum_{t=1}^{n} |Q_T(\theta)|^{p-2} Q_T(\theta) \frac{\partial^2 Q_T(\theta)}{\partial \theta_k \partial \theta_1} & \ldots & \sum_{t=1}^{n} |Q_T(\theta)|^{p-2} Q_T(\theta) \frac{\partial^2 Q_T(\theta)}{\partial \theta_k \partial \theta_k} 
\end{bmatrix} \]

is a \( k \times k \) matrix.
Second derivative methods have been shown to provide an answer in situations where first derivative methods fail. However, in many practical cases the additional computational burden is not justified and moreover, given our class of estimators, the wider applicability to all values of $p > 1$ of the first approach is preferred. For a specific problem a Newton-Raphson method can be implemented whenever $p \geq 2$ and the matrix $\Psi''(\theta)$ of second partials is non-singular.

Using the conventional terminology we refer to the vector $-(D'WD)^{-1}(\theta_r)(D'WQ)_n(\theta_r)$ as the direction of search $P_r(\theta)$ and to $\lambda_r$ as the step length. The choice of $\lambda_r$ will guarantee that at each iterative step, the next value $\theta_{r+1}$ satisfies

1. $\theta_{r+1} = \theta_r - \lambda_r P_r(\theta)$ is in the interior of $\Omega$,

2. $\Psi_n(\theta_{r+1}) \leq \Psi_n(\theta_r)$.

The topic of choosing $\lambda_r$ has been widely discussed in the literature, specifically see Ortega and Rheinboldt [15]. These authors discuss a variety of methods to insure that conditions 1 and 2 above are satisfied for a given direction vector $P_r(\theta_r) \neq 0$. Their exposition emphasizes the considerable independence between the choice of step length and that of the direction of search. What is important is that at each iterative step the value of the objective function $\Psi_n(\theta)$ is reduced (recall that Lemma 3.2 outlines the conditions under which this reduction is made possible by a choice of $\lambda_r$). If one is interested in theoretical convergence then the choice of $\lambda_r$ must decrease $\Psi_n(\theta_r)$ by a sufficient amount. This comment will become
clear when we define the algorithm formally and prove convergence of
the generated sequence \( \{\theta_n\}_{r=1}^{\infty} \).

We start off with an initial guess \( \theta_* \) in the interior of \( \Omega \) such that

1. \( \theta_* \) is in \( V \subset \text{int}(\Omega) \),
2. \( \psi_n(\theta_*) < \inf \psi_n(\theta) \) for \( \theta \) a boundary point in \( \bar{V} \),
3. \( P_0(\theta_*) \neq 0 \).

Then at the \( r \)-th iteration define the next value of \( \theta \) denoted by
\( \theta_{r+1} \) as

\[
\theta_{r+1} = \begin{cases} 
\theta_r - \lambda_r (D'WD)^{-1} (\theta_r) (D'WQ) (\theta_r) & \rho(\theta_r) \neq 0 \\
\theta & \rho_r(\theta_r) = 0
\end{cases}
\]

where \( \lambda_r \) satisfies

\[
\psi_n(\theta_r - \lambda_r P_r(\theta_r)) = \min_{\theta \in V_r} \psi_n(\theta)
\]

and

\[
V_r = \{ \theta | \theta = \theta_r - \lambda P_r(\theta_r), \lambda \in [0,1+\varepsilon], \varepsilon > 0 \}
\]

is a closed set in \( \bar{V} \).

When we refer to the sequence \( \{\theta_n\}_{r=1}^{\infty} \), we mean that each \( \theta_r \)
was generated as described above for \( r = 1, 2, \ldots \) and that \( \theta_* = \theta_r \)
for \( r = 0 \), satisfies the conditions specified for the initial guess.
3.3 Convergence of The Algorithm

The literature is not lacking in proofs of convergence of iterative techniques that belong to the class of gradient methods with a positive definite direction matrix. For a specified direction of search and step length choice, and in the context of minimizing

$$\sum_{t=1}^{n} (y_t - f(x_t, \theta))^2,$$

Ortega and Rheinboldt [15] and Gallant [8] give proofs of convergence without requiring the existence of the second partial derivatives of $\psi_n(\theta)$ (in their case $p$ is 2). Gallant [8] chooses $\lambda_r$ such that

$$\psi_n(\theta_r - \lambda_r \rho_r) = \min_{\theta \in V_r} (\psi_n(\theta))$$

and

$$V_r = \{ \theta | \theta \text{ is in a closed convex set } V \text{ and } \theta = \theta_r - \lambda \rho_r \text{ for } \lambda \in [0,1] \}.$$  

The initial guess $\theta_0$ is in $V$.

On various examples we tried to reduce the number of iterations necessary to locate the minimum of $\psi_n(\theta)$ by considering values of $\lambda$ greater than 1. We extended the set $V$ to include these values.

Our proof of convergence is similar to Gallant's [8]. This proof makes no use of the existence of the second partials of $\psi_n(\theta)$. Recall that for $1 < p < 2$ if $Q(x_t, y_t, \theta) = 0$ for some $t = 1, 2, \ldots, n$, the second partials of $\psi_n(\theta)$ may become infinity.
Theorem 3.2

Let Assumptions 2, 4, 6, 6a, and 13 be satisfied and let
\[ \{\theta^r_n\}_{r=1}^\infty \]
be the sequence generated by the algorithm, then

1. \( \theta^r_n \) is an interior point of \( \Omega \) and \( \theta^r_n \)
in \( V \) for \( r = 1, 2, \ldots \),

2. The sequence \( \{\theta^r_n\}_{r=1}^\infty \) converges to \( \tilde{\theta} \),

3. \( \tilde{\theta} \) is in the interior of \( \Omega \) and
\[ \left( \frac{\partial \psi_n(\tilde{\theta})}{\partial \theta} \right) = 0. \]

Proof

Since \( \theta^\ast \) is an interior point in \( V \subset \Omega \), Assumptions 4 and 13
and the fact that \( P_0(\theta^\ast) \neq 0 \) imply, through Lemma 3.2, that there
is a \( \lambda' > 0 \) such that \( [\theta^\ast, \theta^\ast + \lambda' P_0(\theta^\ast)] \) is in \( V \) and
\( \psi_n(\theta) < \psi_n(\theta^\ast) \) for all \( \theta \) in \( [\theta^\ast, \theta^\ast + \lambda' P_0(\theta^\ast)] \). Recall
\[ V_0 = [\theta|\theta = \theta^\ast - \lambda P_0(\theta^\ast), \lambda \in [0,1+\varepsilon], \varepsilon > 0} \subset \tilde{V} \]
and such that \( V_0 \) is a closed set. Hence, the set
\[ [\theta^\ast, \theta^\ast + \lambda'' P_0(\theta^\ast)] \] is in \( V_0 \) for some \( \lambda'' < \min[\lambda', 1+\varepsilon] \). Since
\( V_0 \) is compact and \( \psi_n(\theta) \) is real valued continuous in \( \theta \), there
exists a \( \theta' \) in \( V_0 \) such that
\[ \psi_n(\theta') = \min_{\theta \in V_0} \psi_n(\theta). \]
By construction \( \theta' = \theta^\ast - \lambda_0 P_0(\theta^\ast) \) for some \( 0 \leq \lambda_0 \leq 1+\varepsilon \) and
\( \theta' \in \tilde{V} \). Note that
\[ \min_{V_0} \psi_n \leq \min_{[\theta^\ast, \theta^\ast + \lambda'' P_0(\theta^\ast)]} \psi_n(\theta). \]
Thus, let $\theta_1 = \theta'$. By Assumption 6a,

$$\psi_n(\theta_1) \leq \psi_n(\theta_\ast) < \inf \psi_n(\theta),$$

where $\theta$ is any boundary point in $\tilde{V}$. Hence $\theta_1$ is an interior point in $V$. Note that in the case where $P_0(\theta_\ast) = 0$, $\theta_n = \theta_\ast$ for all $r = 1, 2, \ldots$.

Let $P_1(\theta_1) \neq 0$. It was just shown that $\theta_1$ is in the interior of $V$. Substitute $\theta_\ast$ by $\theta_1$ in the preceding paragraph and similarly generate $\theta_2, \theta_3, \ldots$, where each $\theta_r$ is an interior point in $V$. Hence, conclusion 1 follows.

The sequence $\{\psi_n(\theta_r)\}_{r=1}^\infty$ is monotone non-increasing with non-negative terms. Hence, $\psi_n(\theta_r) \rightarrow \psi_\ast$ as $r \rightarrow \infty$. Since $\{\theta_r\}$ is a closed, bounded set it contains a convergent subsequence $\{\theta_{n_s}\}$, $\theta_{n_s} \rightarrow \tilde{\theta}$ as $s \rightarrow \infty$. By continuity of $\psi_n(\theta)$ we conclude

$$\psi_n(\theta_{n_s}) \rightarrow \psi_n(\tilde{\theta}) = \psi_\ast$$

as $s \rightarrow \infty$, and hence $\tilde{\theta}$ is in interior of $V$.

The function

$$P_r(\theta) = (D'W^D)^{-1}(\theta)(D'WQ)_n(\theta)$$

is continuous in $\theta$ by Assumption 2 and 4. Since $(D'W^D)_n(\theta)$ is positive definite for $\theta$ in $V$ by Assumption 13, then

$$\lim_{s \rightarrow \infty} P_s(\theta_s) = \lim_{s \rightarrow \infty} [(D'W^D)^{-1}(\theta_s)(D'WQ)_n(\theta_s)]$$

$$= (D'W^D)^{-1}(\tilde{\theta})(D'WQ)(\tilde{\theta}) = P(\tilde{\theta}).$$
Since $\bar{\theta}$ is in the interior of $\Omega$ and Assumptions 4 and 13 hold, by the conclusion of Lemma 3.2 we can find a value $\lambda_0^*$ and an $\eta^*, 0 < \eta^* < \lambda_0^* < \lambda''$, depending on the sequence $\{(x_t, y_t)\}_{t=1}^\infty$ such that

$$\psi_n(\bar{\theta} - \eta^* P(\bar{\theta})) - \psi_n(\bar{\theta}) < -\gamma < 0, \quad \gamma > 0$$

if $P(\bar{\theta}) \neq 0$.

Let $s$ be sufficiently large and let $\varepsilon < \gamma$, by continuity of $P(\theta)$ and convergence of $\theta_s$ as $s \to \infty$, there exists a $\delta(\varepsilon) > 0$, such that for $\eta^* \in (0, \lambda_0^*)$

$$\| (\theta_s - \eta^* P_s) - (\bar{\theta} - \eta^* P(\bar{\theta})) \| < \delta(\varepsilon)$$

implies

$$(\psi_n(\theta_s - \eta^* P_s) - \psi_n(\bar{\theta} - \eta^* P(\bar{\theta}))) < \varepsilon < \gamma$$

and

$$(\psi_n(\theta_s - \eta^* P_s) - \psi_n(\bar{\theta})) < \varepsilon + (-\gamma) < -c^2$$

where $c \neq 0$.

Recall that

$$\psi_n(\theta_{s+1}) = \min_{v_s} \psi_n(\theta_s - P_s) \leq \psi_n(\theta_s - \eta^* P_s)$$

Thus,

$$\psi_n(\theta_{s+1}) - \psi_n(\bar{\theta}) < -c^2$$
\[ \psi_n(\theta_{s+j}) - \psi_n(\bar{\theta}) < - c^2 \]

for all \( j > 1 \) and \( s \) sufficiently large.

This conclusion contradicts the fact previously derived that \( \psi_n(\theta_s) \to \psi_n(\bar{\theta}) \) as \( s \to \infty \) whenever \( P(\bar{\theta}) \neq 0 \), and we conclude \( P(\bar{\theta}) = 0 \). Thus, \( (D'WD)_n^{-1}(D'WQ)_n(\bar{\theta}) = 0 \), which implies

\[ (D'WQ)_n(\bar{\theta}) = \frac{\partial \psi_n(\bar{\theta})}{\partial \theta} = 0. \]

Thus we have that

\[ \lim_{r \to \infty} \psi_n(\theta_r) = \psi_n(\bar{\theta}) \]

where \( \bar{\theta} \) is a stationary point of \( \psi_n(\theta) \). By Assumption 6a, we will now show that the sequence \( \{\theta_r\}_{r=1}^\infty \) also converges to this same \( \bar{\theta} \).

Consider any convergent subsequence of \( \{\theta_r\}_{r=1}^\infty \), let it be \( \{\theta_s'\} \). Then \( \theta_s' \to \theta' \) in \( \Omega \), and by continuity of \( \psi_n(\theta) \), \( \psi_n(\theta_s') \to \psi_n(\theta') \). By convergence of \( \{\psi_n(\theta_r)\} \), we obtain that

\[ \psi_n(\theta') = \psi_n(\theta) \]

for \( \theta' \) the limit of any convergent subsequence. In the same manner as was concluded for \( \bar{\theta} \) we must conclude

\[ \frac{\partial \psi_n(\bar{\theta})}{\partial \theta} = \frac{\partial \psi_n(\theta')}{\partial \theta} = 0. \]

If Assumption 6 holds this can not occur unless \( \theta' = \bar{\theta} \) and thus \( \{\theta_r\}_{r=1}^\infty \to \bar{\theta} \).
Chapter 7 contains numerical illustrations of the performance of the algorithm under different models and different choices of $p$. In some cases the value $\theta'$ generated is compared to the minimizing value obtained from other computing techniques. In the actual implementation of the algorithm an approximating process was used instead of the minimization required for each choice of $\lambda_r$. 
4. GENERAL RESULTS

4.1 Introduction

In this chapter it will be shown that the sequence of estimators \( \{\hat{\theta}_n\}^{\infty}_{n=1} \) converges with probability one to a point \( \theta_0 \) in \( \Omega \), where \( \theta_0 \) is the value in \( \Omega \) defined in Assumption 7. That is,

\[
E_{\theta^*} \rho(Q(x,y,\theta_0)) < E_{\theta^*} \rho(Q(x,y,\theta))
\]

for any \( \theta \neq \theta_0 \) and \( \theta \) in \( \Omega \). Under two different sets of sufficient conditions, the asymptotic distribution of \( \sqrt{n} (\hat{\theta}_n - \theta_0) \) is derived. This is the \( k \)-variate normal with mean 0 and variance-covariance matrix \( G^{-1}(\theta_0) \Lambda(\theta_0) G^{-1}(\theta_0) \).

When \( p \geq 2 \), asymptotic normality follows from a set of sufficient conditions which include the existence and finiteness of the second partials of \( \Psi_n(\theta) \) at all points \( (x,y,\theta) \) in \( S \times \Omega \). This turns out to be a stronger condition than is necessary for normality to follow. Thus, for the choice of \( 1 < p < 2 \), when this latter assumption on the second partials of \( \Psi_n(\theta) \) may not be satisfied, it is possible to formulate a set of weaker assumptions. Of course, this weaker set applies for all values of \( p > 1 \), yet both approaches are presented, since for a particular model and a choice of \( p \geq 2 \) the stronger set of conditions is much easier to verify.

The predictor for the endogenous variable \( y \) and its asymptotic properties are presented in Section 4.4.
Recall that two ways for viewing the sequence of inputs \( \{x_t\}_{t=1}^{\infty} \) were presented in Chapter 2, both were required to guarantee that Assumptions 14 and 15 be satisfied. Under the first approach, \( \{x_t\}_{t=1}^{\infty} \) is a realization of a sequence of random variables and \( \{(x_t, y_t)\}_{t=1}^{\infty} \) denotes independent and identically distributed vectors \((x_t, y_t)\). As was previously noted, in this case, Assumption 14 is guaranteed by the conclusion of the Strong Law of Large Numbers, and the Multivariate Central Limit Theorem implies Assumption 15.

