The Use of Semi-parametric Methods in Achieving Robust Inference

José Manuel de Matos Passos

A thesis submitted to the University of Bristol in accordance with the requirements for the degree of Ph.D. in the Faculty of Social Sciences, Department of Economics

May 1996

Obra entregue para apreciação da equivalência ao grau de doutor em Matemática Aplicada à Economia e à Gestão concedida pela Universidade Técnica de Lisboa no Instituto Superior de Economia e Gestão ao abrigo do D.L. 283/83, de 21 de Junho.

Lisboa, 1 de Abril de 1998.

O Presidente do júri

ABSTRACT

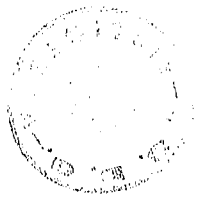
This thesis focuses on some topics in semi-parametric econometrics, particularly the use of semi-parametric methods of estimation to obtain robust inference.

Chapter two proposes a study of the finite-sample performance of the heteroskedastic and autocorrelation consistent covariance matrix estimators (HAC). This performance is accessed through the bias of the first moment of HAC type estimators and the quality of the asymptotic normal approximation to the exact finite-sample distributions of HAC type Wald statistics of scalar linear hypothesis.

In Chapter three, the use of the non-overlapping deleted$-l$ jackknife is used to propose a new approach to estimate the covariance matrix of the least square estimator in a linear regression model. This estimator is robust to the presence of heteroskedasticity and autocorrelation in the errors.

Chapter four deals with improved estimation of regression coefficients through an alternative and efficient method of estimation regression models under heteroskedasticity of unknown form. Kernel and average derivative estimation are used to estimate the conditional variance of the response variable where this conditional variance is assumed to be in an index form.

Chapter five is concerned with the estimation of duration models under unobserved heterogeneity. This is a typical problem in microeconometrics and is in general due to differences among individuals. It is suggested a method of estimation based on a roughness penalty approach.



ACKNOWLEDGEMENTS

I am deeply indebted to my supervisor Andrew Chesher for his support and guidance that made this work possible.

I also would like to thank Whitney Newey, Bernard Silverman, Roger Klein, Garry Phillips and Oliver Linton for helpful comments on several parts of the thesis.

This work would not have been possible without the three years leave kindly granted by the Departamento de Matemática do Instituto Superior de Economia e Gestão (ISEG) da Universidade Técnica de Lisboa. I am also grateful to my colleague Joao Santos Silva, my M.Sc. supervisor Bento Murteira and to the President of the Scientific Council of ISEG Carlos da Silva Ribeiro for their friendship, support and encouragement.

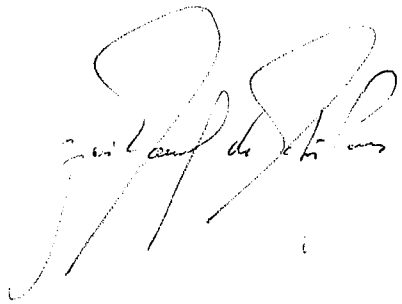
Financial support from Banco de Portugal postgraduate scholarship fund is gratefully acknowledged. Without it this work would not have been possible.

Finally, a very special thank to my wife Helena and my son Ricardo for their support during these three years in Bristol.

AUTHOR'S DECLARATION

The work presented in this thesis was carried out in the Department of Economics at the University of Bristol and it is entirely my own.

The views expressed in this thesis are those of the author and not of the University of Bristol.



CONTENTS

CHAPTER 1 - AN OVERVIEW

1. THE USE OF SEMI-PARAMETRIC METHODS IN ACHIEVING ROBUST INFERENCE.

This thesis focuses on some topics in semi-parametric econometrics, particularly the application of semi-parametric methods of estimation to obtain robust inference. It is a relatively wide area, including a large variety of econometric work and a detailed and comprehensive study of this subject is clearly beyond the scope of this thesis. Only some topics will be addressed here.

Traditionaly, the statistical analysis of economic data is based on a specification of a parametric model depending on a finite number of unknown parameters. This parametric model is usually described as a functional relation between a dependent variable, y, and a set of observable covariates, x, and an unobservable error, ε, where $y = g(x,\theta,\varepsilon)$, θ is an unknown parameter and $g(.)$ is a known function. For this specification to be completed the parametric approach specifies the distribution of the error term. With these elements in hand, the maximum likelihood method can be applied to the estimation of θ. If the model is correctly specified, it is well known that the maximum likelihood estimator (m.l.e.) has all desirable properties as consistency and efficiency.

However this framework has its limitations and can be misleading if not applied carefully. The literature is rich with examples concerning the consequences of departures from the initial assumptions about the parametric model considered. This can lead to incorrect variance estimates or even inconsistency of the estimator. If some of the initial assumptions fails or can not be assumed or if there are components of the model that can not be parameterized, the semi-parametric approach appears as a valuable instrument [for a survey in semi-parametric methods in econometrics consider for example Robinson (1988) and Powell (1992)].

This thesis is concerned with the application of semi-parametric methods in two situations: misspecification of the assumptions concerning the error term as the independence and identically distributed (i.i.d) assumption and misspecification of the model by neglecting heterogeneity, in the context of duration models. The purpose is to use semi-parametric methods in order to make inferences that are robust to these problems. Chapters two to four address the first problem and chapter five addresses the second.

Most of these chapters were originally developed as discussion papers and essays that have been organized in order to produce the five chapters of this thesis, where each one is self contained and independent of the others.

Both microeconomic and macroeconomic data present characteristics that can contribute to violate the assumption of i.i.d. errors. The reasons are heterogeneous population in the first case and temporal dependency in the latter. It is well known that under these circumstances, the usual least square estimator is still consistent but the standard errors are incorrectly estimated. For this reason and because the pattern of the errors are unknown, the traditional procedure is to correct the standard errors through the estimation of robust standard errors. Among the estimators presented in

the literature, the estimator proposed in Newey and West (1987) has been suggested. However there is a lack of knowledge concerning its finite-simple properties and this problem will be studied here.