Alternatively, \( \{x_t\}_{t=1}^{\infty} \) was considered as a sequence of constants. Propositions 4.1 and 4.2 show that such a sequence can be generated so that Assumptions 14 and 15 hold and so that there exists a measure \( u_n(A) \) on \((X, \mathcal{G}_X)\) converging weakly to a probability measure \( u(X) \) on \((X, \mathcal{G}_X)\).

For convenience we repeat Assumptions 14 and 15.

**Assumption 14**

Let \( g(x, y, \theta) \) be a real valued \( \Theta \) measurable function for each \( \theta \) in \( \Omega \), such that

\[
E_{\theta X} |g(x, y, \theta)| < \infty,
\]

then

\[
\frac{1}{n} \sum_{t=1}^{n} g(x_t, y_t, \theta) \rightarrow E_{\theta X} g(x, y, \theta)
\]

with probability one as \( n \rightarrow \infty \).
Assumption 15

Let \( g(x, y, \theta) \) be a \( k \times 1 \) vector such that \( g_1(x, y, \theta) \) is a real valued \( \Theta \) measurable function for all \( \theta \) in \( \Omega \) and

\[
0 < E_{g_1 g_1}(x, y, \theta) < \infty
\]

for \( \theta \) in \( \Omega \), and all \( i = 1, 2, \ldots, k \), then

\[
\left( \frac{1}{n} \sum_{t=1}^{n} (g(x_t, y_t, \theta) - E_{g_1 g_1}(x, y, \theta)) \right) \overset{L}{\longrightarrow} N_k(0, \Sigma)
\]

where \( \Sigma \) is the \( (k \times k) \) matrix with \((i, j)\) element

\[
[S]_{i,j} = E_{g_1}((g_1(x, y, \theta) - E_{g_1 g_1}(x, y, \theta))(g_j(x, y, \theta)) - E_{g_1 g_j}(x, y, \theta))
\]

Proposition 4.1

Let the sequence \( \{x_t\}_{t=1}^{\infty} \) be generated as follows:

\[
\{x_t\}_{t=1}^{\infty} = \{x_1, x_2, \ldots, x_r, x_{r+1}, x_{r+2}, \ldots, x_{2r}, \ldots, x_{(m-1)r+1}, \ldots, x_{mr}\}
\]

and

\[
x_h = x_{(i-1)r+h} \quad \text{for} \ i = 1, 2, \ldots, m ;
\]

\[
h = 1, 2, \ldots, r ;
\]

\[
n = mr
\]
for \( n \) in \( \{1, 2, \ldots\} \). In addition, let \( X \) be a Borel subset of \( \mathbb{R}^m \) and \( \mathcal{G}_X \) be the collection of \( m \)-dimensional Borel subsets of \( X \).

Let

\[
 u_n(A) = \frac{1}{n} \sum_{t=1}^{n} I_A(x_t)
\]

for any \( A \) in \( \mathcal{G}_X \) and \( x_t \) in \( \{x_t\}_{t=1}^{n} \), then

\[
 u_n(A) \xrightarrow{\text{weakly}} u_x(A)
\]

where

\[
 u_x(A) = \sum_{h=1}^{r} \frac{1}{r} \sum_{i=1}^{m} I_A(x_h)
\]

is a probability measure and

\[
 \{x_h\}_{h=1}^{r} = \{x_1, x_2, \ldots, x_r\}.
\]

**Proof**

Consider

\[
 u_n(A) = \frac{1}{n} \sum_{t=1}^{n} I_A(x_t)
\]

for any \( A \) in \( \mathcal{G}_X \). Then \( u_n(A) \) is a probability measure on \( (X, \mathcal{G}_X) \), and

\[
 \frac{1}{n} \sum_{t=1}^{n} I_A(x_t) = \frac{1}{r} \sum_{h=1}^{r} \frac{1}{m} \sum_{i=1}^{m} I_A(x_{(i-1)r+h})
\]

\[
 = \frac{1}{r} \sum_{h=1}^{r} \frac{1}{m} \sum_{i=1}^{m} I_A(x_h)
\]

\[
 = \frac{1}{r} \sum_{h=1}^{r} I_A(x_h).
\]
Hence,

\[ u_n(A) = u^X(A) \] for each \( n = 1, 2, \ldots \), and \( A \) in \( \mathcal{G}_x \)

and the conclusion follows. \( \square \)

**Proposition 4.2**

Let \( \{x_t\}_{t=1}^{\infty} \) be generated as described in Proposition 4.1. For each \( x_t \) let \( y_t \) be a random variable with distribution function \( F_{y(x_t, \theta^*)} \) for all \( t = 1, 2, \ldots \), and let the \( y_t \) be independent. Let

\[ u^X(A) = \frac{1}{r} \sum \frac{1}{r} \mathbb{I}_A(x_h) \]

for

\[ \{x_h\}_{h=1}^{r} = \{x_1, x_2, \ldots, x_r\} \]

and let \( g \) be a \((k \times 1)\) real valued \( \mathcal{B} \) measurable function such that

\[ \int \int g^2(x, y, \theta) dF_{y(x, \theta^*)} du_x < \infty \]

for \( i = 1, 2, \ldots, t \). Then Assumption 14 and 15 are satisfied.

**Proof**

For each \( n = mr \) consider the sum

\[ \frac{1}{n} \sum_{t=1}^{n} g(x_t, y_t, \theta) = \frac{1}{r} \sum_{h=1}^{r} \frac{1}{m} \sum_{i=1}^{m} g(x_{i-1}+h, y_{(i-1)r+h}, \theta) \]

Since for each \( h = 1, 2, \ldots, r \),

\[ \{y_t, y_{r+h}, y_{2r+h}, \ldots, y_{(m-1)r+h}\} \]
are independent and identically distributed the Strong Law of Large Numbers implies

\[ \frac{1}{m} \sum_{i=1}^{m} g(x_{(i-1)r+h}, y_{(i-1)r+h}, \theta) \to E_y(x_h, \theta^*) g(x_h, y, \theta) \]

with probability one as \( m \to \infty \). Hence, with probability one

\[ \frac{1}{n} \sum_{t=1}^{n} g(x_t, y_t, \theta) \to \frac{1}{r} \sum_{h=1}^{r} E_y(x_h, \theta^*) g(x_h, y, \theta) \]

\[ = E_{\theta^*} g(x, y, \theta) \]

as \( n \to \infty \).

Similarly,

\[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} g(x_t, y_t, \theta) = \frac{1}{\sqrt{r}} \sum_{h=1}^{r} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} g(x_{h(i-1)r+h}, \theta) \]

By the Multivariate Central Limit Theorem, in Rao [16, p. 128]

\[ \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \frac{1}{\sqrt{m}} g(x_h, y_{(i-1)r+h}, \theta) - E_y(x_h, \theta^*) g(x_h, y, \theta) \]

\[ \xrightarrow{L} N_k(0, \frac{1}{r} \Sigma_h) \]

where \( \Sigma_h \) is the \((k \times k)\) matrix with \((i, j)\) term given by

\[ [\Sigma_h]_{i, j} = E_y(x_h, \theta^*) \left((g_i(x_h, y, \theta) - E_y(x_h, \theta^*) g_i(x_h, y, \theta)) \times \right. \]

\[ \left. \times (g_j(x_h, y, \theta) - E_y(x_h, \theta^*) g_j(x_h, y, \theta)) \right) \]

For each \( h' \), the characteristic function of

\[ \frac{1}{\sqrt{r}} \sum_{h=1}^{r} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (g(x_{h'}, y_{(i-1)r+h'}, \theta) - E_y(x_h, \theta^*)) \]
is
\[ iu'\left(\Sigma_{h=1}^{r} Z_{m_{h}}\right)u = \prod_{h=1}^{r} Ee_{m_{h}} = C_{m}(u) \]
where the \( k \times 1 \) vector \( Z_{m_{h}} \) is
\[ Z_{m_{h}} = \sqrt{\frac{1}{r}} \sqrt{\frac{1}{m}} \sum_{i=1}^{m} \left( g(x_{h}, y_{(i-1)r+h}, \theta) - E_{y}(x_{h}, \theta) g(x_{h}, y_{(i-1)r+h}, \theta) \right) \]
and the \( \{Z_{m_{h}}\}_{h=1}^{r} \) are independent. Hence,
\[ \lim_{m \to \infty} C_{m}(u) = \prod_{h=1}^{r} Ee_{m_{h}} \]
which is the characteristic function of a \( k \)-variate normal with mean 0 and variance-covariance matrix \( \frac{1}{r} \Sigma_{h=1}^{r} \Sigma_{h} \). Hence,
\[ \sqrt{n} \sum_{i=1}^{n} \left( g(x_{i}, y_{i}, \theta) - E_{y} g(x_{i}, y_{i}, \theta) \right) \xrightarrow{L} N_{k}(0, \Sigma) \]
where the \((i,j)\) element of \( \Sigma \) is
and the expectation is taken with respect to $u_x$. \]

### Lemma 4.3

Let Assumption 2 be satisfied, then for each $\theta$ in $\Omega$

$$\lim_{\delta \to 0} E_{\theta^*} \inf_{\|\theta' - \theta\| < \delta} \rho(x, y, \theta') = E_{\theta^*} \rho(x, y, \theta)$$

**Proof**

Let $\delta_n$ be any convergent subsequence such that $\delta_n \to 0$ as $n_k \to \infty$, and let

$$z_{n_k} = \inf_{\|\theta' - \theta\| < \delta_n} \rho(x, y, \theta')$$

Then,

$$0 \leq z_{n_k} \leq \rho(x, y, \theta)$$

By continuity of $\rho$ in $\theta$ (Assumption 2), and by the Monotone Convergence Theorem

$$\lim_{k} E_{\theta^*} z_{n_k} = E_{\theta^*} \rho(x, y, \theta)$$

Since this is true for any convergent subsequence $\delta_n$, then

$$\lim_{\delta \to 0} E_{\theta^*} \inf_{\|\theta' - \theta\| < \delta} \rho(x, y, \theta') = E_{\theta^*} \rho(x, y, \theta)$$

for each $\theta$ in $\Omega$. \]
The conclusion of Lemma 4.3 is used in the following proof, which follows the ideas in Theorem 1 of Huber [10].

**Theorem 4.1**

Let the conclusion of Lemma 4.3 be satisfied, and let Assumptions 1, 2, 7, 8, and 14 hold, then with probability one

\[ \hat{\theta}_n \to \theta_0 \]

as \( n \to \infty \), where \( \{\hat{\theta}_n\}_{n=1}^{\infty} \) is the sequence of estimators and \( \theta_0 \) is defined in Assumption 7.

**Proof**

Let \( \theta_0 \in A(\theta_0, \delta_0) \subset \text{int}(\Omega) \). By compactness of \( \Omega \), the set \( \Omega/A(\theta_0, \delta_0) \) is a compact subset of \( \mathbb{R}^k \). Assumptions 2 and 8 imply continuity over \( \Omega \) of \( E_{\theta_0} \rho(x, y, \theta) \) and by the conclusion of Lemma 4.3, for any \( \varepsilon > 0 \) there exists a \( \delta(\varepsilon) > 0 \) such that for \( \theta \in \Omega/A(\theta_0, \delta) \).

\[ \|\theta' - \theta\| < \delta(\varepsilon) \] implies

\[ |E_{\theta_0} \rho(x, y, \theta') - E_{\theta_0} \rho(x, y, \theta)| < \varepsilon \]

and

\[ |E_{\hat{\theta}_n} \inf_{\|\theta' - \theta\| < \delta(\varepsilon)} \rho(x, y, \theta') - E_{\theta_0} \rho(x, y, \theta)| < \varepsilon . \quad (4.1) \]

Let \( A(\theta, \delta(\varepsilon)) \) be

\[ \{\theta' \mid \|\theta' - \theta\| < \delta(\varepsilon)\} , \]

then the \( A(\theta, \delta) \) form a covering of \( \Omega/A(\theta_0, \delta_0) \) and there exists a
finite set \( \{ \theta_1, \theta_2, \ldots, \theta_R \} \) such that

\[
\bigcup_{i=1}^{R} A(\theta_i, \delta(\varepsilon))
\]

are a finite covering of \( \Omega/A(\theta_0, \delta_0) \).

Given any \( \bar{\theta} \) in \( \Omega/A(\theta_0, \delta_0) \), then \( \bar{\theta} \) is in \( A(\theta_i, \delta(\varepsilon)) \) for some \( i = 1, 2, \ldots, R \) and by relation \((4.1)\) there exists a \( \delta(\varepsilon) \) such that for all \( i \),

\[
E_{\bar{\theta}^*} \left( \inf_{A(\theta_i, \delta(\varepsilon))} \rho(x, y, \theta_i) \right) > E_{\bar{\theta}^*} \rho(x, y, \theta_i) - \varepsilon.
\]

Let \( \varepsilon > 0 \) be chosen so that

\[
\inf_{\theta \in \Omega/A(\theta_0, \delta_0)} E_{\bar{\theta}^*} \rho(x, y, \theta) = E_{\bar{\theta}^*} \rho(x, y, \bar{\theta})
\]

\[
\geq E_{\bar{\theta}^*} \rho(x, y, \theta_0) + 4\varepsilon
\]

where \( \theta_0 \) is the value in \( \text{int}(\Omega) \) such that

\[
E_{\bar{\theta}^*} \rho(x, y, \theta_0) < E_{\bar{\theta}^*} \rho(x, y, \theta)
\]

for any \( \theta \) in \( \Omega \), \( \theta \neq \theta_0 \), and \( \bar{\theta} \) is in \( \Omega/A(\theta_0, \delta_0) \).

Then, it follows that

\[
\inf_{\theta' \in A(\theta_i, \delta(\varepsilon))} E_{\bar{\theta}^*} \rho(x, y, \theta') \geq E_{\bar{\theta}^*} \rho(x, y, \theta_i) \inf_{A(\theta_i, \delta(\varepsilon))} \rho(x, y, \theta')
\]

\[
> E_{\bar{\theta}^*} \rho(x, y, \theta_i) - \varepsilon
\]

\[
\geq E_{\bar{\theta}^*} \rho(x, y, \theta_0) + 3\varepsilon
\]

for all \( i = 1, 2, \ldots, R \).
By the conclusion of Lemma 4.3, if \( A(\theta, \delta) = \{ \theta' \mid \| \theta' - \theta \| < \delta \} \)

\[
E_{\theta^*} \inf_{A(\theta, \delta)} \rho(x, y, \theta') \rightarrow E_{\theta^*} \rho(x, y, \theta)
\]
as \( \delta \rightarrow 0 \), for \( \theta \) in \( \Omega \), and by Assumptions 8 and 14 there exists an \( n_{\delta} \) depending on the sequence \( \{(x_t, y_t)\}_{t=1}^{\infty} \) such that with probability one for all \( n \geq n_{\delta} \) and all \( i = 1, 2, \ldots, R \),

\[
\frac{1}{n} \sum_{t=1}^{n} \inf_{\theta' \in A(\theta_i, \delta(\varepsilon))} \rho(x_t, y_t, \theta') \geq E_{\theta^*} \inf_{\theta' \in A(\theta_i, \delta(\varepsilon))} \rho(x, y, \theta') - \varepsilon
\]

and

\[
\inf_{\theta' \in A(\theta_1, \delta(\varepsilon))} \frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \theta') \geq \frac{1}{n} \sum_{t=1}^{n} \inf_{\theta' \in A(\theta_1, \delta(\varepsilon))} \rho(x_t, y_t, \theta') \\
\geq E_{\theta^*} \rho(x, y, \theta_0) + 2\varepsilon
\]

Hence, with probability one, for all \( n \geq n_{\delta} \) and some \( i = i_0 \)

\[
\inf_{\theta \in \Omega/A(\theta_0, \delta_0)} \frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \theta) = \frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \bar{\delta}) \\
\geq \frac{1}{n} \sum_{t=1}^{n} \inf_{\theta' \in A(\theta_0, \delta(\varepsilon))} \rho(x_t, y_t, \theta') \\
\geq E_{\theta^*} \rho(x, y, \theta_0) + 2\varepsilon \quad (4.2)
\]

for \( \bar{\delta} \) in \( A(\theta_0, \delta(\varepsilon)) \subset \Omega/A(\theta_0, \delta_0) \).
Similarly, with probability one for all $n \geq n_0 \geq n_0'$,

$$\frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \theta) < E_{\theta^*} \rho(x, y, \theta_0) + \varepsilon$$

and

$$\inf_{\theta \in \Omega} \frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \theta) \leq \frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \theta_0)$$

$$< E_{\theta^*} \rho(x, y, \theta_0) + \varepsilon.$$

Hence, with probability one, for $n \geq n_0'$, $\hat{\theta}_n : S^n \to \Omega$ satisfies

$$\inf_{\theta \in \Omega} \frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \theta) = \frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \hat{\theta}_n)$$

$$< E_{\theta^*} \rho(x, y, \theta_0) + \varepsilon$$

and by relation (4.2)

$$\frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \hat{\theta}_n) < \inf_{\theta \in \Omega/A(\theta_0', \delta_0')} \frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \theta).$$

Thus, since $\hat{\theta}_n$ is a value in $\Omega$, $\hat{\theta}_n$ is in $A(\theta_0', \delta_0')$ with probability one for all $n \geq n_0'$. $\square$

The previous theorem relies on the existence of a unique value $\theta = \theta_0$ in the interior of $\Omega$ such that

$$E_{\theta^*} \rho(x, y, \theta_0) < E_{\theta^*} \rho(x, y, \theta)$$

for $\theta$ in $\Omega$, $\theta \neq \theta_0$. Corollary 4.1 extends this result to the case when there is more than one value in $\Omega$ for which $E_{\theta^*} \rho(x, y, \theta)$ attains its minimum.
Corollary 4.1

Let the conditions of Theorem 4.1 be satisfied, where Assumption 7 is substituted by the following:

\[ \gamma = \left\{ \theta \in \Omega \mid E_{\theta^*} \rho(x, y, \theta) = \inf_{\theta \in \Omega} E_{\theta^*} \rho(x, y, \theta) \right\} \]

and \( \theta \) in \( \gamma \) implies \( \theta \) is in \( \text{int}(\Omega) \).