Chapter two is concerned with the study of the Heteroskedasticity and Autocorrelation Consistent Covariance Matrix (HAC) estimators for the least-squares regression coefficients. Among the HAC type estimators the Newey and West (1987) estimator is considered to show the relation between its finite-sample performance and the design generated by the regressors. This performance is assessed through the bias of the first moment of HAC type estimators and the quality of the asymptotic normal approximation to the exact finite-sample distributions of HAC type Wald statistics of scalar linear hypothesis. In this case Imhof procedure is used. A slight modification of the Newey and West estimator, based on a bias correction in the case of homoskedastic and non-autocorrelated errors, is also presented.

The recent development in non-parametric and semi-parametric methods of estimation opened the way to a growing research in econometrics. Having in mind these new resources and the fact that HAC estimators can be severely biased in small samples, a rather different approach is suggested. The deleted-l Jackknife with non-overlapping blocks and Average Derivative Estimation techniques are used in chapter three and four respectively to develop estimators that are robust to heteroskedasticity and/or autocorrelation in the error term. These estimators can be viewed as alternatives and generalizations of some estimators presented in the literature.

In part three it is proposed a new approach to estimating the covariance matrix of the least square (henceforth LS) estimator in a linear regression model. This approach was inspired by a previous paper of Carlstein (1986) concerning the use of subseries values for estimating the variance of a sample mean, based on dependent stationary data. This concept is extended to the estimation of the standard error of

the LS estimator in regression models. To deal with it and to avoid problems related to a possible lack of degrees of freedom, the non-overlapping deleted$-l$ jackknife (or grouped jackknife) is used instead. This estimator is robust to the presence of heteroskedasticity and autocorrelation in the errors. In another way it generalizes the heteroskedasticity consistent covariance matrix estimator (or the deleted-1 jackknife estimator, as is it known in the literature), to situations with non-independent stationary errors. It is also an alternate to the estimator presented by Newey and West (1987). An application concerning the finite-sample performance of this estimator is also presented.

Chapter four deals with improved estimation of regression coefficients through an alternative and efficient method of estimation regression models under heteroskedasticity of unknown form. In this case, the pattern of heteroskedasticity is first estimated non-parametrically and then used as weights *via* weighted (non-linear) least square estimation. Kernel smoothing and average derivative estimation (henceforth ADE) will be used to estimate the conditional variance of the response variable, where this conditional variance is assumed to be in an index form. This method is presented as a generalisation of Carrol's (1982) estimator where the conditional variance of the response variable is not restricted to be a function of the conditional mean of the response variable. A comparison is made through a Monte Carlo simulation.

Chapter five is concerned with the estimation of duration models when unobservable individual effects are present. This problem can be viewed as a mixture model where the density of the unobservables is the unknown mixing density that is to be estimated using some appropriated method. Two different approaches have been considered in the literature: one uses a parametric specification of the heterogeneity by assuming some mixing density; the other one uses nonparametric methods as, for example, the non-parametric maximum likelihood estimation. The first approach is too restrictive; the second one provides good estimates of the structural parameters but

the underlying mixing density is poorly estimated. In this chapter it is suggested an alternative approach towards this last problem, using a roughness penalty approach.

2. REFERENCES

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence, *The Annals of Statistics*, 14, 1171-1179.

Carrol, R. J. (1982). Adapting for Heteroskedasticity in Linear Models. *Annals od Statistics*, 10, 1224-1233.

Newey, W. and West, K. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica*, 55, 703-708.

Powell, J. (1992). Estimation of Semiparametric Models. Draft, Princeton University.

Robinson, P. (1988). Semiparametric Econometrics: A Survey. Journal of Applied Econometrics, 3, 35-51.

CHAPTER 2 - FINITE-SAMPLE PERFORMANCE OF THE HETEROSKEDASTICITY AND AUTOCORRELATION CONSISTENT COVARIANCE MATRIX ESTIMATORS

1. INTRODUCTION

In the econometric framework, the consequences of autocorrelated and heteroskedastic errors have been studied in the literature for a long time. These are common problems associated with time series and cross sectional models, respectively, as it can be seen when some useful diagnostic tests are applied. In this context it is well known that if the errors are heteroskedastic and/or autocorrelated, the usual estimator is still consistent but the standard errors are erroneous.

If the form of autocorrelation and/or heteroskedasticity is known, there are appropriated techniques to minimise this problem. However this is not what happens in general and several works have been developed to find a consistent estimator of the standard errors, robust to this kind of unknown error structure.

The key to finding such an estimator is due to Eicker (1963). He has shown in the case of heteroskedasticity of unknown order, that consistent estimation of the covariance matrix estimator of the least-squares (LS) regression coefficients does not require consistent estimation of the variance of the disturbance term. White (1980), MacKinnon and White (1985) and Chesher and Jewitt (1987), among others, have studied

heteroskedasticity consistent covariance matrix (henceforth HCCM) estimators. But when the errors are not independent other estimators will be needed. In a pioneering paper concerning Generalised Method of Moments estimation Hansen (1982) derived a heteroskedasticity and autocorrelation consistent covariance (henceforth HAC) estimator. Since then several works have addressed this subject. Hansen and Singleton (1982), have applied this estimator in a non-linear rational expectations model; White (1984) dedicated a whole chapter to studying the properties of Hansen's type estimator for different structures of the errors; White and Domowitz (1984) applied this estimator to the non-linear regression; Wooldridge (1991) suggest the use of this estimator in robust diagnostics for non-linear models of conditional means and conditional variance; Newey and West (1987) presented a positive semi-definite matrix estimator; Andrews (1991) showed that all these estimators can be viewed as kernel estimators, presenting simultaneously a Monte Carlo study for different types of kernels. Moreover he advocates an HAC estimator that uses the Quadratic-Spectral kernel. In this context the estimators proposed by Hansen (1982) [see also White (1984)], Newey and West (1987) and Gallant (1987, pag. 533) correspond to estimators using truncated, Bartlett and Parzen kernels, respectively. Recently, Andrews and Monahan (1992) proposed a slightly different version and named it prewhitened kernel estimator, with vector autoregressions employed in the prewhitening stage.

When one suspects for non-iid errors and little is known about their structure, the use of HAC type estimators in test statistics have been suggested by the authors being cited as a way to compute correct standard errors and therefore correct inferences. Furthermore its facility of computation is now accessable through econometric packages like Shazam. However little is known about their finite-sample properties. In a previous paper, Chesher and Jewitt (1987) developed finite-sample results for the HCCM estimators, showing that substantial bias can occur when the regression design contains points of high leverage. In what follows their results are extended to HAC estimators.

The remainder of this Chapter is organised as follows. In Section 2, are presented the HAC type estimators (in particular the Newey and West estimator) and some notation related to them, in the context of linear models. Additionally it is shown that HAC type estimators can be decomposed in a HCCM estimator and in an Autocorrelation Consistent Covariance estimator. The expectation and variance of this estimator are presented in Section 3. Section 4 is dedicated to the finite-sample performance of the Newey and West estimator, given particular attention to the effects of leverage points in the data. In this context three different error structures are considered: independent, AR(1) and MA(1) errors. Section 5 is dedicated to the evaluation of the quality of the first-order asymptotic approximation to the null distribution of the Wald test, using the Newey and West estimator. Finally, Section 6 summarises the main conclusions and presents some possible directions for future research.

2. HAC TYPE ESTIMATORS IN THE LINEAR MODEL

The present work deals with the linear model,

$$y = X\beta + \varepsilon \ , \tag{1}$$

where y is an $n \times 1$ vector of observations, X is a $n \times k$ matrix of full column rank k with rows x_i, β is a $kx1$ vector of unknown parameters and ε is a $n \times 1$ vector of errors with $E(\varepsilon \mid X) = 0$ and $E(\varepsilon\varepsilon' \mid X) = \Omega$ positive definite of order $n \times n$. Additionally it is assumed that, conditional on X, the errors are covariance stationary. Under these assumptions, the covariance matrix of the LS estimator of β, $\widehat{\beta} = (X'X)^{-1}X'y$, conditional on X is given by

$$\Sigma = (X'X)^{-1}X'\Omega X(X'X)^{-1}. \tag{2}$$

In this expression $X'\Omega X$ can be decomposed in a suitable way by,

$$X'\Omega X = \sum_{i=0}^{n} \gamma(0)\, x_i'x_i + \sum_{j=1}^{n-1}\sum_{i=j+1}^{n} \gamma(j)\,(x_i'x_{i-j} + x_{i-j}'x_i)\ , \tag{3}$$

where $\gamma(j) = E(\varepsilon_i\varepsilon_{i-j})$ is the autocovariance function of order j and $\gamma(0)$ the error variance that is assumed constant from now on.

In the LS context, the direct estimation of Ω, substituting the unknown errors, ε_i, by the residuals, $\widehat{\varepsilon}_i$, leads to a degenerate estimator, due to the orthogonality condition, $X'\widehat{\varepsilon} = 0$. If the errors are covariance stationary, the autocovariance function, $E(\varepsilon_i\varepsilon_j)$, is a function of the difference, $\mid i-j \mid$. Additionally if one limits the attention to the case in which the autocovariance function of order $\mid i-j \mid$ decreases as $\mid i-j \mid$ tends to infinity, it seems reasonable to estimate $X'\Omega X$ considering only the most significant autocovariances. In the particular case of linear models, Hansen's (1982) estimator can be expressed as follows,

$$X'\widehat{\Omega}X = \sum_{i=0}^{n} \widehat{\varepsilon}_i^2\, x_i'x_i + \sum_{j=1}^{m}\sum_{i=j+1}^{n} \widehat{\varepsilon}_i\widehat{\varepsilon}_{i-j}(x_i'x_{i-j} + x_{i-j}'x_i)\ , \tag{4}$$

where m is a lag truncation that in general equals the number of non-zero autocorrelations of $x_i'\varepsilon_i$. If the structure of the errors is known a *priori*, for example if it is a moving average of order q, then such lag should be equal to q. When this structure is unknown White and Domowitz (1984) and Andrews (1991) proposed ap-

propriated methods to choose the optimal value of m. However Hansen's estimator has the drawback of not always being positive semi-definite. To solve this problem, two techniques have been usually presented in the literature: time domain techniques due to Cumby, Huizinga and Obstfeld (1983) and Newey and West (1987), among others, and frequency domain techniques suggested initially by Hansen (1982) and adopted later by Andrews (1991, 1992)[1]. In what follows time the domain technique is considered, particularly the Newey and West estimator due to its simplicity of computation. These authors proposed a positive semi-definite HAC estimator, simply by smoothing the samples autocovariances in (4),

$$X'\widehat{\Omega}X = \sum_{i=0}^{n} \widehat{\varepsilon}_i^2 \, x_i' x_i + \sum_{j=1}^{m} \sum_{i=j+1}^{n} \kappa(j,m) \, \widehat{\varepsilon}_i \widehat{\varepsilon}_{i-j} \, (x_i' x_{i-j} + x_{i-j}' x_i), \tag{5}$$

where

$$\kappa(j,m) = 1 - \frac{j}{m+1} \, , \tag{6}$$

is the weight associated with the autocovariance of order j (Having in mind the dependence of $\kappa(j,m)$ on m, henceforth κ_j is considered for simplifying purposes). Using the methodology of Andrews (1991), the Newey and West estimator can be written as,

$$\frac{1}{n} X'\widehat{\Omega}X = \sum_{j=-m}^{m} \varphi\left(\frac{j}{S_n}\right) \widehat{\Gamma}(j) \, , \tag{7}$$

[1]This technique is motivated by the fact that when $x_i'\varepsilon_i$ is second order stationary, the estimator $(1/n)X'\widehat{\Omega}X$ is equal to $2\pi f(0)$ where $f(0)$ is the estimator of the spectral density of $x_i'\varepsilon_i$ at frequency zero.

where

$$\hat{\Gamma}(j) = \begin{cases} \frac{1}{n} \sum_{i=j+1}^{n} x_i' \hat{\varepsilon}_i \, \hat{\varepsilon}_{i-j} \, x_{i-j} & , \; j \geq 0 \; , \\ \\ \frac{1}{n} \sum_{i=-j+1}^{n} x_{i+j}' \hat{\varepsilon}_{i+j} \, \hat{\varepsilon}_i \, x_i & , \; j < 0 \; , \end{cases}$$

$\varphi(\cdot)$ is a kernel[2] and S_n is a bandwidth parameter, with $S_n = m + 1$.