Then, with probability one, for all \( n \geq n_0 \), where \( n_0 \) depends on the sequence \( \{(x_t, y_t)\}_{t=1}^{\infty} \), the estimator \( \theta_n \) is in any neighborhood of \( \theta \) in \( \gamma \).

Proof

\( \gamma \) is a compact subset of \( \Omega \). Consider the spheres

\[ A(\theta, \delta) = \{ \theta' \mid \| \theta' - \theta \| < \delta \} \]

for some \( \delta > 0 \) and \( \theta \) in \( \gamma \).

Then there exists a finite union of them which covers the set \( \gamma \), say

\[ \gamma \subseteq \bigcup_{i=1}^{R} A(\theta_i, \delta) \cdot \]

The set

\[ \Omega/ \bigcup_{i=1}^{R} A(\theta_i, \delta) \]

is also compact. Similarly to the proof of Theorem 4.1, with probability one for all \( n \geq n_0 \)

\[ \inf_{\theta \in \Omega/ \bigcup_{i=1}^{R} A(\theta_i, \delta)} \frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \theta) \geq E_{\theta^*} \rho(x, y, \theta') + 2\delta \]
and 

$$\inf_{\theta \in \Omega} \frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \theta) < E_{\theta^*} \rho(x, y, \theta') + \epsilon$$

for all $\theta'$ in $\gamma$.

Hence, with probability one for all $n \geq n_0$ and all $\theta'$ in $\gamma$,

$$\frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \hat{\theta}_n) < E_{\theta^*} \rho(x, y, \theta') + \epsilon$$

$$< \inf_{\theta \in \Omega \cup \bigcup_{i=1}^{I} A(\theta_i, \delta)} \frac{1}{n} \sum_{t=1}^{n} \rho(x_t, y_t, \theta)$$

and therefore $\hat{\theta}_n$ is in

$$\bigcup_{i=1}^{I} A(\theta_i, \delta).$$

If $\hat{\theta}_n$ is in the interior of $\Omega$ for each $n = 1, 2, \ldots$, then at $\theta = \hat{\theta}_n$

$$\frac{\partial y_n(\theta)}{\partial \theta} = 0.$$

Huber [10] gives sufficient conditions such that any sequence $[\theta_n']_{n=1}^{\infty}$ satisfying the above equation converges with probability one to a point $\theta_0$ in $\Omega$. Instead of Assumption 7, an equivalent assumption is for $\theta_0$ to be the unique value in $\Omega$ satisfying

$$E_{\theta^*} \frac{\partial \rho(x, y, \theta)}{\partial \theta} = 0.$$

This result is presented in Theorem 4.2.
Theorem 4.2

Let Assumptions 1, 2, 4, 9, and 14 be satisfied, and let $\theta_0$ in $\Omega$ be the unique solution to

$$E_{\theta^*} \left( \frac{\partial p(x, y, \theta)}{\partial \theta} \right) = 0$$

where

$$\frac{\partial p(x, y, \theta)}{\partial \theta}$$

is integrable for all $\theta$ in $\Omega$, then any sequence of measurable functions $\{\theta_n\}_{n=1}^{\infty}$, where $\theta_n: \mathbb{R}^n \rightarrow \Omega$ and

$$\frac{\partial \psi_n(\theta_n)}{\partial \theta} = 0,$$

converges to $\theta_0$ with probability one as $n \rightarrow \infty$.

Proof

It can be shown that this is a special case of Theorem 3 in Huber [10] when Assumption 14 substitutes for the result implied by the Strong Law of Large Numbers.

4.3 Asymptotic Normality of $\hat{\theta}_n$

Throughout this section $\theta_0$ denotes the parameter value in the interior of $\Omega$ such that $\hat{\theta}_n \rightarrow \theta_0$ with probability one as $n \rightarrow \infty$. Theorem 4.3 gives sufficient conditions for the asymptotic normality of $\sqrt{n} (\hat{\theta}_n - \theta_0)$. The same conclusion follows from Theorem 4.4 under a set of stronger assumptions.
Theorem 4.3

Let Assumptions 2, 3, 4, 5, 7, 10, 11, 12, 14, and 15 be satisfied and let

\[
\frac{\partial \rho(x, y, \theta)}{\partial \theta_i} - \frac{\partial \rho(x, y, \theta_0)}{\partial \theta_i} - \sum_{j=1}^{k} \frac{\partial^2 \rho(x, y, \theta_0)}{\partial \theta_i \partial \theta_j} (\theta^j - \theta_0^j) \leq m^1(x, y, \delta) \|	heta - \theta_0\| \tag{4.3}
\]

for all \(i = 1, 2, \ldots, k\), all \(\theta \in \Omega(\theta_0, \delta)\) and each \(0 < \delta < \varepsilon'\), where \(m^1(x, y, \delta)\) is a measurable function for each \(0 < \delta < \varepsilon'\)

such that as \(n \to \infty\)

\[
\frac{1}{n} \sum_{t=1}^{n} m^1(x_t, y_t, \theta) \to M(\delta)
\]

for almost all realizations \(\{(x_t, y_t)\}_{t=1}^{\infty}\) and

\[M(\delta) \to 0 \text{ as } \delta \to 0.\]

Then,

\[
\sqrt{n} (\hat{\theta}_n - \theta_0) \overset{L}{\to} N_k(0, G^{-1}(\theta_0) A(\theta_0) G^{-1}(\theta_0)).
\]

Proof

Assumptions 2 and 4 insure the existence and measurability of the first partials of \(\rho(x, y, \theta)\) in a neighborhood of \(\theta = \theta_0\), say in \(A(\theta_0, \delta)\) for each \(0 < \delta < \varepsilon'\) and some \(\varepsilon' > 0\). Since \(\theta_0\) is an interior point of \(\Omega\), \(A(\theta_0, \delta) \subset \text{int}(\Omega)\) for some \(\varepsilon' > 0\) and all \(0 < \delta < \varepsilon'\). Convergence of \(\hat{\theta}_n\) to \(\theta_0\) with probability one implies that for all \(n\) sufficiently large, and some \(0 < \delta < \varepsilon'\), \(\hat{\theta}_n\) is in \(A(\theta_0, \delta)\) almost surely.
Hence, at $\theta = \hat{\theta}_n$, for all $n$ sufficiently large and almost all realizations $\{(x_t, y_t)\}_{t=1}^{\infty}$

$$\left( \frac{\partial \psi_n(\hat{\theta}_n)}{\partial \hat{\theta}_i} \right) = 0$$

and relation (4.3) becomes

$$\frac{\partial \rho(x, y, \theta_0)}{\partial \theta_i} - \sum_{j=1}^{k} \sum_{l=1}^{p} \frac{\partial^2 \rho(x, y, \theta_0)}{\partial \theta_j \partial \theta_i} (\hat{\theta}_i - \theta_{0j}) = 0$$

$$\leq \lambda_i^{(x, y, \delta)} \| \hat{\theta}_n - \theta_0 \|$$

for all $i, j = 1, 2, \ldots, k$. Note that Assumptions 3 and 5 imply the existence of the second partials of $\rho(x, y, \theta)$, at $\theta = \theta_0$ except on a set of measure zero.

It follows that

$$\frac{1}{n} \sum_{t=1}^{n} \frac{\partial \rho(x_t, y_t, \theta_0)}{\partial \theta_i} + \sum_{j=1}^{k} \sum_{l=1}^{p} \frac{\partial^2 \rho(x_t, y_t, \theta_0)}{\partial \theta_j \partial \theta_i} (\hat{\theta}_i - \theta_{0j})$$

$$\leq \frac{1}{n} \left( \sum_{t=1}^{n} \lambda_i^{(x_t, y_t, \delta)} \right) \| \hat{\theta}_n - \theta_0 \|$$

or equivalently

$$\frac{1}{n} \sum_{t=1}^{n} \frac{\partial \rho(x_t, y_t, \theta_0)}{\partial \theta_i} + \sum_{j=1}^{k} \frac{1}{n} \sum_{l=1}^{p} \frac{\partial^2 \rho(x_t, y_t, \theta_0)}{\partial \theta_j \partial \theta_i} (\hat{\theta}_i - \theta_{0j})$$

$$= \lambda_n^{(x, y, \delta)} \frac{1}{n} \sum_{t=1}^{n} \lambda_i^{(x_t, y_t, \delta)} \| \hat{\theta}_n - \theta_0 \|$$

where $0 < |\lambda_n^{(x, y, \delta)}| \leq 1$ for $i = 1, 2, \ldots, k$ and $0 < \delta < \varepsilon'$.

Hence,
In matrix notation,
\[
\begin{align*}
\Omega &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \rho(x_t, y_t, \theta)}{\partial \theta_i} + \frac{k}{n} \sum_{j=1}^{n} \sum_{t=1}^{n} \frac{\partial^2 \rho(x_t, y_t, \theta)}{\partial \theta_i \partial \theta_j} \nabla_n (\hat{\theta}_n - \theta_0) \\
&- \frac{\lambda_n'(\delta)}{||\hat{\theta}_n - \theta_0||} \sum_{t=1}^{n} \frac{1}{n} m^i(x_t, y_t, \delta) \nabla_n (\hat{\theta}_n - \theta_0) \\
&+ \sum_{j=1}^{k} \frac{1}{n} \sum_{t=1}^{n} \frac{\partial^2 \rho(x_t, y_t, \theta)}{\partial \theta_i \partial \theta_j} \nabla_n (\hat{\theta}_n - \theta_0). 
\end{align*}
\]

Consider the second term in the above expression. Firstly,
\[
\begin{align*}
(\sum_{t=1}^{n} \frac{\lambda_n'(\delta)}{n} m(x_t, y_t, \delta))
\end{align*}
\]
is a \((k \times 1)\) vector with \(i^{th}\) element given by
\[
\frac{\lambda_n'(\delta)}{n} \sum_{t=1}^{n} \frac{1}{n} m^i(x_t, y_t, \delta),
\]
where
\[
|\lambda_n'(\delta)| \sum_{t=1}^{n} \frac{1}{n} m^i(x_t, y_t, \delta) \leq |\sum_{t=1}^{n} \frac{1}{n} m^i(x_t, y_t, \delta)|
\]
for all $i = 1, 2, \ldots, k$, and $(\hat{\theta}_n - \theta_0)^n/\|\hat{\theta}_n - \theta_0\|$ is a $1 \times k$ vector with $j^{th}$ component satisfying

$$\frac{|\hat{\theta}^j_n - \theta^j_0|}{\|\hat{\theta}_n - \theta_0\|} \leq 1.$$ 

Thus, the second term is a $k \times k$ matrix $\frac{1}{n} M_n(\delta)$ with $(i,j)$ element $\frac{1}{n} [M_n(\delta)]_{i,j}$ such that

$$\left| \frac{1}{n} \sum_{t=1}^{n} \lambda_n(x_t, y_t, \delta) \cdot \frac{(\hat{\theta}^i_n - \theta^i_0)}{\|\hat{\theta}_n - \theta_0\|} - \frac{1}{n} \sum_{t=1}^{n} m^i(x_t, y_t, \delta) \right| \leq \frac{1}{n} \sum_{t=1}^{n} m^i(x_t, y_t, \delta).$$

By relation (4.3) let $0 < \delta < \varepsilon'$ such that for any $\gamma > 0$

$$M(\delta) < \frac{\gamma}{2}.$$

Then for all $n$ sufficiently large $\hat{\theta}_n$ is in $A(\theta_0, \delta)$ with probability one and for almost all realizations $\{(x_t, y_t)\}_{t=1}^{\infty}$ and

$$\left| \frac{1}{n} [M_n(\delta)]_{i,j} \right| \leq \left| \frac{1}{n} \sum_{t=1}^{n} m^i(x_t, y_t, \delta) \right| < M(\delta) + \frac{\gamma}{2} = \gamma.$$

Hence, with probability one,

$$\frac{1}{n} [M_n(\delta)]_{i,j} \to 0$$

as $n \to \infty$ for $i,j = 1, 2, \ldots, k$. Thus,

$$\frac{1}{\sqrt{n}} \left( \sum_{t=1}^{n} \frac{\partial \rho(x_t, y_t, \theta)}{\partial \theta} \right) = \left[ \frac{1}{n} M_n(\delta) - \frac{1}{n} q_n(\theta_0') \right] \sqrt{n} (\hat{\theta}_n - \theta_0) \cdot$$

Assumptions 7, 11, and 15 imply

$$\frac{1}{\sqrt{n}} \left( \sum_{t=1}^{n} \frac{\partial \rho(x_t, y_t, \theta)}{\partial \theta} \right) \overset{L}{\to} N_k(0, \Lambda(\theta_0')).$$
where \( \Lambda(\theta_0) \) has \((i,j)\) element

\[
E_{\theta^*} \left( \frac{\partial^2 \rho(x,y,\theta)}{\partial \theta_i \partial \theta_j} \right) < \infty
\]

for \( i, j = 1, 2, \ldots, k \). It follows then, that

\[
\left[ \frac{1}{n} M_n(\delta) - \frac{1}{n} G_n(\theta_0) \right] V_n (\hat{\theta}_n - \theta_0) \xrightarrow{L} N_k(0, \Lambda(\theta_0)) .
\]

Consider

\[
R_n(\delta) = \begin{cases} 
0 & \text{det} \left( \frac{1}{n} M_n(\delta) - \frac{1}{n} G_n(\theta_0) \right) = 0 \\
\left( \frac{1}{n} M_n(\delta) - \frac{1}{n} G_n(\theta_0) \right)^{-1} & \text{det} \left( \frac{1}{n} M_n(\delta) - \frac{1}{n} G_n(\theta_0) \right) \neq 0 .
\end{cases}
\]

By Assumptions 10 and 14, \( \frac{1}{n} G_n(\theta_0) \) with \((i,j)\) element

\[
\frac{1}{n} \sum_{t=1}^n \frac{\partial^2 \rho(x_t,y_t,\theta_0)}{\partial \theta_i \partial \theta_j} \rightarrow G(\theta_0)
\]

with \((i,j)\) element

\[
E_{\theta^*} \left( \frac{\partial^2 \rho(x,y,\theta_0)}{\partial \theta_i \partial \theta_j} \right) ,
\]

with probability one as \( n \to \infty \) for each \( i, j = 1, 2, \ldots, k \).