When the structure of the errors is unknown a priori, the estimator of $X'\Omega X$ is still consistent if m increases at some appropriate rate with the sample size[3] and the errors obey some regularity conditions concerning finite fourth moments and certain mixing sequences [see for example White (1984) and Newey and West (1987)]. Keener and Kmenta (1991) have proposed an alternative and easily proof by restricting the size of the error correlations. Recently, Hansen (1992) in the context of kernel estimation showed a consistency proof under mild conditions about the errors' structure that requires only the existence of second moments.

Putting $X'\hat{\Omega}_H X = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 x_i' x_i$ and $X'\hat{\Omega}_A X = \sum_{j=1}^{m} \sum_{i=j+1}^{n} \kappa_j \, \hat{\varepsilon}_i \hat{\varepsilon}_{i-j} \, (x_i' x_{i-j} + x_{i-j}' x_i)$ in expression (5), the estimator of the covariance matrix becomes,

$$\hat{\Sigma} = (X'X)^{-1} X' \hat{\Omega}_H X (X'X)^{-1} + (X'X)^{-1} X' \hat{\Omega}_A X (X'X)^{-1}, \qquad (8)$$

[2]In Newey and West $\varphi(\cdot)$ is the Bartlett kernel,

$$\varphi(x) = \begin{cases} 1 - |\, x \,| & , \quad |\, x \,| \leq 1 \; , \\ \\ 0 & , \quad \text{otherwise} \; . \end{cases}$$

[3]For consistency purposes $m = o(n^{1/4})$ [see for example Davidson and Mackinnon (1993), pag.611].

where the first factor on the right hand side of (8) is the HCCM estimator due to White (1980), and the second term is an autocorrelation consistent covariance matrix estimator.

In the following, the finite-sample performance of the Newey and West estimator and its relation to extreme leverage points are studied. This estimator is considered, among the HAC type estimators, for two main reasons: because it is one of the most well known and by the conclusion of Andrews (1991) that the differences (in performance) between HAC estimators are not large for the kernels considered in his simulation study.

3. MOMENTS OF THE HAC ESTIMATORS

In this Section, the first and second moments of the HAC type estimators are derived, for any chosen combination, w, of the parameters in $\widehat{\beta}$. Therefore, $\widehat{Var(w'\widehat{\beta})} = \widehat{\Sigma}_w$ is a scalar and thus can be expressed in vectorial form as,

$$\widehat{\Sigma}_w = (z'_w \otimes z'_w) vec(\widehat{\Omega})$$

where $z'_w = w'(X'X)^{-1}X'$ is a $1 \times k$ vector, $\otimes$ is the kroneker product and $vec(\widehat{\Omega})$ is a $n^2 \times 1$ vector formed by stacking the columns of $\widehat{\Omega}$. The first and second moments of $\widehat{\Sigma}_w$ evolve products of quadratic forms in normal variables and the results of Magnus (1978) will apply.

Theorem 1: Considering m_i as the ith column of $M = I - X(X'X)^{-1}X'$, the first and second moment of the HAC estimators are given by the following expressions,

1. $E(\hat{\Sigma}_w) = (z'_w \otimes z'_w)vec(\Xi)$, where Ξ is a nxn matrix defined as $\Xi = E(\hat{\Omega}) = [E(\hat{\Omega}_{ij})] = [\kappa_{i-j}\, m'_i \Omega m_j]$, $i,\ j = 1, 2, \ldots, n$, with $\kappa_0 = 1$;

2. $Cov(\hat{\Sigma}_w \hat{\Sigma}_v) = (z'_w \otimes z'_w)\Gamma(z_v \otimes z_v)$, where Γ is a $n^2 \times n^2$ matrix defined as $\Gamma = E[vec(\hat{\Omega})vec(\hat{\Omega})'] - E[vec(\hat{\Omega})]E[vec(\hat{\Omega})'] = [\kappa_{i-j}\,\kappa_{l-k}\,(m'_j \Omega m_l\, m'_k \Omega m_i + m'_j \Omega m_k\, m'_l \Omega m_i)]$, $i, j, l, k = 1, 2, \ldots, n$, with $\kappa_0 = 1$.

Proof: For the ith residual, $\hat{\varepsilon}_i = m'_i \varepsilon$. From Section 2 ε can be expressed as $\varepsilon = \Omega^{1/2} u$, with $u \sim N(0, I)$. From now on let $s_i = \Omega^{1/2} m_i$ and $A_{ij} = s_i s'_j + s_j s'_i$. To prove part 1 note that $\hat{\Omega}_{ij} = \kappa_{i-j} \hat{\varepsilon}_i \hat{\varepsilon}_j = \kappa_{i-j} m'_i \Omega^{1/2} u u' \Omega^{1/2} m'_j = \kappa_{i-j} u'\ s_i s'_j u$ is a quadratic form in normal variables. Applying lemma 6.1 of Magnus, $E(\hat{\Omega}_{ij}) = \kappa_{i-j}(1/2)E[u' A_{ij} u] = \kappa_{i-j}(1/2)tr(A_{ij}) = \kappa_{i-j} m'_i \Omega m_j$, proving part 1. To prove 2 one needs to compute the expectation of $vec(\hat{\Omega})vec(\hat{\Omega})'$, with generic element $\hat{\omega}_{ij}\hat{\omega}_{kl} = \kappa_{i\ \ j}\, \hat{\varepsilon}_i \hat{\varepsilon}_j \kappa_{l\ \ k} \hat{\varepsilon}_l \hat{\varepsilon}_k$. Following the same procedure, these elements can be expressed as a product of quadratic forms in normal variables, i.e., $\hat{\omega}_{ij}\hat{\omega}_{kl} = \kappa_{i-j}\kappa_{l-k}\, u' s_i s'_j u\, u' s_l s'_k u$. Rewriting this term as $\hat{\omega}_{ij}\hat{\omega}_{kl} = \kappa_{i-j}\kappa_{l-k}\,(1/4)\, u' A_{ij} u\ u' A_{lk} u$, one can apply lemma 6.2 of Magnus. Thus, $E(\hat{\omega}_{ij}\hat{\omega}_{kl}) = \kappa_{i-j}\kappa_{l-k}\,(1/4)[tr(A_{ij})tr(A_{lk}) + 2tr(A_{ij}A_{lk})] = \kappa_{i-j}\kappa_{l-k}(m'_j \Omega m_i\, m'_k \Omega m_l + m'_j \Omega m_l\, m'_k \Omega m_i + m'_j \Omega m_k\, m'_l \Omega m_i)$ and finally, using the proof of part 1, the generic element of Γ follows $\square$.