Hence, by Slutsky's Theorem

\[
\left( \frac{1}{n} M_n(\delta) - \frac{1}{n} G_n(\theta_0) \right) \xrightarrow{P} G(\theta_0) .
\]

Similarly, the inverse of a matrix is a continuous function of its elements and hence since \( \text{det}(G(\theta_0)) \neq 0 \) by Assumption 12, then

\[
R_n(\delta) \xrightarrow{P} G^{-1}(\theta_0) .
\]
Let

\[ Z_n(\delta) = \sqrt{n} (\hat{\theta}_n - \theta_0) - R_n(\delta) \frac{1}{\sqrt{n}} M_n(\delta) \]

- \frac{1}{\sqrt{n}} G_n(\theta_0) N \sqrt{n} (\hat{\theta}_n - \theta_0) .

Then, \( Z_n(\delta) \xrightarrow{P} 0 \) and

\[ R_n(\delta) \frac{1}{\sqrt{n}} M_n(\delta) - \frac{1}{\sqrt{n}} G_n(\theta_0) N \sqrt{n} (\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0, G^{-1}(\theta_0) \Lambda(\theta_0) G^{-1}(\theta_0)) . \]

\( x \in G^{-1}(\theta_0) \))

imply the conclusion of the theorem and

\[ \sqrt{n} (\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0, G^{-1}(\theta_0) \Lambda(\theta_0) G^{-1}(\theta_0)) . \]

In the following theorem, we assume the stronger set of assumptions for deriving asymptotic normality of \( \sqrt{n} (\hat{\theta}_n - \theta_0) \).

**Theorem 4.4**

Let Assumption 10 hold in a neighborhood of \( \theta_0 \) and let Assumptions 2, 4, 5, 7, 11, 12, 14, and 15 be satisfied, let

\[ \sup_{\|\theta' - \theta_0\| \leq \delta} \frac{\partial^2 \rho(x, y, \theta')}{\partial \theta_j \partial \theta_i} - \frac{\partial^2 \rho(x, y, \theta_0)}{\partial \theta_j \partial \theta_i} \]

\[ \leq h_{i,j}(x, y, \delta) \] (4.4)

where \( h_{i,j}(x, y, \delta) \) is a \( B \) measurable function for each \( \theta \) in \( \Omega \) such that for all \( i, j \),
Then the conclusion of Theorem 4.3 follows.

Proof

Assumptions 2, 4, and 5 guarantee continuity of the second partials of $\rho(x,y,\theta)$ with respect to $\theta$. Relation (4.4) implies that $\frac{\partial^2 \rho(x,y,\theta)}{\partial \theta_j \partial \theta_i}$ is finite valued, except perhaps on a set of measure zero not depending on $\theta$. Note that this is the condition violated by the case when $1 < p < 2$.

Representing the partials of $\rho$, by a first order Taylor's approximation around $\theta_0$, then for $\theta$ in $A(\theta_0, \delta)$ and each $i = 1, 2, \ldots, k$, we obtain

$$\frac{\partial \rho(x,y,\theta)}{\partial \theta_i} - \frac{\partial \rho(x,y,\theta_0)}{\partial \theta_i} = \sum_{j=1}^{k} \frac{\partial^2 \rho(x,y,\bar{\theta})}{\partial \theta_j \partial \theta_i} (\theta^j - \theta_0^j)$$

where

$$\bar{\theta} = \lambda(\theta) + (1 - \lambda)\theta_0$$

for $\lambda \in [0,1]$ and $\bar{\theta} \in A(\theta_0, \delta)$. Jennrich [12] has shown that there is a $\delta$ measurable function of $(x,y)$ corresponding to $\bar{\theta}$.

Hence,

$$\frac{\partial \rho(x,y,\theta)}{\partial \theta_i} - \frac{\partial \rho(x,y,\theta_0)}{\partial \theta_i} = \sum_{j=1}^{k} \frac{\partial^2 \rho(x,y,\bar{\theta})}{\partial \theta_j \partial \theta_i} (\theta^j - \theta_0^j)$$

$$= \sum_{j=1}^{k} \left( \frac{\partial^2 \rho(x,y,\bar{\theta})}{\partial \theta_j \partial \theta_i} - \frac{\partial^2 \rho(x,y,\theta_0)}{\partial \theta_j \partial \theta_i} \right) (\theta^j - \theta_0^j)$$

and

$$E_{\theta} h_{i,j}(x,y,\delta) \to 0 \text{ as } \delta \to 0.$$
Let
\[ m^i(x, y, \delta) = \sum_{j=1}^{k} \sup_{A(\theta_0, \delta)} \left| \frac{\partial^2 \rho(x, y, \theta)}{\partial \theta_j \partial \theta_l} - \frac{\partial^2 \rho(x, y, \theta_0)}{\partial \theta_j \partial \theta_l} \right| \]
then for each \( 0 < \delta < \varepsilon \) such that \( A(\theta_0, \varepsilon) \) is in \( \text{int}(\Omega) \),
\[ m^i(x, y, \delta) \leq \sum_{j=1}^{k} h_{i,j}(x, y, \delta) \]
and \( m^i(x, y, \delta) \) is \( \mathfrak{B} \) measurable for each \( 0 < \delta < \varepsilon \). By the above relation and relation (4.4),
\[ 0 \leq E_{\theta^*} m^i(x, y, \delta) \leq \sum_{j=1}^{k} E_{\theta^*} h_{i,j}(x, y, \delta) \to 0 \]
as \( \delta \to 0 \).
\[ 0 \leq \frac{1}{n} \sum_{t=1}^{n} m^i(x_t, y_t, \delta) = \sum_{t=1}^{n} \frac{1}{n} \sum_{j=1}^{k} \sup_{A(\theta_0, \delta)} \left| \frac{\partial^2 \rho(x_t, y_t, \theta)}{\partial \theta_j \partial \theta_l} - \frac{\partial^2 \rho(x_t, y_t, \theta_0)}{\partial \theta_j \partial \theta_l} \right| \to E_{\theta^*} m^i(x, y, \delta) \]
with probability one as \( n \to \infty \), by Assumption 14. Hence, relation (4.3) in Theorem 4.3 is satisfied and together with the other assumptions the conclusion of the theorem follows. \( \square \)
The asymptotic distribution derived is in terms of the parameter value \( \theta_0 \) and expectations of measurable functions of \((x,y)\). We now investigate the asymptotic distribution when

(a) \( G(\theta) \) is evaluated at \( \theta = \hat{\theta}_n \), and

(b) if an estimate is substituted for the variance-covariance matrix of the asymptotic distribution.

The properties of this estimator are derived from the conclusion of Lemma 4.2.

**Lemma 4.1**

Let \( h(y,\theta) \) be a measurable function of the real variable \( y \) for each value of \( \theta \) in \( \Omega \) a compact subset of \( \mathbb{R}^k \), let \( y_1, y_2, \ldots \) be a sequence of independent and identically distributed random variables, and let \( u_n \) be a sequence of functions with range in \( \Omega \) such that

\[
u_n = u_n(y_1, y_2, \ldots, y_n)
\]
is measurable on the space of \( y_1, y_2, \ldots, y_n \); \( n = 1, 2, \ldots \), if

1. \( E_{\theta^*}(h(y,\theta)) \) is continuous over \( \Omega \),

2. For each \( \theta \) in a countable base of \( \Omega \);

\[
\lim_{\delta \to 0} (E_{\theta^*} \sup_{\|\theta-\theta^*\|<\delta} |h(y,\theta') - h(y,\theta)|) = 0
\]

for \( \theta \) in \( \Omega \) then with probability one, for any \( \varepsilon > 0 \) there is an \( n_0 \), depending upon the sequence \( \{y_t\}_{t=1}^{\infty} \), such that for all
\[
\sum_{t=1}^{n} \frac{1}{n} h(y_t, \theta) - E_{\theta^*} h(y, \theta) < \varepsilon, \quad \theta \in \Omega
\]

and

\[
\sum_{t=1}^{n} \frac{1}{n} h(y_t, u_n) - E_{\theta^*} h(y, u_n) < \varepsilon.
\]

**Proof**

The proof is in Mickey [14].

As a consequence of the above lemma, we formulate Lemma 4.2.

**Lemma 4.2**

Let \( h(x, y, \theta) \) be a \( \mathcal{B} \) measurable function of \((x, y)\) for each \( \theta \in \Omega \), let \( \hat{\theta}_n \) be as defined in Chapter 3, let Assumptions 1 and 14 hold, and let

1. \( E_{\theta^*} h(x, y, \theta) \) be continuous over \( \Omega \),

2. \( \lim_{\delta \to 0} \left( \sup_{\|\theta - \theta'\| < \delta} |h(x, y, \theta) - h(x, y, \theta')| \right) = 0 \),

\( \theta \in \Omega \),

then, with probability one, for any \( \varepsilon > 0 \) there is an \( n_0 \), depending on the sequence \( \{(x_t, y_t)\}_{t=1}^{\infty} \) such that for all \( n \geq n_0 \),

\[
\sum_{t=1}^{n} \frac{1}{n} h(x_t, y_t, \theta) - E_{\theta^*} h(x, y, \theta) < \varepsilon,
\]

\( \theta \in \Omega \), and

\[
\sum_{t=1}^{n} \frac{1}{n} h(x_t, y_t, \hat{\theta}_n) - E_{\theta^*} h(x, y, \hat{\theta}_n) < \varepsilon.
\]
Proof

By Assumption 1, \( \Omega \) is a compact subset of \( \mathbb{R}^k \). The estimator \( \hat{s}_n \) is a \( \mathcal{B} \) measurable function with arguments

\[
\{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \}
\]

and range in \( \Omega \) for each \( n = 1, 2, \ldots \). The proof then is equivalent to that of Mickey [14] for Lemma 4.1 with Assumption 14 guaranteeing the conclusion of the Strong Law of Large Numbers.

Corollary 4.1

Let the conditions of Theorem 4.4 or Theorem 4.3 be satisfied, and let

\[
\mathbb{E}_{\theta^*} \left[ \frac{\partial^2 \varphi(x, y, \theta)}{\partial \theta_j \partial \theta_l} \right] < \infty
\]

be continuous at \( \theta = \theta_0 \), then

1. \( G(\hat{s}_n) \sqrt{n}(\hat{s}_n - \theta_0) \overset{L}{\rightarrow} N(0, \Lambda(\theta_0)) \).

If

\[
\frac{1}{n} g_n(\hat{s}_n) \rightarrow g(\theta_0)
\]

with probability one (or in probability) and the conclusion of Theorem 4.3 or Theorem 4.4 is satisfied then

2. \( \frac{1}{n} g_n(\hat{s}_n) \sqrt{n}(\hat{s}_n - \theta_0) \overset{L}{\rightarrow} N(0, \Lambda(\theta_0)) \).

Proof

Let

\[
G(\hat{s}_n) \sqrt{n}(\hat{s}_n - \theta_0) = G(\hat{s}_n) G^{-1}(\theta_0) G(\theta_0) \sqrt{n}(\hat{s}_n - \theta_0).
\]
By Theorem 4.3 or 4.4,

\[ G(\theta_0) \mathcal{W} n(\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0, \Lambda(\theta_0)). \]

Since \( \hat{\theta}_n \to \theta_0 \) with probability one and \( G(\theta) \) is finite valued continuous at \( \theta_0 \) by Slutsky's Theorem

\[ G(\hat{\theta}_n) \xrightarrow{P} G(\theta_0) \]

and

\[ G(\hat{\theta}_n) G^{-1}(\theta_0) \xrightarrow{P} I. \]

Thus,

\[ G(\hat{\theta}_n) \mathcal{W} n(\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0, \Lambda(\theta_0)). \]

Conclusion 2 follows similarly since

\[ \frac{1}{n} G_n(\hat{\theta}_n) \mathcal{W} n(\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0, \Lambda(\theta_0)). \]

Note that the conditions of Lemma 4.2 guarantee the existence of a consistent estimator of the dispersion matrix

\[ G^{-1}(\theta_0) \cdot \Lambda(\theta_0) \cdot G^{-1}(\theta_0). \]

What is required is that for all \( n \) sufficiently large, for any \( \varepsilon > 0 \) and all \( i, j = 1, 2, \ldots, k \),

\[ \left| \frac{1}{n} \sum_{t=1}^{n} \left( \frac{\delta^2 \rho(x_t, y_t, \hat{\theta}_n)}{\delta \theta_j \delta \theta_i} - \frac{\delta^2 \rho(x_t, y_t, \theta_0)}{\delta \theta_j \delta \theta_i} \right) \right| < \varepsilon \]

with probability one and hence \( \frac{1}{n} G_n(\hat{\theta}_n) \) is a strongly consistent estimator of \( G(\theta_0) \). However, for the case of \( 1 < p < 2 \),
Condition 2 of this lemma is not satisfied for
\[ h(x, y, \theta) = \frac{\delta^2 p(x, y, \theta)}{\delta \theta_i \delta \theta_j}. \]

Note that the stronger conditions of Theorem 4.4 imply those of Lemma 4.2.

4.4 The Predictor

The random variable \( y \) is usually considered as the endogenous variable determined by the model, the vector \( x \), as the non-stochastic exogenous variables. Generating predictions for \( y \) requires that we can derive a function \( f(x, \theta) \) analytically or numerically, such that the predictor \( \hat{y} \) is defined as

\[ \hat{y} = f(x, \hat{\theta}_n). \]

For the general implicit model the predictor \( \hat{y} \) is that value satisfying

\[ Q(x, \hat{y}, \hat{\theta}_n) = 0. \]

The conditions under which a solution \( \hat{y} \) exists are given in the Implicit Function Theorem, which we present for convenience. The proof is in [4].

Theorem 4.6 (Implicit Function Theorem)

Let \( (a, b)^T, \ (a \in \mathbb{R}^m, \ b \in \mathbb{R}^n) \) be an interior point of a set \( A \) in \( \mathbb{R}^m \times \mathbb{R}^n \) and suppose that the function \( f:A \rightarrow \mathbb{R}^n \) satisfies the following conditions:
1. \( f(a,b) = 0 \),

2. \( f \) is continuously differentiable in an open set 
   \( G \) containing \((a,b)^T\),

3. \( \text{Det } J[f(a,b)] \neq 0 \), where \( J[f(a,b)] \) has \((i,j)^{th}\)
   element \( \frac{\partial f_i(a,b)}{\partial b_j} \),

then, there exist intervals

\[
M = [a_1 - \alpha, a_1 + \alpha] \times \cdots \times [a_m - \alpha, a_m + \alpha],
\]

\[
N = [b_1 - \beta, b_1 + \beta] \times \cdots \times [b_n - \beta, b_n + \beta]
\]

and a continuous function \( \phi: M \rightarrow N \) such that \( y = \phi(x) \) is the
only solution lying in \( M \times N \) of the equation \( f(x,y) = 0 \). Moreover
\( \phi \) is continuously differentiable in \( \text{int}(M) \) and

\[
D\phi(x) = -(D_2^T f(x,\phi(x)))^{-1} D_1 f(x,\phi(x))
\]

for \( x \in \text{int}(M) \). Note that \( D\phi(x) \) is the matrix with \((i,j)^{th}\)

element \( \frac{\partial \phi_i(x)}{\partial x_j} \), \( D_1^T f(x,\phi(x)) \) has \((i,j)^{th}\)

element \( \frac{\partial f_i}{\partial x_j} \),

\( D_2 f(x,\phi(x)) \) has \((i,j)^{th}\)

element \( \frac{\partial f_i}{\partial y_j} \) evaluated at \( y = \phi(x) \).

In our case if we consider \( x \) as fixed, Condition 3 requires

\[
\left| \frac{\partial Q_x(\theta_0, y_0)}{\partial y} \right| \neq 0,
\]

Condition 1 that \( Q_x(y_0, \theta_0) = 0 \) for some \( y_0(x) \) in \( R \). Then we
can solve for \( y = \phi_x(\theta) \) for \( \theta \) in a neighborhood of \( \theta_0 \) and
\[
\left( \frac{\partial \phi_x(\theta)}{\partial \theta} \right)^T = \frac{-1}{\partial \phi_x(\theta, \phi_x(\theta))} \cdot \frac{\partial \phi_x(\theta, \phi_x(\theta))}{\partial \theta} \cdot \frac{\partial \phi_x(\theta, \phi_x(\theta))}{\partial \theta} \cdot \ldots
\]

Hence, if the conditions of the Implicit Function Theorem hold for \( Q \) given a fixed \( x \), and if \( n \) is sufficiently large

\[ y = \phi_x(\hat{\theta}_n) \]

exists with probability one for \( \hat{\theta}_n \) in a neighborhood of \( \theta_0 \).

**Corollary 4.2**

Let the condition of Theorem 4.6 hold for the function \( Q_x(y, \theta) \) with domain \( A \subset \mathbb{R} \times \mathbb{R}^k \), let \( y_0 \) be the point in \( A \) such that

\[ Q(x_0, \theta_0) = 0. \]

Let \( \hat{\theta}_n \to \theta_0 \) with probability one as \( n \to \infty \), and let the conclusion of Theorem 4.3 or Theorem 4.4 be satisfied, then

1. \( \hat{y} \to \phi_x(\theta_0) \) in probability as \( n \to \infty \),
2. \( \sqrt{n}(\hat{y} - \phi_x(\theta_0)) \to N(0, \left( \frac{\partial \phi}{\partial \theta} \right)^T G^{-1}(\theta_0) A(\theta_0) G^{-1}(\theta_0) \left( \frac{\partial \phi}{\partial \theta} \right))^\prime) \).

**Proof**

By Theorem 4.6 there is a neighborhood of \( \theta_0 \) such that \( \phi_x(\theta) \) is defined, is continuously differentiable in \( \theta \) and

\[ y = \phi_x(\theta) \]
exists. By the almost sure convergence of $\hat{\theta}_n$ to $\theta_0$ we can find an $n_0$ sufficiently large such that

$$\|\hat{\theta}_n - \theta_0\| < \varepsilon$$

for all $n \geq n_0$ and the solution $\hat{y}_n$ to

$$Q_x(y, \hat{\theta}_n) = 0$$

exists for all $n \geq n_0$ with probability one.

The function $\phi_x(\theta)$ is real valued continuous in $\theta$, hence by Stutsky's Theorem conclusion (a) follows and

$$\phi_x(\hat{\theta}_n) \to \phi_x(\theta_0)$$

in probability as $n \to \infty$.

Existence and continuity of the partials of $\phi_x(\theta)$ allow a first order Taylor approximation of $\phi_x(\theta)$ in a neighborhood of $\theta_0$. At $\theta = \hat{\theta}_n$ then,

$$\hat{y}_n = \phi_x(\hat{\theta}_n) = \phi_x(\theta_0) + \left(\frac{\partial \phi_x}{\partial \theta}(\hat{\theta}_n)\right)^T(\hat{\theta}_n - \theta_0)$$

for

$$\tilde{\theta}_n = \theta_0 + \alpha(\hat{\theta}_n - \theta_0),$$

$\alpha \in [0,1]$. Then

$$(\hat{y}_n - \phi_x(\theta_0)) = \left(\frac{\partial \phi_x}{\partial \theta}(\theta_0)\right)^T(\hat{\theta}_n - \theta_0)$$

$$+ \left(\frac{\partial \phi_x}{\partial \theta}(\tilde{\theta}_n) - \frac{\partial \phi_x}{\partial \theta}(\theta_0)\right)^T(\tilde{\theta}_n - \theta_0)$$
By continuity of $\frac{\partial \phi_x}{\partial \theta}$, $\frac{\partial \phi_x}{\partial \theta}$ is measurable and

$\frac{\partial \phi_x(\hat{\theta}_n)}{\partial \theta} - \frac{\partial \phi_x(\theta_0)}{\partial \theta} \to 0$ in probability as $n \to \infty$. By the Multivariate Central Limit Theorem and the conclusion of Theorem 4.3 or Theorem 4.4, it follows that

$\frac{\partial \phi_x(\theta_0)}{\partial \theta} \frac{\partial \phi_x(\theta_0)}{\partial \theta}^T \frac{\partial \phi_x(\theta_0)}{\partial \theta} \to 0$,

and hence

$\sqrt{n} (\hat{\theta}_n - \theta_0)$

$\xrightarrow{L} \text{N}(0, \frac{\partial \phi_x(\theta_0)}{\partial \theta} \frac{\partial \phi_x(\theta_0)}{\partial \theta}^T \text{G}^{-1}(\theta_0) \Lambda(\theta_0) \text{G}^{-1}(\theta_0) \frac{\partial \phi_x(\theta_0)}{\partial \theta})$.

Note that when

$\hat{y}_n = f(x, \hat{\theta}_n)$

then $\hat{y}_n = f(x, \hat{\theta}_n)$ is consistent if $\hat{\theta}_n \to \theta^*$ the true parameter value, and $\sqrt{n} (\hat{y}_n - f(x, \theta^*))$ is asymptotically normal. In the next chapter we consider this specific model in more detail.
5. THE CASE OF \( Q(\mathbf{x}_t, y_t, \theta^*) = y_t - f(\mathbf{x}_t, \theta^*) \)

In the single equation regression model, the defining relationship among variables and parameters is given by

\[
Q(\mathbf{x}_t, y_t, \theta^*) = y_t - f(\mathbf{x}_t, \theta^*) = \varepsilon_t.
\]

Alternatively, the model in reduced form is

\[
y_t = f(\mathbf{x}_t, \theta^*) + \varepsilon_t.
\]

The arguments \( \mathbf{x}_t \) and the random error \( \varepsilon_t \) satisfy the model specifications presented in Chapter 2, and the function \( f: (X \times \Omega) \to \mathbb{R} \) is known explicitly. The unknown parameter \( \theta^* \) is in the interior of \( \mathbb{R}^k \). For this specific case the estimator \( \hat{\theta}_n \) satisfies

\[
\min_{\theta \in \Omega} \sum_{t=1}^{n} |y_t - f(\mathbf{x}_t, \theta)|^p = \sum_{t=1}^{n} |y_t - f(\mathbf{x}_t, \hat{\theta}_n)|^p
\]

for a choice of \( p \geq 1 \).

The almost sure convergence of \( \{\hat{\theta}_n\}_{n=1}^{\infty} \) to a point \( \theta_0 \) in \( \Omega \) and the asymptotic distribution of \( \sqrt{n}(\hat{\theta}_n - \theta_0) \) were derived in Chapter 4. Under conditions imposed on the distribution function of \( \varepsilon_t \), it will be shown that some of the major assumptions are satisfied by the regression model. Furthermore, the estimator \( \hat{\theta}_n \) is strongly consistent and \( \sqrt{n}(\hat{\theta}_n - \theta^*) \) is asymptotically normal with mean zero and variance covariance matrix \( G^{-1}(\theta^*) \Lambda(\theta^*) G^{-1}(\theta^*) \).
Lemma 5.1

Consider the model

\[ y_t - f(x_t, \theta^*) = \epsilon_t, \]

and let Assumptions 1, 2, 4, 8, 16 and 17 be satisfied, then

1. For a given choice of \( p \geq 1 \), there exists a value \( \theta_0 \) in \( \Omega \) such that

\[ E_{\theta^*} \left| y - f(x, \theta_0) \right|^p \leq E_{\theta^*} \left| y - f(x, \theta) \right|^p, \ \theta \in \Omega. \]

2. In particular, for \( p = 1 \), then \( \theta_0 = \theta^* \) and

\[ E_{\theta^*} \left| y - f(x, \theta^*) \right| < E_{\theta^*} \left| y - f(x, \theta) \right| \]

for \( \theta \) in \( \Omega \).

If in addition, Assumption 18 is satisfied, then for all \( p > 1 \)

\[ E_{\theta^*} \left| y - f(x, \theta^*) \right|^p < E_{\theta^*} \left| y - f(x, \theta) \right|^p \]

for \( \theta \) in \( \Omega \).

Proof

Firstly consider the case for \( p = 1 \).

The model is given by

\[ y_t - f(x_t, \theta^*) = \epsilon_t. \]
At $\theta \neq \theta^*$, $Q(x_t, y_t, \theta)$ is $y_t - f(x_t, \theta)$, or

$$y_t - f(x_t, \theta) = e_t - (f(x_t, \theta) - f(x_t, \theta^*)) .$$

By Assumptions 2 and 8, $E_{\theta^*}|Q(x, y, \theta)|$ is a finite valued continuous function of $\theta$ over $\Omega$. Since $\Omega$ is compact, by Assumption 1, conclusion 1 follows, that is, there is a value of $\theta_0$ in $\Omega$ (depending on $p$) such that

$$E_{\theta^*}|y - f(\bar{x}, \theta^*) - (f(\bar{x}, \theta_0) - f(\bar{x}, \theta^*))|$$

$$\leq E_{\theta^*}|y - f(\bar{x}, \theta^*) - (f(\bar{x}, \theta) - f(\bar{x}, \theta^*))|$$

(5.1)

for all $\theta$ in $\Omega$. We know that

$$E_{\theta^*}|y - f(\bar{x}, \theta^*) - (f(\bar{x}, \theta_0) - f(\bar{x}, \theta^*))|$$

$$\geq E \min_{\theta \in \Omega} E(|y - f(\bar{x}, \theta^*) - (f(\bar{x}, \theta) - f(\bar{x}, \theta^*))|)$$

(5.2)

and by Assumption 16 the above equality is satisfied at $\theta = \theta^*$. Thus,

$$E_{\theta^*}|y - f(\bar{x}, \theta^*) - (f(\bar{x}, \theta_0) - f(\bar{x}, \theta^*))|$$

$$\geq E E(|y - f(\bar{x}, \theta^*)|)$$

for all $x$ in $X$.

Hence, in particular let $\theta = \theta^*$. Relations (5.1) and (5.2) imply
and therefore,

\[ \min_{\theta \in \Omega} E_{\theta^*} |y - f(x, \theta^*) - (f(x, \theta) - f(x, \theta^*))| \]

\[ = E_{\theta^*} |y - f(x, \theta^*)| = E_{\theta^*} |Q(x, y, \theta^*)| . \]

Suppose \( \theta_0 \neq \theta^* \). By Assumption 17, \( u_x(C(\theta_0)) > 0 \) implies

\[ E_{\theta^*} |y - f(x, \theta^*) - (f(x, \theta_0) - f(x, \theta^*))| \]

\[ = \int_{C(\theta_0)} E(|\epsilon - f(x, \theta_0) - f(x, \theta^*)| |x|) dP_x \]

\[ + \int_{C^c(\theta_0)} E(|\epsilon| |x|) dP_x > E_{\theta^*} |y - f(x, \theta^*)| . \]

Hence, \( \theta = \theta^* \) is the unique value in \( \Omega \) such that

\[ \min_{\theta \in \Omega} E_{\theta^*} |Q(x, y, \theta)| = E_{\theta^*} |y - f(x, \theta^*)| \]

and conclusion 2 follows.

The case of \( p > 1 \).

For each \( x \), let \( \Delta_x(\theta) = (f(x, \theta) - f(x, \theta^*)) \). Then \( \Delta_x(\theta) \) is continuous over \( \Omega \), \( \Delta_x(\theta) : \Omega \to D \subset R \). By Assumption 2 and 18, \( D \) is a compact connected subset in \( R \). For each \( x \), consider

\[ \rho(x, y, \theta) = |y - f(x, \theta)|^p = |y - f(x, \theta^*) - \Delta|^p \]
as a function of \( y \) and \( \Delta \), and let

\[
E_x(\Delta) = E(p_x(y, \Delta) | x) = E(|y - f(x, \theta^*) - \Delta|^p | x)
\]

Assumptions 4 and 8 guarantee that we can interchange the order of integration and differentiation and hence differentiating with respect to \( \Delta \),

\[
\frac{\partial E(|y - f(x, \theta^*) - \Delta|^p | x)}{\partial \Delta} = \frac{\partial (|y - f(x, \theta^*)|^p | x)}{\partial \Delta}
\]

where

\[
\frac{\partial |y - f(x, \theta^*) - \Delta|^p}{\partial \Delta} = \begin{cases} 
-p(y - f(x, \theta^*) - \Delta)^{p-1} & y - f(x, \theta^*) > \Delta \\
0 & y - f(x, \theta^*) = \Delta \\
p(\Delta - (y - f(x, \theta^*)))^{p-1} & y - f(x, \theta^*) < \Delta 
\end{cases}
\]

Note that the function \( E_x(\Delta) \geq 0 \), is continuous in \( \Delta \) for each \( x \) and by compactness of \( D \), \( E_x(\Delta) \) attains the minimum at some value \( \Delta_0 \) in \( D \) (depending on \( x \)). For a fixed \( x \) and each \( (y - f(x, \theta^*)) \), by connectedness of \( D \), \( p_x(\Delta) \) is a convex function of \( \Delta \), for all \( p > 1 \) and hence \( E_x(\Delta) \) is a convex function of \( \Delta \). This and Assumption 16 imply:

\[
\frac{\partial E_x(\Delta)}{\partial \Delta} = \begin{cases} 
< 0 & \Delta < 0 \\
= 0 & \Delta = 0 \\
> 0 & \Delta > 0 
\end{cases}
\]

for all \( \Delta \) in \( \text{int}(D) \) and all \( x \).
We know that $\Delta = 0$ is attained at $\theta = \theta^*$ in the interior of $\Omega$ for all $\mathbf{x}$, and hence $\Delta = 0$ is in the interior of $D$ and

$$E_\mathbf{x}(0) < E_\mathbf{x}(\Delta)$$

for any $\Delta$ in $\text{int}(D)$. Consider $\Delta_1^I$, $\Delta_2^I$, boundary points of $D$. The function $E_\mathbf{x}(\Delta)$ is continuous in $\Delta$ and thus, $E_\mathbf{x}(\Delta_{n_k}) \uparrow E(\Delta_1^I)$ for $\Delta_1^I > 0$ and $E_\mathbf{x}(\Delta_{n_k}) \downarrow E(\Delta_2^I)$ for $\Delta_2^I < 0$ and $\Delta_{n_k} \rightarrow \Delta_1^I$, $\Delta_{n_k} \rightarrow \Delta_2^I$ any convergent subsequence. Hence, for all $n_k \geq n_{k_0}$ and $n \geq n_{k_0}$,

$$E_\mathbf{x}(0) < E_\mathbf{x}(\Delta_{n_k}) \leq E_\mathbf{x}(\Delta_1^I)$$

and

$$E_\mathbf{x}(0) < E_\mathbf{x}(\Delta_{n_k}) \leq E_\mathbf{x}(\Delta_2^I).$$

Thus, $\Delta^*_\mathbf{x} = 0$ is the unique value of $\Delta$ in $D$ such that $E_\mathbf{x}(\Delta)$ attains its minimum at $\Delta_0 = 0$ for all $\mathbf{x}$. Note that the value of $\Delta$ minimizing $E_\mathbf{x}(\Delta)$ is not a function of $\mathbf{x}$ nor of the choice of $p > 1$.

Hence,

$$E(|y - f(\mathbf{x}, \theta^*) - \Delta_0|^P|\mathbf{x}) < E(|y - f(\mathbf{x}, \theta^*) - \Delta|^P|\mathbf{x})$$

for $\Delta \neq \Delta_0$, $\Delta$ in $D$ and $\Delta_0 = 0$.