4. BIAS OF THE HAC ESTIMATORS

In this Section one shows that the finite-sample performance of the Newey and West estimator is related to the design of the matrix X. The importance of this design is more general than the purposes of this Chapter and should be evaluated, by using some diagnostic measure, in any applied econometric work. Such an importance can be explained briefly as follows [see Pollock (1979), Chapter 5]. Due to the fact that in the regression context the vector y does not belong to the space spanned by the columns of X, $M(X)$, the system $y = X\beta$ is impossible. Therefore, the solution to this problem can not be exact but approximated, where this approximation is

made in two main steps: first y is projected into $\widehat{y} \in M(X)$ by a projection matrix H, where $\widehat{y} = Hy$; the value of β is derived in the second step by minimising the distance between y and[4] $\widehat{y} = X\widehat{\beta}$. Considering the problem under the above explanation, the relation between estimated and true values, depends on the elements of the projection matrix, as can be seen by the following expressions:

$$\widehat{y}_i = h_{ii}y_i + \sum_{j \neq i}^{n} h_{ij}y_j \ , \qquad \widehat{\varepsilon}_i = (1 - h_{ii})\varepsilon_i - \sum_{j \neq i}^{n} h_{ij}\varepsilon_j$$

Knowing that h_{ii} is a measure of the distance of the point x_i from the bulk of the points in X, one remote point leads the regression hyperplane to pass near this point[5]. This fact can be responsible for some undesirable non-linearities in the data pattern and therefore data points with leverage can worsen the finite-sample performance of the model. [6].

As pointed by Huber (1973), $\partial\widehat{y}_i/\partial y_i = h_{ii}$ and the inverse of h_{ii} can be thought as the equivalent number of observations that determine $\widehat{y}_i$. He also suggests points with $h_{ii} > 0.2$ to be classified as high leverage points. Others, like Belsley et al. (1980) define as high leverage observations with corresponding diagonal element of H greater than two times the mean of the h_{ii}'s (note that $\overline{h} = (1/n)\sum_1^n h_{ii} = k/n$). Good references in these topics are Huber (1981), Cook and Weiseberg (1982) and Chatterjee and Hadi (1987).

To assess the effect of leverage points on the finite-sample performance of the Newey and West estimator one requires some measure that allows us to confront it under slight perturbations made to the design. To deal with this, one will use the

[4]With this procedure one finds $II = X(X'X)^{-1}X'$.

[5]When h_{ii} is high, $\widehat{y}_i$ is approximately equal to y_i and consequently the correspondent residual is approximately equal to zero.

[6]One factor associated with the presence of leverage points is measurement error.

proportionate bias that is no more than the bias scaled by the true value. Thus, for any chosen combination, w, of the parameters in β, $w\beta$, the proportionate bias of $\widehat{Var(w'\hat{\beta})} = \hat{\Sigma}_w$ is given by,

$$pb(\hat{\Sigma}_w) = \frac{E(\hat{\Sigma}_w) - \Sigma_w}{\Sigma_w} ,$$

or,

$$pb(\hat{\Sigma}_w) = \frac{z'Qz}{z'\Omega z} , \tag{9}$$

where $z = z_w$ is considered for simplifying purposes and

$$Q = E(\hat{\Omega}) - \Omega .$$

Defining h_i and Ω_i as the ith column of H and Ω, respectively, the generic element of Q becomes,

$$Q_{i,i-j} = \begin{cases} -2\Omega_i' h_i + h_i'\Omega h_i & , \ j = 0 \quad , \ i = 1,2,\ldots,n \\ \kappa_j \left(\Omega_{i,i-j} - h_i'\Omega_{i-j} - \right. & \\ \left. \Omega_i' h_{i-j} + h_i \Omega h_{i-j}\right) - \Omega_{i,i-j} & , \ j = 1,\ldots,m \quad , \ i = j+1,\ldots,n \\ -\Omega_{i,i-j} & , \ j > m \quad , \ i = j+1,\ldots,n \end{cases} \tag{10}$$

The dependence of $Q_{i,i-j}$ on the design of X is evident. What is not evident is the

amount and the direction of this dependence. In a heteroskedastic model with non autocorrelated errors Chesher and Jewitt (1987) found appropriate bounds, in the finite-sample case, for the bias of the White's HCCM estimator and have shown a straight relation between this bounds and the design of the X matrix. However, if the errors are not independent their methodology[7] does not seem to be possible because the matrix Q is not of a diagonal type, presenting a more complex structure. Therefore, as an alternative, the bias of the Newey and West estimator and its relations with the design of the X matrix will be studied via examples of three particular cases: independent, AR(1) and MA(1) errors. In each case, homoskedastic errors are assumed. These examples allow us to draw a better idea of how important the design can be to explain the bias and if this importance depends on a particular structure of the errors.

The main example considered in this Section, is the linear model (1) and the MacKinnon and White (1985) data[8], with 50 observations and three regressors, $X = [X_1\ X_2\ X_3]$, where X_1 is a vector of ones and X_2 and X_3 are the rate of growth of real U.S. disposable income and the U.S. treasury bill rate, respectively, seasonally adjusted, for the period 1963-3 to 1974-4. The dependent variable is determined by the linear relation (1), given a particular Ω matrix.

4.1 Choice of the lag truncation

One of the problems related to the computation of the Newey and West estimate is the derivation of the optimal value for the lag truncation parameter. White and Domowitz (1984) and Andrews (1991) gave some insight to this problem[9]. Andrews (1991) in the context of kernel HAC estimators has derived automatic bandwidth

[7] In particular the derivation of an algebraical expression for the eigenvalues, λ, solution to the characteristic polynomial $|Q - \lambda\Omega| = 0$.

[8] See table 3.

[9] Recently Newey and West (1994) suggested a non-parametric method for automatically selecting the number of autocovariances to use in computing a HAC estimator.

estimators, where the value of the optimal bandwidth parameter is a function of the number of observations and the structure of $X'\widehat{\varepsilon}$ [see Andrews(1991), pgs. 832-35, in particular his expressions (6.2) and (6.4) to (6.8)]. The structure of $X'\widehat{\varepsilon}$ can be assessed specifying k univariate approximate parametric models for $\{x_{at}\widehat{\varepsilon}_t\}$ for $a = 1, ..., k$. The estimated parameters of these models are next used to compute the optimal m. However, in general, there is no a *priori* information about the best approximated parametric model and this fact can be seen as an inconvenience of the application of Andrews' method.

Another possible problem, not yet studied, is the performance of these approximate parametric models when the design of X has one or more leverage points and therefore, its effects on the derivation of the optimal value of m. Suppose for example that the sequences of observations $\{x_{at}\widehat{\varepsilon}_t\}_{t=1}^{n}$, for each $a = 1, ..., k$, are generated by a zero-mean AR(1) process. If contamination is added to the observed value $x_{al}\widehat{\varepsilon}_l$, $l = 1, ..., n$, one says that one isolated *additive outlier* occurs, in the terminology of Fox (1972). In this case the point $(x_{al-1}\widehat{\varepsilon}_{l-1}, x_{al}\widehat{\varepsilon}_l)$ is an outlier in the response variable and $(x_{al}\widehat{\varepsilon}_l, x_{al+1}\widehat{\varepsilon}_{l+1})$ a leverage point. As a consequence, the usual LS method will yield biased parameter estimates, meaning that the resulting observations no longer obey the AR(1) model. Finally, Andrews' procedure gives an optimal value for m whichever the direction w has. As the following example suggests, this value should depend on this direction, particularly when the data contains one or more leverage points.

Using the Mackinnon and White data and the model (1) with AR(1) errors, figures 1a and 1b show the proportionate bias of the Newey and West HAC estimator of β_2 and β_3, respectively, as a function of correlation coefficient and lag truncation. Table[10] 1 shows the optimal value of the lag truncation associated with each value

[10]For each ρ this table gives the values of m in which the absolute value of the proportionate bias is minimum. $m^*(\beta_i)$, corresponds to the optimal lag truncation associated with the Newey and

of the correlation coefficient[11]. By optimal value, one means the value for which the proportionate bias is minimum. As can be seen, the shape of the proportionate bias and the optimal value of the lag truncation associated with β_2 and β_3 are very different. This difference is due to the presence of a high leverage point in the Mackinnon and White data (the 48^{th} data point) in the direction of β_2 [see Chesher and Austin (1991), pag. 160][12]. When measured by the diagonal elements of the hat matrix, the correspondent value associated with the 48^{th} data point is $h_{48,48} = 0.39$, approximately two times the limit value suggest by Huber and 6.5 times the mean of the h_{ii}'s.

Table 1-Optimal value for the optimal lag truncation as a function of ρ
(AR(1) Errors, Mackinnon and White data)

ρ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$m^*(\beta_2)$	0	0	0	0	0	1	2	5	10	16
$m^*(\beta_3)$	0	1	2	3	4	5	6	8	12	13

The same conclusions can be drawn for the case with MA(1) errors. The difference is only on the optimal value for the lag truncation that should be approximately equal to one [see table[13] 2 and figures 2a and 2b].

West estimator of $Var(\widehat{\beta}_i)$, $i = 1, 2$.

[11]Note that what these pictures show are not simulated but exact values.

[12]Deleting this point, the proportionate bias and the optimal lag truncation associated with β_2 and β_3 are approximately equal.

[13]θ is the parameter of the MA(1) process.

Table 2-Optimal value for the optimal lag truncation as a function of θ and ρ (MA(1) Errors, Mackinnon and White data)

ρ	0.00	0.10	0.192	0.275	0.345	0.400	0.441	0.470	0.488	0.497
θ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$m^*(\beta_2)$	0	0	0	0	0	0	0	0	0	0
$m^*(\beta_3)$	0	1	1	2	2	2	2	2	2	2

In the following a fixed lag truncation for different structures of the errors is considered. This simplification does not change the conclusions achieved for the relation between bias and leverage.

4.2 Independent and Homoskedastic Errors

With independent errors the generic element of Q in expression (10) simplifies to,

$$Q_{i,i-j} = \begin{cases} -h_{ii} & , \ j = 0 & , \ i = 1, 2, \ldots, n \\ -\kappa_j \, h_{i,i-j} & , \ j = 1, 2, \ldots, m & , \ i = j+1, \ldots, n \\ 0 & , \ j > m & , \ i = j+1, \ldots, n \end{cases} \tag{11}$$

and the proportionate bias (9) becomes,

$$pb(\widehat{\Sigma}_w) = -\frac{1}{z'z} \left(\sum_{i=1}^{n} z_i^2 h_{ii} + 2 \sum_{j=1}^{m} \sum_{i=j+1}^{n} \kappa_j \, h_{i,i-j} \, z_i \, z_{i-j} \right) . \tag{12}$$

However this expression does not give us a good guidance to the knowledge of how large it can be. Following Chesher and Jewitt (1987), a better way is to bound the above expression by using the following mathematical device,

$$\sup_z \frac{z'Qz}{z'z} = \max(\lambda_i) \ , \qquad \inf_z \frac{z'Qz}{z'z} = \min(\lambda_i) \ , \tag{13}$$

where the λ_i's are solutions of the characteristic equation, $\mid Q - \lambda I \mid = 0$. Unfortunately the computation of the eigenvalues by $\mid Q - \lambda I \mid = 0$ is not algebraically workable due to the fact that Q is a band matrix, with bandwidth equal to the lag truncation.