We will now show that $\theta^*$ is the only value in $\Omega$ minimizing

$$E_{\theta^*}|y - f(\mathbf{x}, \theta^*) - (f(\mathbf{x}, \theta) - f(\mathbf{x}, \theta^*))|^P$$

over $\Omega$. 

Similar to the case of $p = 1$, let $\theta_0 \neq \theta^*$ and

$$
E_{\theta^*}|y - f(x, \theta^*) - \Delta(\theta_0)|^P = E_{\theta^*}|y - f(x, \theta^*)|^P.
$$

Then, by Assumption 17

$$
\int_{C(\theta_0)} E(|y - f(x, \theta^*) - \Delta(\theta_0)|^P|x) dP_x
$$

$$
> E_{\theta^*}|y - f(x, \theta^*)|^P
$$

which is a contradiction unless $u_x(C(\theta_0)) = 0$ and hence $\theta_0 = \theta^*$. Thus, conclusion 3 follows.

**Theorem 5.1**

Let $\hat{\theta}_n$ be the estimator of $\theta^*$, where

$$y_t - f(x_t, \theta^*) = \epsilon_t$$

for all sample vectors $\{(x_t, y_t)\}_{t=1}^n$ and $n = 1, 2, \ldots$. Let the conditions of Lemma 5.1 and Assumption 14 be satisfied, then with probability one

$$\hat{\theta}_n \rightarrow \theta^*$$

as $n \rightarrow \infty$. 
Proof

The conclusion of Lemma 5.1 says that at $\theta^*$ in the interior of $\Omega$

$$E_{\theta^*}|y - f(x, \theta^*)|^p < E_{\theta^*}|y - f(x, \theta)|^p$$

for all $\theta \neq \theta^*$, $\theta$ in $\Omega$. Hence Assumption 7 of Theorem 4.1 is satisfied. The rest of the Assumptions of Theorem 4.1 are satisfied by the conditions of Lemma 5.1 and Assumption 14. Hence, the conclusion of Theorem 4.1 follows. 

5.2 Asymptotic Normality

In this section we shall be concerned with showing that the conditions of Theorem 4.3 are satisfied whenever

$$Q(x_t, y_t, \theta^*) = y_t - f(x_t, \theta^*) = \varepsilon_t,$$

and specific conditions are imposed on the distribution function of $\varepsilon_t$. In the previous section it was required for this distribution to be in the family of symmetric distributions with a unique median, and hence the strong consistency of $\hat{\theta}_n$ was derived. Assumption 19 will add an additional constraint on this family.

Some of the expressions derived in Chapter 4 are presented below for the specific regression model.

Consider the partial derivatives of $p(x, y, \theta)$ with respect to $\theta$. Then,
\[ a | Q_t(\theta) |^p \frac{\partial}{\partial \theta_i} = \left\{ \begin{array}{ll} -p(y_t - f_t(\theta))^p - 1 \frac{\partial f_t(\theta)}{\partial \theta_i} & y_t - f_t(\theta) > 0 \\ p(f_t(\theta) - y_t)^p - 1 \frac{\partial f_t(\theta)}{\partial \theta_i} & y_t - f_t(\theta) < 0 \\ 0 & y_t - f_t(\theta) = 0 \end{array} \right. \]

where \( f_t(\theta) \) denotes \( f(x_t, \theta) \).

Equivalently,

\[ \frac{\partial}{\partial \theta_i} \frac{\partial | Q_t(\theta) |^p}{\partial \theta} = \left\{ \begin{array}{ll} -p|y_t - f_t(\theta)|^{p-2}(y_t - f_t(\theta)) \frac{\partial f_t(\theta)}{\partial \theta_i} & y_t - f_t(\theta) \neq 0 \\ 0 & y_t - f_t(\theta) = 0 \end{array} \right. \]

for \( i = 1, 2, \ldots, k \) and \( t = 1, 2, \ldots, n \).

The general expression for \( \frac{\partial^2 \rho_t(\theta)}{\partial \theta_i \partial \theta_j} \) derived in Chapter 3 becomes,
\[
\frac{\partial^2 Q_t(\theta)}{\partial \theta_i \partial \theta_j} = \begin{cases} 
\begin{align*}
p(p-1)(y_t - f_t(\theta))^{p-2} & \frac{\partial f_t(\theta)}{\partial \theta_i} \frac{\partial f_t(\theta)}{\partial \theta_j} \\
- p(y_t - f_t(\theta))^{p-1} & \frac{\partial^2 f_t(\theta)}{\partial \theta_i \partial \theta_j}
\end{align*}
\end{cases}
\]

\[\begin{align*}
t & - f_t(\theta) > 0 \\
(1) & - f_t(\theta) < 0 \\
0 & - f_t(\theta) = 0 \\
p & \geq 2 \\
\infty & - f_t(\theta) = 0 \\
1 & < p < 2 \\
\frac{\partial f_t(\theta)}{\partial \theta_i} - \frac{\partial f_t(\theta)}{\partial \theta_j} & \neq 0
\]

or
for $i, j = 1, 2, \ldots, k$ and $t = 1, 2, \ldots, n$.

Note that the following results hold for the regression model:

1. Assumptions 8 and 20 imply Assumption 9 for all $p > 1$.

2. Assumptions 8 and 20 imply Assumption 20(a) for $1 < p \leq 2$.

3. Assumptions 16, 19, and 20 imply Assumption 10 for all $p > 1$. 
4. Assumptions 1, 2, 4, 8, 16, 17, 18 imply

Assumption 7 for all \( p \geq 1 \).

The next result will be presented in Lemma 5.2. Under the conditions specified by the next lemma, the regression model satisfies the conditions of Theorem 4.3 and asymptotic normality of \( \hat{\theta}_n - \theta^* \) for \( 1 < p < 2 \) follows.

**Lemma 5.2**

Let Assumptions 1, 2, 3, 4, 5, 8, 10, 14, 16, 17, 19, 20, 20a, 2c, and 20b be satisfied, and let

\[
Q(x_t, y_t, \theta^*) = y_t - f(x_t, \theta^*)
\]

then except almost everywhere \((x, y)\), there exists an \( \varepsilon > 0 \), and a measurable function of \((x, y)\), \( m^i(x, y, \delta) \), for each \( i = 1, 2, \ldots, k \), and each \( 0 < \delta < \varepsilon \), such that

1. \[
| \frac{\partial p(x_t, y_t, \theta)}{\partial \theta_i} - \frac{\partial p(x_t, y_t, \theta^*)}{\partial \theta_i} | \leq \frac{k}{\delta} \sum_{j=1}^{\delta^2} m^i(x_t, y_t, \delta) \left| \theta_j - \theta_j^* \right| \]

\[
\leq m^i(x_t, y_t, \delta) \| \theta - \theta^* \|
\]

for \( i, j = 1, 2, \ldots, k \) and all \( \theta \) in \( A(\theta^*, \delta) \).

2. \[
\frac{1}{n} \sum_{t=1}^{n} m^i(x_t, y_t, \delta) \rightarrow 0 \quad \text{with probability one as} \quad n \rightarrow \infty
\]

and \( \delta \rightarrow 0 \).
Proof

Consider

\[
\frac{\partial^2 \phi_t(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} - \sum_{j=1}^{k} \frac{\partial^2 \phi_t(\boldsymbol{\theta}^*)}{\partial \theta_i \partial \theta_j} (\theta_j - \theta_j^*)
\]

\[
= p \left( |y_t - f_t(\theta^*)|^{p-2} (y_t - f_t(\theta^*)) \frac{\partial f_t(\theta^*)}{\partial \theta_i} \right)
\]

\[
- |y_t - f_t(\theta)|^{p-2} (y_t - f_t(\theta)) \frac{\partial f_t(\theta)}{\partial \theta_i}
\]

\[
- (p-1) \sum_{j=1}^{k} \left[ |y_t - f_t(\theta^*)|^{p-2} (y_t - f_t(\theta^*)) \frac{\partial f_t(\theta^*)}{\partial \theta_j} \frac{\partial f_t(\theta^*)}{\partial \theta_i} \right]
\]

\[
+ |y - f_t(\theta^*)|^{p-2} (y_t - f_t(\theta^*)) \frac{\partial^2 f_t(\theta^*)}{\partial \theta_j \partial \theta_i} (\theta_j - \theta_j^*)
\]

for \( i, j = 1, 2, \ldots, k; \ t = 1, 2, \ldots, n \) and all \( \theta \) in \( A(\theta^*, \delta) \).

In order to eliminate the absolute value signs we will consider firstly two general cases:

**Case I**

\( y_t - f(x_t, \theta^*) > 0 \) and all possible values of \( (y_t - f(x_t, \theta)) \)

for \( \theta \) in \( \Omega \),

**Case II**

\( y_t - f(x_t, \theta^*) < 0 \) and all possible values of \( (y_t - f(x_t, \theta)) \).

The goal is to define a bounding function of

\[
\left| \frac{\partial \phi_t(\boldsymbol{\theta})}{\partial \theta_i} - \frac{\partial \phi_t(\boldsymbol{\theta}^*)}{\partial \theta_i} \right|
\]

of the form
for almost all \((x_t, y_t)\), and such that for some \(\varepsilon > 0\), and each
\[
0 < \delta < \varepsilon \quad \text{and} \quad \Theta \in A(\theta^*, \delta),
\]

\[
\left| \frac{\partial \rho_t(\Theta)}{\partial \theta_i} - \frac{\partial \rho_t(\theta^*)}{\partial \theta_i} \right| \leq h_i(x_t, y_t, \delta) \|\Theta - \Theta^*\|.
\]

In all cases we consider only the values \(1 < p \leq 2\). Then, \(0 < (p-1) \leq 1\) and \((p-2) \leq 0\). In the following derivations \(f(x_t, \Theta)\) and \(f(\Theta)\) denote the same value of the function. We will not use the subscript \(t\) to distinguish the observations.

**Case I**

\[
y - f(x, \Theta^*) > 0
\]

1. \(y > f(x, \Theta) \geq f(x, \Theta^*)\)

(a) \((y - f(x, \Theta))^{p-1} \frac{\partial f(\Theta)}{\partial \theta_i} \geq (y - f(x, \Theta^*)) \frac{\partial f(\Theta^*)}{\partial \theta_i}\),

where \(\frac{\partial f(\Theta)}{\partial \theta_i} \leq 0\) and \(\frac{\partial f(\Theta^*)}{\partial \theta_i} < 0\). Note that for Case I since \(1 < p \leq 2\),

\[
0 < (y - f(x, \Theta))^{p-1} \leq (y - f(x, \Theta^*))^{p-1}
\]

and

\[
(y - f(x, \Theta))^{p-2} \geq (y - f(x, \Theta^*))^{p-2}.
\]
\[ \left| \frac{\partial p(\theta)}{\partial \theta_i} - \frac{\partial p(\theta^*)}{\partial \theta_i} \right| = p(y - f(\theta))^{p-1} \frac{\partial f(\theta)}{\partial \theta_i} \]

- \( p(y - f(\theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta_i} \)

= \( p(y - f(\theta))^{p-2}(y - f(\theta)) \frac{\partial f(\theta)}{\partial \theta_i} \)

- \( p(y - f(\theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta_i} \)

\leq p |y - f(\theta^*)|^{p-2} \left[ (y - f(\theta)) \frac{\partial f(\theta)}{\partial \theta_i} \right.

- \( (y - f(\theta^*)) \frac{\partial f(\theta^*)}{\partial \theta_i} \)

\leq p |y - f(\theta^*)|^{p-2} \left[ (y - f(\theta^*)) \left( \frac{\partial f(\theta)}{\partial \theta_i} - \frac{\partial f(\theta^*)}{\partial \theta_i} \right) \right.

+ \frac{\partial f(\theta)}{\partial \theta_i} \left( f(\theta^*) - f(\theta) \right) \right].

(b) \( (y - f(x, \theta))^{p-1} \frac{\partial f(\theta)}{\partial \theta_i} \geq y - f(x, \theta) \frac{\partial f(\theta^*)}{\partial \theta_i} \),

where

\[ \frac{\partial f(x)}{\partial \theta_i} > 0, \quad \frac{\partial f(\theta^*)}{\partial \theta_i} \geq 0. \]

\[ \left| \frac{\partial p(\theta)}{\partial \theta_i} - \frac{\partial p(\theta^*)}{\partial \theta_i} \right| = p(y - f(\theta))^{p-1} \frac{\partial f(\theta)}{\partial \theta_i} \]

- \( p(y - f(\theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta_i} \)

\leq p |y - f(\theta^*)|^{p-1} \left[ \frac{\partial f(\theta)}{\partial \theta_i} - \frac{\partial f(\theta^*)}{\partial \theta_i} \right].
(c) \( (y - f(x,\theta))^{p-1} \frac{\partial f(\theta)}{\partial \theta_1} \leq (y - f(x,\theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta_1} \),

where

\[
\frac{\partial f(\theta)}{\partial \theta_1} < 0, \quad \frac{\partial f(\theta^*)}{\partial \theta_1} \leq 0.
\]

Note that this case is equivalent to the preceding case and hence

\[
|\frac{\partial p(\theta)}{\partial \theta_1} - \frac{\partial p(\theta^*)}{\partial \theta_1}| \leq p|y - f(\theta^*)|^{p-1}[\frac{\partial f(\theta^*)}{\partial \theta_1} - \frac{\partial f(\theta)}{\partial \theta_1}].
\]

Cases (b) and (c) do not occur whenever \( \partial f(x,\theta)/\partial \theta \) is a constant function of \( \theta \).