Alternatively, if z is the eigenvector associated with the eigenvalue λ it is well known that $(Q - \lambda I)\, z = 0$. Therefore, for each $i = 1, 2, \ldots, n$, and $z_i \neq 0$,

$$(q_{ii} - \lambda)\, z_i + \sum_{j \neq i} q_{ij}\, z_j = 0 \ ,$$

and the generic expression for the eigenvalues becomes,

$$\lambda = q_{ii} + \sum_{j \neq i} \alpha_{ij}\, q_{ij} \ ,$$

or, by (11),

$$\lambda = -h_{ii} - \sum_{\substack{j \neq i \\ |i-j| \leq m}} \kappa_{|i-j|}\, \alpha_{ij}\, h_{ij} \ , \tag{14}$$

with $\alpha_{ij} = z_j/z_i$. Applying Gerschgorin's *circle theorem* [see for example Strang (1980)], "*every eigenvalue of* Q *lies in at least one of the circles* C_i, $i = 1, 2, ..., n$, *where* C_i *has its centre at the diagonal entry* q_{ii} *and its radius equal to the absolute sum along the rest of the row*". Therefore, assuming from now on that $\kappa_{|i-j|}$ are the weights defined by Newey and West, this theorem implies that

$$| \lambda + h_{ii} | \leq \sum_{\substack{j \neq i \\ |i-j| \leq m}} \kappa_{|i-j|} \mid h_{ij} \mid, \tag{15}$$

and the eigenvalues of Q lie in the ith circle, i.e., λ is bounded by,

$$-h_{ii} - \sum_{\substack{j \neq i \\ |i-j| \leq m}} \kappa_{|i-j|} \mid h_{ij} \mid \leq \lambda \leq -h_{ii} + \sum_{\substack{j \neq i \\ |i-j| \leq m}} \kappa_{|i-j|} \mid h_{ij} \mid . \tag{16}$$

To better evaluate this inequality, in particular the maximum and minimum value of the right and left hand side, respectively, one will make use of the following lemma [see for example, Cook and Weisberg (1982)],

Lemma 1: The hat matrix, H, has the properties:

1. if $h_{ii} = 1$ then $h_{ij} = 0$ for all $j \neq i$;
2. if $h_{ii} = 0$ then $h_{ij} = 0$ for all[14] j;
3. $h_{ij}^2 \leq h_{ii}(1 - h_{ii})$ for all $j \neq i$ and
4. if $h_{ij}^2 = h_{ii}(1 - h_{ii})$ then $| h_{ij} |$ is at its maximum value and $h_{ik}^2 = 0$ for all $k \neq i$, j.

[14]Note that when X contains a constant column $h_{ii} > n^{-1}$.

Proof: Due to the fact that H is an idempotent matrix, h_{ii} can be written as $h_{ii} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + h_{ij}^2 + \sum_{r \neq i,j}^{n} h_{ir}^2$ and 1 and 2 follows immediately. To prove part 3 and 4 note that the above equality can be written as $h_{ii}(1 - h_{ii}) = h_{ij}^2 + \sum_{r \neq i,j}^{n} h_{ir}^2$ [see Cook and Weisberg (1982) and Chatterjee and Hadi (1988)]. □

Whichever the values of $\kappa_{|i-j|}$ and h_{ij} can be, it follows from *lemma 1* that when $h_{ii} \to 1$, $\lambda_i \to -1$ and when $h_{ii} \to 0$, $\lambda_i \to 0$. As a first conclusion, downward bias equal to its minimum value is attainable, when $h_{ii} = 1$.

Part 3 and 4 of *lemma 1* show a straight relation between h_{ii} and h_{ij}. Thus, for a given value of h_{ii} the maximum of λ depends only on the sum of the right hand side of (16). Applying *lemma 1* to this sum, its maximum value is attained at $\mid h_{ij} \mid = h_{ii}^{1/2}(1 - h_{ii})^{1/2}$ and[15] $\mid i - j \mid = 1$. Therefore, having in mind these results and expression (13) one has,

$$pb(\widehat{\Sigma}_w) \leq \max\{ -h_{ii} + \kappa_1 h_{ii}^{1/2}(1 - h_{ii})^{1/2} \} . \tag{17}$$

Considering the Newey and West estimator one has, $\kappa_1 = m/(m+1)$. In particular, if $m = 1$, λ is limited above by 0.0591, when $h_{ii} = 0.053$. The question that arises at this point is to know if this limit can be attained. A theoretical answer to this question was not possible through this paper. However, from the examples considered, some of which are presented below, there are reasons to believe that the maximum eigenvalue of Q is not greater than zero.

Additionally, inequality (16) or (17) shows a straight relation between proportionate bias and leverage. An illustration can be seen in the following example. Using the Mackinnon and White data and the linear model (1) with $N(0,1)$ errors, one will

[15]Note that $\kappa_{|i-j|}$ decreases with $|i-j|$.

control the leverage associated with the 48^{th} data point in the direction of β_2. For $m = 1$ figure 3 shows the proportionate bias and the bounds for the proportionate bias[16] as a function of the leverage associated with the 48^{th} data point. If the proportionate bias is sensitive to the leverage associated with the 48^{th} data point then it should change, as $h_{48,48}$ changes. In fact when $h_{48,48}$ increases, the proportionate bias associated with β_2 tends towards its minimum value. However such a dependence is not straightforward. It depends also on the particular direction considered. In this sense the Newey and West estimator is drastically downward biased if it is computed in a direction for which there exists an observation with a high leverage.

In a second example, the relation between leverage points and the proportionate bias is assessed by considering a simple model with one regressor,

$$x_i = \begin{cases} \delta & , \quad i = 1 \\ 1 & , \quad i \text{ odd} \\ -1 & , \quad i \text{ even} \end{cases}$$

where $\delta \in \Re$. In this case the proportionate bias, for all ω, becomes,

$$pb(\widehat{\Sigma}) = -\frac{\delta^4 + n - 1}{(\delta^2 + n - 1)^2} - \frac{2}{(\delta^2 + n - 1)^2}\left[\kappa_1(\delta + n - 2) + \right.$$

$$\left. + \kappa_2(\delta + n - 3) + \cdots + \kappa_m(\delta + n - m - 1)\right]$$

For fixed n, when[17] $\delta \to \infty$, $h_{11}(\delta) = (\delta^2 \ / \ (\delta^2 + n - 1)) \to 1$ and $pb(\widehat{\Sigma}) \to -1$. This

[16]With non-autocorrelated errors the proportionate bias is expressed by $z'Qz/z'z$. The supremum and infimum of this ratio over z are given by the maximum and minimum of the λ_i's values, respectively, solutions of the characteristic equation $\mid Q - \lambda I \mid = 0$.

[17]In a regression problem with one regressor $\sum_{i=1}^{n} h_{ii} = 1$. Therefore when $h_{11} \to 1$, $h_{ii} \to 0$ for all $i \neq 1$.

relation is pictured in figure 4 and as can be seen, the proportional bias is negative. When $\delta \to 1$, $h_{ii}(\delta) \to 1/n$, the design becomes well balanced, where each point contributes equally to the regression line. In this case the proportionate bias depends only on the value of the lag truncation with a maximum value of $-1/n$ (when $m = 0$) and a minimum value that tends towards -1 as m approaches n. The value of the proportionate bias is only explained by the autocorrelation structure of the residuals that are never independent, even though when errors are independent.

4.3 Non-independent and homoskedastic errors

In this subsection, model (1) and the Mackinnon and White data are considered in two situations: AR(1) errors and MA(1) errors. In the case of AR(1) errors, figures 5a and 5b show the proportionate bias of the Newey and West and LS estimators associated with $\widehat{\beta}_2$ and $\widehat{\beta}_3$, respectively, as a function of the correlation coefficient, with $m = 2$ and $\rho \in [0, 0.9]$ in steps of 0.1. For the moment, three points should be kept in mind: 1) The LS variance estimator is not dominated by the Newey and West estimator and for small values of ρ it performs better[18]; 2) the shape of the proportional bias associated with the variance estimators of $\widehat{\beta}_2$ and[19] $\widehat{\beta}_3$ are very different and 3) the absolute value of the bias increases significantly when ρ approaches the upper limit[20]. As seen above, this difference in the shape of the proportionate bias is due to the presence of a leverage point in the direction of β_2. Therefore, if the proportionate bias is sensitive to this kind of observations, then deleting this point should have some effect on it. In fact the deletion of this point leads to a change in the shape and, in general, a reduction (in absolute value) in the proportionate bias associated to β_2 [see figs.6a and 6b].

[18]A slight modification of the NW estimator with improved results for small values of ρ will be presented in the next section.

[19]Selecting $w' = [0\ 1\ 0]$ and $w' = [0\ 0\ 1]$ respectively.

[20]It should be noted that part of this increasing is due to the fact that the value of the lag truncation is fixed.

Another interesting problem is the knowledge of the proportionate bias for all possible directions $w \in \Re^k$, considering expressions similar to (13), with the difference that the denominator is now given by $z'\Omega z$. Letting, for example $\lambda_{\max} = \max(\lambda_i)$, the vectors z^* that satisfy,

$$\frac{z'Qz}{z'\Omega z} = \lambda_{\max} ,$$

are the eigenvectors associated with $\lambda_{\max}$, solutions to $(Q - \lambda_{\max}\Omega)\, z = 0$. To these vectors the proportionate bias (9) attains its maximum value. However because the matrix X is fixed, expression (13) should be evaluated with care. The main reason is that both $\lambda_{\max}$ ($\lambda_{\min}$) and z depend on the design of X and therefore the supremum (infinum) over z should be calculated subject to the restriction of X fixed. Otherwise, if X changes, resulting for example from the evaluation of expressions similar to (13), the eigenvalues no longer will be the same. In this sense and having in mind the relation $z^* = X(X'X)^{-1}w^*$, the direction for which the proportionate bias attends its maximum becomes equal to $w^* = X'z^*$ (i.e., a linear combination of the rows of X, where the parameters of this combination are the elements of the eigenvector associated with $\lambda_{\max}$)[21]. The proportionate bias and its bounds, viewed as function of ρ, are pictured in figure 7.

In order to eliminate some of the drawbacks associated with the evaluation of the supremum and infinum of (9), the effects of leverage points on the bounds of the proportionate bias will be assessed by means of,

$$\sup_w \frac{w'Aw}{w'Bw} = \max(\lambda_i) , \qquad \inf_w \frac{w'Aw}{w'Bw} = \min(\lambda_i) , \tag{18}$$

[21]The same kind of reasoning could be made to $\min(\lambda_i)$.

with $A = (X'X)^{-1}X'QX(X'X)^{-1}$ and $B = (X'X)^{-1}X'\Omega X(X'X)^{-1}$. The advantage of this expression is that w does not depend on X and all the dynamic of this matrix is incorporated in A and B. Therefore these bounds can be controlled in an appropriated manner by changing the design of the X matrix. As before these effects will be assessed controlling the leverage associated with the 48^{th} point. Figures 8 a), b) and c) present these results for $m = 2$ and ρ equal to 0.2, 0.5 and 0.8, respectively. The first conclusion extracted from these pictures is that leverage points seem to affect both the lower and upper bounds. However, the way in which each one of these bounds is affected depends on the value of ρ. The second conclusion is that the effect on the lower bound is more important for small values of ρ. When $\rho = 0.8$, for example, this bound does not seem to be influenced for $h_{48,48} < 0.9$, approximately. However, while the lower bound remains constant the upper bound change with $h_{48,48}$. The third conclusion is that $h_{48,48}$ does not affect both, lower and upper bounds altogether (when one changes the other remains constant). Finally there is a possibility of upward bias for a moderate leverage associated with a high value of ρ. This is the case of the proportionate bias of the Newey and West estimator of $Cov(\widehat{\beta}_2)$, showed in figure 8c. However, this possibility should be evaluated with care, due to the non-monotonicity of the upper bound, relatively to $h_{48,48}$. When $\rho = 0.8$ the maximum bias is attainable for $h_{48,48}$ approximately equal to 0.5.

Figures 9 a), b) and c) present the same results for the case of MA(1) errors, with $m = 1$ and θ equal to $0.2, 0.5$