(d) \( (y - f(x,\theta))^{p-1} \frac{\partial f(\theta)}{\partial \theta_1} \leq (y - f(x,\theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta_1} \),

where

\[
\frac{\partial f(\theta)}{\partial \theta_1} \geq 0, \quad \frac{\partial f(\theta^*)}{\partial \theta_1} > 0.
\]

Cases (a) and (d) are equivalent. Therefore,

\[
|\frac{\partial p(\theta)}{\partial \theta_1} - \frac{\partial p(\theta^*)}{\partial \theta_1}| \leq p|y - f(\theta^*)|^{p-2}[(y - f(\theta^*))(\frac{\partial f(\theta^*)}{\partial \theta_1} - \frac{\partial f(\theta)}{\partial \theta_1})

+ \frac{\partial f(\theta)}{\partial \theta_1} (f(\theta) - f(\theta^*))].
\]

2. \( y > f(x,\theta^*) > f(x,\theta) \)

The above relation implies
\[ 0 < (y - f(x, \theta^*))^{p-1} < (y - f(x, \theta))^{p-1} \]

and

\[ 0 < (y - f(x, \theta))^{p-2} < (y - f(x, \theta^*))^{p-2} , \]

(a) \( (y - f(x, \theta))^{p-1} \frac{\partial f(\theta)}{\partial \theta} \geq (y - f(x, \theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta} \)

where

\[ \frac{\partial f(\theta)}{\partial \theta} \leq 0 , \quad \frac{\partial f(\theta^*)}{\partial \theta} < 0 . \]

This case corresponds to 1(c) with \( \theta^* \) and \( \theta \) interchanged in

\[ \left( \frac{\partial f(\theta^*)}{\partial \theta} - \frac{\partial f(\theta)}{\partial \theta} \right) . \]

Hence,

\[ |\frac{\partial f(\theta)}{\partial \theta} - \frac{\partial f(\theta^*)}{\partial \theta}| < p |y - f(\theta^*)|^p \left( \frac{\partial f(\theta)}{\partial \theta} - \frac{\partial f(\theta^*)}{\partial \theta} \right) . \]

(b) \( (y - f(x, \theta))^{p-1} \frac{\partial f(\theta)}{\partial \theta} \geq (y - f(x, \theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta} \)

where

\[ \frac{\partial f(\theta)}{\partial \theta} > 0 , \quad \frac{\partial f(\theta^*)}{\partial \theta} > 0 . \]

Interchange \( \theta^* \) and \( \theta \) in the terms

\[ \left( \frac{\partial f(\theta^*)}{\partial \theta} - \frac{\partial f(\theta)}{\partial \theta} \right) \quad \text{and} \quad (f(\theta^*) - f(\theta)) \]

of Case 1(d).
\[
\left| \frac{\partial p(\theta)}{\partial \theta_i} - \frac{\partial p(\theta^*)}{\partial \theta_i} \right| < p \left| y - f(\theta^*) \right|^{p-2} \left( y - f(\theta^*) \right) \left( \frac{\partial f(\theta^*)}{\partial \theta_i} - \frac{\partial f(\theta)}{\partial \theta_i} \right) \\
+ (f(\theta^*) - f(\theta)) \frac{\partial f(\theta)}{\partial \theta_i} \right].
\]

(c) \((y - f(x, \theta))^p \frac{\partial f(\theta)}{\partial \theta_i} \leq (y - f(x, \theta^*))^p \frac{\partial f(\theta^*)}{\partial \theta_i} \),

where
\[
\frac{\partial f(\theta)}{\partial \theta_i} < 0, \quad \frac{\partial f(\theta^*)}{\partial \theta_i} < 0.
\]

Similar to 1(a) and interchanging \(\theta^*\) and \(\theta'\) we get
\[
\left| \frac{\partial p(\theta)}{\partial \theta_i} - \frac{\partial p(\theta^*)}{\partial \theta_i} \right| < p \left| y - f(\theta^*) \right|^{p-2} \left( y - f(\theta^*) \right) \left( \frac{\partial f(\theta^*)}{\partial \theta_i} - \frac{\partial f(\theta)}{\partial \theta_i} \right) \\
+ (f(\theta^*) - f(\theta)) \frac{\partial f(\theta)}{\partial \theta_i} \right].
\]

(d) \((y - f(x, \theta))^p \frac{\partial f(\theta)}{\partial \theta_i} \leq (y - f(x, \theta^*))^p \frac{\partial f(\theta^*)}{\partial \theta_i} \),

where
\[
\frac{\partial f(\theta)}{\partial \theta_i} > 0, \quad \frac{\partial f(\theta^*)}{\partial \theta_i} \leq 0.
\]

By 2(a) we obtain,
\[
\left| \frac{\partial p(\theta)}{\partial \theta_i} - \frac{\partial p(\theta^*)}{\partial \theta_i} \right| < p \left| y - f(\theta^*) \right|^{p-2} \left( \frac{\partial f(\theta^*)}{\partial \theta_i} - \frac{\partial f(\theta)}{\partial \theta_i} \right) .
\]
Note that Case 2(a) and 2(d) do not occur when $\frac{\partial f(\theta)}{\partial \theta}$ is a constant function of $\theta$.

Only the cases where $\frac{\partial f(\theta)}{\partial \theta_1}$ and $\frac{\partial f(\theta^*)}{\partial \theta_1}$ have the same sign (if both are different from zero) have been considered. Under 1 and 2 above, it is possible that for some values of $\theta$, these derivatives have opposite signs. The only applicable cases then are as follows:

1. $y > f(x, \theta) \geq f(x, \theta^*)$

   (a) $(y - f(x, \theta))^{p-1} \frac{\partial f(\theta)}{\partial \theta_1} \geq (y - f(x, \theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta_1},$

   where

   $$\frac{\partial f(\theta^*)}{\partial \theta_1} \leq 0, \quad \frac{\partial f(\theta)}{\partial \theta_1} \geq 0.$$

   $$|\frac{\partial f(\theta)}{\partial \theta_1} - \frac{\partial f(\theta^*)}{\partial \theta_1}| = p(y - f(\theta))^{p-1} \frac{\partial f(\theta)}{\partial \theta_1}$$

   $$- p(y - f(\theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta_1}$$

   $$< p(y - f(\theta^*))^{p-1} (\frac{\partial f(\theta)}{\partial \theta_1} - \frac{\partial f(\theta^*)}{\partial \theta_1}).$$

   (b) $(y - f(x, \theta))^{p-1} \frac{\partial f(\theta)}{\partial \theta_1} \leq (y - f(x, \theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta_1},$

   where

   $$\frac{\partial f(\theta)}{\partial \theta_1} \leq 0, \quad \frac{\partial f(\theta^*)}{\partial \theta_1} \geq 0.$$
This is equivalent to the case above. Hence,

\[
|\frac{\partial p(\theta)}{\partial \theta_1} - \frac{\partial p(\theta^*)}{\partial \theta_1}| \leq p(y - f(\theta^*))^{p-1}(\frac{\partial f(\theta^*)}{\partial \theta_1} - \frac{\partial f(\theta)}{\partial \theta_1}) .
\]

2. \( y > f(\theta^*, \theta) > f(\theta) \).

(a) \((y - f(\theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta_1} \leq (y - f(\theta)) \frac{\partial f(\theta)}{\partial \theta_1} ,
\]

where

\[
\frac{\partial f(\theta^*)}{\partial \theta_1} \leq 0 , \quad \frac{\partial f(\theta)}{\partial \theta_1} \geq 0 .
\]

(b) \((y - f(\theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta_1} \geq (y - f(\theta))^{p-1} \frac{\partial f(\theta)}{\partial \theta_1} ,
\]

where
Similarly to 2(a) we obtain,

$$
\left| \frac{\partial f(x)}{\partial \theta} (\theta^*) - \frac{\partial f(x)}{\partial \theta} (\theta) \right| < p(y - f(\theta^*))^{p-2} (y - f(\theta^*)) \left( \frac{\partial f(\theta^*)}{\partial \theta} - \frac{\partial f(\theta)}{\partial \theta} \right) + \frac{\partial f(\theta)}{\partial \theta} (f(\theta^*) - f(\theta))
$$

3. \( f(x, \theta^*) < y \leq f(x, \theta) \).

Hence,

$$
0 < f(x, \theta) - y < f(x, \theta) - f(x, \theta^*)
$$

and

$$
0 < y - f(x, \theta^*) \leq f(x, \theta) - f(x, \theta^*)
$$

$$
\left| \frac{\partial f(x)}{\partial \theta} (\theta^*) - \frac{\partial f(x)}{\partial \theta} (\theta) \right| = p(y - f(\theta^*))^{p-1} \frac{\partial f(\theta^*)}{\partial \theta} (f(\theta) - y)^{p-1} \frac{\partial f(\theta)}{\partial \theta} + (f(\theta) - y)^{p-1} \frac{\partial f(\theta)}{\partial \theta}
\leq p|f(\theta) - y|^{p-1}\left| \frac{\partial f(\theta)}{\partial \theta} \right| + p|y - f(\theta^*)|^{p-1}\left| \frac{\partial f(\theta^*)}{\partial \theta} \right|
\leq p|f(\theta) - f(\theta^*)|^{p-1}\left| \frac{\partial f(\theta)}{\partial \theta} \right|
$$
Hence, for each \((x_t, \theta)\) in \(X \times \Omega\) we can define a function 

\[ h_i(x_t, y_t, \theta^*) \]

which bounds 

\[ \left| \frac{\partial \phi(\theta) - \partial \phi(\theta^*)}{\partial \theta_i} \right| \]

for \(i = 1, 2, \ldots, k\) and all \((x, y)\) such that \(y > f(x, \theta^*)\).

Similarly, we can derive the equivalent cases for \(y < f(x, \theta^*)\).

That is,

**Case 1:** \(y < f(x, \theta) \leq f(x, \theta^*)\)

**Case 2:** \(y < f(x, \theta^*) < f(x, \theta)\).

**Case 3:** \(f(x, \theta) < y < f(x, \theta^*)\).

For all \((x_t, y_t)\) such that \(y_t \neq f(x_t, \theta^*)\), then

\[ h_i(x, y, \theta) = 2|y - f(\theta^*)|^{p-2} \left[ |y - f(\theta^*)| \left| \frac{\partial^2 f(\theta)}{\partial \theta_i^2} \right| \right. \]

\[ + \left. |f(\theta) - f(\theta^*)| \left| \frac{\partial f(\theta)}{\partial \theta_i} \right| \right] \]
for $i = 1, 2, \ldots, k$.

Consider the set $\{\theta' | \theta \in A(\theta^*, \delta)\}$ where $0 < \delta < \varepsilon$ and $A(\theta^*, \delta)$ is in the interior of $\Omega$. Then, by the Mean Value Theorem, (through Assumptions 4 and 5) for all $\theta$ in $A(\theta^*, \delta)$ it follows that

$$|f(x, \theta) - f(x, \theta^*)| \leq \sup_{A(\theta^*, \delta)} \left( \frac{\partial f(x, \theta)}{\partial \theta} \right) \|\theta - \theta^*\|$$

and

$$\left| \frac{\partial f(x, \theta)}{\partial \theta_1} - \frac{\partial f(x, \theta^*)}{\partial \theta_1} \right| \leq \sup_{A(\theta^*, \delta)} \left( \frac{\partial^2 f(x, \theta)}{\partial \theta \partial \theta_1} \right) \|\theta - \theta^*\|.$$

Hence, for each $0 < \delta < \varepsilon$ and all $\theta$ in $A(\theta^*, \delta)$

$$h_1(x, y, \delta) = 6|y - f(\theta^*)|^{p-2} \sup_{A(\theta^*, \delta)} \left( \frac{\partial f(\theta)}{\partial \theta} \right) \|\theta - \theta^*\| + \sup_{A(\theta^*, \delta)} \left( \frac{\partial^2 f(\theta)}{\partial \theta \partial \theta_1} \right) \|\theta - \theta^*\| + 2|y - f(\theta^*)|^{p-1} \sup_{A(\theta^*, \delta)} \left( \frac{\partial^2 f(\theta)}{\partial \theta \partial \theta_1} \right) \|\theta - \theta^*\|.$$

Let

$$L_{11}(x, \delta) = \sup_{A(\theta^*, \delta)} \left( \frac{\partial f(\theta)}{\partial \theta} \right) \|\theta - \theta^*\| \cdot \sup_{A(\theta^*, \delta)} \left( \frac{\partial^2 f(\theta)}{\partial \theta \partial \theta_1} \right) \|\theta - \theta^*\|,$$
Hence, with probability one, by Assumption 2, whenever \( \theta \) is \( \Lambda(\theta^*, \delta) \)

\[
|\frac{\partial^2 f(\theta)}{\partial \theta^2} - \frac{\partial^2 f(\theta^*)}{\partial \theta^2} - \sum_{j=1}^{k} \frac{\partial^2 f(\theta_j)}{\partial \theta_j^2} (\theta_j - \theta_j^*)| 
\leq h(x, y, \delta)\|\theta - \theta^*\| + 2|y - f(\theta^*)|\bigg|\frac{\partial f(\theta^*)}{\partial \theta^2}\bigg| \cdot \|\theta - \theta^*\|. 
\]

Using this bounding function we can construct the function \( m(x, y, \delta) \) of the lemma, such that with probability one,

\[
\frac{1}{n} \sum_{t=1}^{n} m(x_t, y_t, \delta) \to E_{\theta^*} m(x, y, \delta) \to 0 
\]

as \( \delta \to 0 \) and \( n \to \infty \).

For each \( \delta \), consider the set

\[
S(\delta) = \{(x, y) | |y - f(\theta^*)| - 2 \max_{\Lambda(\theta^*, \delta)} |f(\theta)| - f(\theta^*)|, \delta| \leq 0\} . 
\]

Then on \( S^C(\delta) \),

\[
|y - f(x, \theta^*)| > 2 \max_{\Lambda(\theta^*, \delta)} |f(\theta) - f(\theta^*)|, \delta| = 2\Delta(x, \delta) 
\]

and hence on \( S^C(\delta) \),
\[ |y - f(\theta)| = |y - f(\theta^*) - (f(\theta) - f(\theta^*))| > \Delta(x, \delta) \]

for all \( \theta \) in \( A(\theta^*, \delta) \).

By Assumption 4, the second partials of \( r(x, y, \theta) \) with respect to \( \theta \), exist on \( S^C(\delta) \), and are real valued continuous functions of \( \theta \) for all \( \theta \) in \( A(\theta^*, \delta) \). Using a second order Taylor's approximation around \( \theta^* \), let \( \theta \) be in \( A(\theta^*, \delta) \) and let \((x_t, y_t)\) be in \( S^C(\delta) \), then

\[
\frac{\partial^2 r_t(\theta)}{\partial \theta_i \partial \theta_j} - \frac{\partial^2 r_t(\theta^*)}{\partial \theta_i \partial \theta_j} \leq \sum_{j=1}^{k} \frac{\partial^2 r_t(\tilde{\theta})}{\partial \theta_j \partial \theta_i} - \frac{\partial^2 r_t(\theta^*)}{\partial \theta_j \partial \theta_i} ||\theta - \theta^*||
\]

where \( \tilde{\theta} \) is in \( A(\theta^*, \delta) \) and

\[
\frac{\partial^2 r_t(\tilde{\theta})}{\partial \theta_j \partial \theta_i} - \frac{\partial^2 r_t(\theta^*)}{\partial \theta_j \partial \theta_i} < 2||y_t - f_t(\tilde{\theta})||^{p-2} \frac{\partial f_t(\tilde{\theta})}{\partial \theta_i} \frac{\partial f_t(\tilde{\theta})}{\partial \theta_j}
\]

\[
- ||y_t - f_t(\theta^*)||^{p-2} \frac{\partial f_t(\theta^*)}{\partial \theta_i} \frac{\partial f_t(\theta^*)}{\partial \theta_j}
\]

\[
+ 2||y_t - f_t(\theta^*)||^{p-2} (y_t - f_t(\theta^*)) \frac{\partial^2 f(\theta^*)}{\partial \theta_i \partial \theta_j}
\]

\[
- ||y_t - f_t(\theta^*)||^{p-2} (y_t - f_t(\tilde{\theta})) \frac{\partial^2 f(\tilde{\theta})}{\partial \theta_i \partial \theta_j}
\]

\[
= 2A_1(q, x, \delta) + 2A_2(q, x, \delta), \hspace{1cm} (5.3)
\]

where
For each $0 < \delta < \varepsilon$ and each $x$ consider the sets

$$S_1(x, \delta) = \{y - f(x, \theta^*) | (x, y) \in S^C(\delta)\},$$

and

$$S_2(x, \delta) = \{y - f(x, \theta^*) | (x, y) \in S^C(\delta)\},$$

where by Assumption 20(a)

$$\sup_{x} \left( \sup_{A(x, \varepsilon^*)} |f(x, \theta) - f(x, \theta^*)| \right) \leq T$$

and

$$T' = \max \{1, 2T\}.$$

Consider

$$S_1^+(x, \delta) = \{(y - f(x, \theta^*)) > 0 | (y - f(\theta^*)) \in S_1(x, \delta)\}.$$

Then, denoting the distribution function of $q$ by $F_q$, by Assumption 19

$$I_1 = \int_{R^m} \int_{S_1^+(x, \delta)} A_1(q, x, \theta) dF_q dP_x,$$

$$\leq M \int_{R^m} \int_{S_1^+(x, \delta)} ||q||^{p-2} - |q - (f(\bar{\theta})) - f(\theta^*)|^{p-2}$$

$$\times |\frac{\partial f(\theta^*)}{\partial \theta_i} \frac{\partial f(\theta^*)}{\partial \theta_j}| d\eta dP_x.$$
For convenience let \( \Delta(x, \delta) \) be denoted as \( \Delta \). By Assumption 20 for \( \theta \) in \( A(\theta^*, \delta) \) and on \( S_1^+(x, \delta) \) if \( (p - r) < 0 \)

\[
0 < |q - (f(\theta) - f(\theta^*))|^{p-r} \leq (q - \Delta)^{p-r}.
\]

Since \( (q - (f(\theta) - f(\theta^*)) \geq \delta > 0 \) on \( S_1^+(x, \delta) \) a first order Taylor's expansion around \( \theta^* \) gives for any \( \theta \) in \( A(\theta^*, \delta) \),

\[
||q - (f(\theta) - f(\theta^*))||^{p-2} - |q|^{p-2} \leq |p - 2||q - (f(\theta) - f(\theta^*))||^{p-3} \sum_{i,j} \frac{\partial^2 f(\theta)}{\partial \theta_i \partial \theta_j} ||\theta - \theta^*||.
\]

Hence,

\[
I_1 \leq \int_{R^m} M^i(\Delta^2 T_{1,p-2} - \Delta_{p-2}^2 \Sigma_{r,1,j}^i(x) + M((T^i)^{p-1} - \Delta_{p-1}^2) \sup_{A(\theta^*, \delta)} \frac{\partial f(\theta^*)}{\partial \theta_i} \frac{\partial f(\theta^*)}{\partial \theta_j} - \frac{\partial^2 f(\theta)}{\partial \theta_i \partial \theta_j} |dF_x|
\]

By the Dominated Convergence Theorem then \( I_1 \to 0 \) as \( \delta \to 0 \). The case for \( S_1^-(x, \delta) \) would be equivalent by Assumption 16.
Now consider the integral of $A_1(q, \theta, \theta_*)$ over the set $S_2(\theta, \delta)$. Call it $I_2$. Then, since

$$|y - f(\theta)| > 1$$
on $S_2(\theta, \delta)$, the second term in $I_2$ is bounded for all $x$ by an integrable function, and by the D.C.T., the integral of this term goes to zero as $\delta \to 0$. That is,

$$\int \int_{S_2(\theta, \delta)} |y - f(\theta^*) - (f(\theta)|$$

$$- f(\theta^*)|^{p-2} - f(\theta^*)$$

$$\frac{\partial f(\theta^*)}{\partial \theta_i} \frac{\partial f(\theta)}{\partial \theta_i}$$

$$\frac{df(\theta^*)}{d\theta_i} \frac{df(\theta)}{d\theta_i}$$

$$dF q dP_x$$

$$\to 0 \text{ as } \delta \to 0.$$  

By Assumption 20, the first term in $I_2$ is

$$\leq \int \int_{S_2(\theta, \delta)} |y - f(\theta^*)|^{p-2}$$

$$- |y - f(\theta)|^{p-2} dF q dP_x$$

and since for all $\theta$ in $A(\theta^*, \delta)$ the integrand is bounded by the integrable function $2r_{1, j}(\theta)$ and
\[ \int_{\mathbb{R}^m} r_{i,j}(x) \int_{S_2(x,\delta)} \left| y - f(\theta^*) \right|^p \, dy \]
Note that by Assumptions 5, 8, 20(b), and the D.C.T. the integral of the first term is

$$\leq \int_{\mathbb{R}^m} \int_{(-\infty, \infty)} |y - f(\theta^*)|^{P-1} \sup_{\lambda(\theta^*, \delta)} \left| \frac{\partial^2 f(\theta^*)}{\partial \theta_j \partial \theta_1} \right| \, dF \, dP_X$$

$$= \frac{\partial^2 f(\theta^*)}{\partial \theta_j \partial \theta_1} \left| \sup_{\lambda(\theta^*, \delta)} \right| \int_{\mathbb{R}^m} \int_{(-\infty, \infty)} |y - f(\theta^*)|^{P-1} \, dF \, dP_X \to 0$$

as $\delta \to 0$. Consider the second term bounded by

$$\int_{\mathbb{R}^m} \left( \sup_{\lambda(\theta^*, \delta)} \left| \frac{\partial^2 f(\theta^*)}{\partial \theta_j \partial \theta_1} \right| \right) \int_{(S_1 \cup S_2)} \left| (y - f(\theta)^{P-2} (y - f(\theta^*)) |dF \, dP_X \right.$$ 

$$\leq \int_{\mathbb{R}^m} \left( \sup_{\lambda(\theta^*, \delta)} \left| \frac{\partial^2 f(\theta^*)}{\partial \theta_j \partial \theta_1} \right| \right)$$

$$\left| \int_{(0, \infty)} |y - f(\theta^*)|^{P-1} - |y - f(\theta)|^{P-1} \right|$$

$$+ \int_{(-\infty, 0)} |y - f(\theta^*)|^{P-1}$$

$$- \left| |y - f(\theta^*)|^{P-1} \right| dF \, dP_X.$$

Similarly, this integral $\to 0$ as $\delta \to 0$. 
Define \( m^i(x, y, \delta) \) as follows:

\[
m^i(x, y, \delta) = \begin{cases}
6|y - f(x, \theta^*)|^2(L_{11}(x, \delta) + L_{12}(x, \delta)) & \text{if } 6|y - f(x, \theta^*)|^2L_{12}(x, \delta) \\
+ 2|y - f(x, \theta^*)|^2L_{12}(x, \delta) & \text{if } 2|y - f(x, \theta^*)|^2L_{12}(x, \delta) \\
\end{cases}
\]

\[
m^i(x, y, \delta) = \begin{cases}
|y - f(x, \theta^*)| \leq 2 \max(\sup_{A(\theta^*, \delta)} |f(x, \theta) - f(x, \theta^*)|, \delta) & \text{if } \sup_{A(\theta^*, \delta)} \sum_{1}^{k} \left| \frac{\partial^2 \rho(x, y, \theta)}{\partial \theta_j \partial \theta_1} - \frac{\partial^2 \rho(x, y, \theta^*)}{\partial \theta_j \partial \theta_1} \right| \\
|y - f(x, \theta^*)| > 2 \max(\sup_{A(\theta^*, \delta)} |f(x, \theta) - f(x, \theta^*)|, \delta) & \text{if } \end{cases}
\]

Then,

\[
E_{\theta^*} m^i(x, y, \delta) \to 0 \text{ as } \delta \to 0,
\]

and, by Assumption 14,

\[
\frac{1}{n} \sum_{t=1}^{n} m^i(x_t, y_t, \delta) \to E_{\theta^*} m^i(x, y, \delta)
\]

with probability one as \( n \to \infty \).

**Theorem 5.2**

Let \( \hat{\theta}_n \) be a consistent estimator of \( \theta^* \), with \( 1 < p \leq 2 \), let the conditions of Lemma 5.2 be satisfied, and let the conclusion of Lemma 5.1 and Assumptions 12 and 15 hold, then if

\[
Q(x_t, y_t, \theta^*) = y_t - f(x_t, \theta^*) = \epsilon_t,
\]
\[ \sqrt{n} (\hat{\theta}_n - \theta^*) \xrightarrow{L} N(0, G^{-1}(\theta^*) \Lambda(\theta^*) G^{-1}(\theta^*)). \]

**Proof**

The conclusion follows from Theorem 4.3. It is easy to verify that the conditions of this theorem are all satisfied. 

The condition to verify for \( p > 2 \) is relation (4.4). That is,

\[
\sup_{\theta - \theta^*} \left| \frac{\partial^2 \rho(x, y, \theta)}{\partial \theta_j \partial \theta_i} - \frac{\partial^2 \rho(x, y, \theta^*)}{\partial \theta_j \partial \theta_i} \right| \leq h_{i,j}(x, y, \delta)
\]

where

\[
E_{\theta^*} h_{i,j}(x, y, \delta) \to 0 \text{ as } \delta \to 0.
\]

Note that in this case \( 0 < p-1 < p \) and \( 0 < p-2 < p \). Assumption 8 implies that

\[
E_{\theta^*} |y - f(\theta)|^{p-2} < \infty
\]

and

\[
E_{\theta^*} |y - f(\theta)|^{p-1} < \infty
\]

for all \( \theta \) in \( \Omega \). This, and Assumptions 5, 20, and 20(b), allow for the conclusion of the D.C.T. to apply and relation (4.4) is satisfied.
6. SUMMARY

The statistical properties of the estimator \( \hat{\theta}_n \) were investigated, where \( \hat{\theta}_n \) satisfies

\[
\min_{\theta \in \Omega} \sum_{t=1}^{n} |Q(x_t, y_t, \theta)|^p = \sum_{t=1}^{n} |Q(x_t, y_t, \hat{\theta}_n)|^p
\]

for a choice of \( p \geq 1 \), and

\[
Q(x_t, y_t, \theta^*) = e_t
\]
is the hypothesized implicit model. An algorithm to compute \( \hat{\theta}_n \) was derived and under general conditions it was shown to converge to a stationary point of the objective function.

The implicit model was more specifically defined by placing restrictions on the function \( Q \) and the parameter space \( \Omega \), the arguments \( (x_t, y_t) \), \( t = 1, 2, \ldots \), and the behavior of the random error. Given the complete model specification and a set of assumptions to be fulfilled by the model it was shown that for any \( p \geq 1 \), \( \hat{\theta}_n \) converges with probability one to a point \( \theta_0 \) in the interior of \( \Omega \) and that for \( p > 1 \), \( \sqrt{n}(\hat{\theta}_n - \theta_0) \) is asymptotically \( k \) variate normal. The point \( \theta_0 \) is defined as that value in \( \Omega \) such that for any \( \theta \neq \theta_0 \), \( \theta \) in \( \Omega \)

\[
E_{\theta^*} |Q(x, y, \theta_0)|^p < E_{\theta^*} |Q(x, y, \theta)|^p.
\]

Moreover, when more than one point in \( \Omega \) minimizes the above expectation, the estimate \( \hat{\theta}_n \) ultimately remains in any neighborhood of these points.
When the model can be solved for the endogenous variable $y$, through analytical or numerical procedures, and equating $e_t$ to zero, a predictor $\hat{y}_n$ is defined for a fixed value of $x$. Under rather general conditions $\hat{y}_n$ exists as a continuous function $\phi_x(\hat{\theta}_n)$, and if $\hat{\theta}_n$ converges to $\theta_0$ with probability one and has an asymptotically normal distribution with mean $\theta_0$, then it follows that $\hat{y}_n$ converges in probability to $\phi_x(\theta_0)$, and that $\sqrt{n}(\hat{y}_n - \phi_x(\theta_0))$ converges in distribution to the univariate normal with mean zero.

Finally, the regression model was considered as a special case of the general implicit model. The model assumptions and additional conditions on a class of distribution functions for the error, and on the parameter space, was sufficient to show that in this case $\theta_0$ corresponds to the true parameter value $\theta^*$ and, hence, the findings presented in the previous paragraphs follow for $\theta_0 = \theta^*$. 
7. NUMERICAL ILLUSTRATIONS

The examples in this chapter were solved using the computing technique developed in Chapter 3. The purpose was to assess the performance of the algorithm and when possible to compare the results with those generated by other computing methods.

The following example was used by Schlossmacher [19] in evaluating a proposed technique for minimizing the sum of absolute values of residuals from a linear regression model.

Given the data

\[
\begin{array}{cccccccc}
    x_t & 12 & 18 & 24 & 30 & 36 & 42 & 48 \\
    y_t & 5.27 & 5.68 & 6.25 & 7.21 & 8.02 & 8.71 & 8.42 \\
\end{array}
\]

the optimal fit using the minimum absolute deviations criteria was

\[ y_t = 3.95 + 0.1098x_t. \]

However, as commented in [19], this example has a non-unique optimal region. In particular, lines with slope \(0.1078 \leq b \leq 0.1250\) and passing through the point \((30, 7.21)\) are optimal. The minimum value for the objective function was reported as 1.65. The results we obtained were as presented in Table 1.
Table 1  A two parameter linear regression model

<table>
<thead>
<tr>
<th>p</th>
<th>a_n</th>
<th>b_n</th>
<th>Initial Values</th>
<th>SRHO</th>
<th>DERV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>3.92</td>
<td>1.096</td>
<td>8.0,-1.0</td>
<td>1.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>.1099</td>
<td>4.0,2.0</td>
<td>1.65</td>
<td></td>
</tr>
<tr>
<td>1.25</td>
<td>3.93</td>
<td>1.080</td>
<td>5.0,1.0</td>
<td>1.30</td>
<td>-.196401D-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.619430D-03</td>
</tr>
<tr>
<td>1.50</td>
<td>3.95</td>
<td>1.052</td>
<td>5.0,1.0</td>
<td>1.01</td>
<td>-.354370D-07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.115185D-05</td>
</tr>
</tbody>
</table>

For p = 1 the residuals corresponding to the observation (30,7.21) were .2284D-06 and .2220D-15 for the two fits respectively.

The nonlinear regression model

\[ y_t = \exp(-\theta x_{1t} \cdot \exp(-\theta x_{2t}^2)) + \epsilon_t \]

was fitted using various choices of p. Since there were no other available results for values of p different from two several initial values for the parameters were considered. Table 2 presents these results.
Table 2  A two parameter nonlinear regression model

<table>
<thead>
<tr>
<th>p</th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>Initial Values</th>
<th>SRHO</th>
<th>DERV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.01</td>
<td>525.6997</td>
<td>872.1269</td>
<td>200,1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>300, 800</td>
<td>0.57177</td>
<td>-1.03530D-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>750,1200</td>
<td>-1.84522D-03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1000,1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.25</td>
<td>610.7786</td>
<td>900.2983</td>
<td>100, 300</td>
<td>0.29875</td>
<td>-4.72908D-10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>750,1200</td>
<td></td>
<td>-1.29813D-09</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1200,1300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.50</td>
<td>706.8473</td>
<td>931.0162</td>
<td>100, 300</td>
<td>0.15128</td>
<td>-7.40393D-11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200,1000</td>
<td></td>
<td>-2.37167D-10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1200,1300</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>750,1200</td>
<td></td>
<td>-1.86779D-11</td>
</tr>
<tr>
<td>2.0</td>
<td>813.8722</td>
<td>961.0026</td>
<td>1000,1000</td>
<td>0.039806</td>
<td>6.81658D-11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>750,1200</td>
<td></td>
<td>-1.86779D-11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200,1000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bard [1] reported $\hat{\theta}_1 = 813.4583$, $\hat{\theta}_2 = 960.9063$, SRHO = 0.039806 for $p$ equal to 2. The author approached the problem using the algorithm by Marquardt as well as an approximation to a second derivative method combined with different choices for the step length. The direction vector derived by Bard was equivalent to the one in our algorithm when $p = 2$.

A third type of implicit model was considered.
\[ Q(x_t, y_t, \theta) = \theta_1 y_t + e_1 x_{1t} + \theta_2 x_{2t} = \epsilon_t. \]

In this case, the function is algebraically implicit in the variable \( y \), yet for specific values of \( \theta_1, \theta_2 \) and \( X_1 \) and \( X_2 \) there exists a solution for \( y \) in terms of \( X_1, X_2, \theta_1, \theta_2 \).

To generate values for the \( y_t \) it was assumed that the true parameter values were \( \theta_1 = 1.0, \theta_2 = -8.0 \), and that the \( X_1 \) and \( X_2 \) were known constants. Solution for the \( y_t \) was obtained by solving the equation with \( \epsilon_t = 0 \). To the generated values an error term from the uniform distribution \((-1/64, 1/64)\) was added so that the model would not apply exactly to the observed sequence \((x_t, y_t)\).

The results were equivalent across different choices of \( p \), probably due to the very good fit of the model. With a starting value of \( \theta_1 = 4.0, \theta_2 = -11.0 \), the solution \((0.9999, -8.000)\) was generated for \( p = \{1.1, 1.5, 2.5\} \).

In carrying out the computations for these three types of implicit models an approximation was used in the choice of step length \( \lambda_r \) at each iteration. The value of \( \theta \) at the \((r + 1)\) iteration was chosen to be

\[ \theta_{r+1} = \theta_r - \lambda_r p_r(\theta_r) \]

where \( \lambda_r = \max_j (C_j) \) for \( j = 1, 2, \ldots, j_0 \) such that

\[ \varphi_n(\theta_{r+1}) < \varphi_n(\theta_r). \]

The value of \( C \) was chosen as .6 in all cases.
8. LITERATURE CITED


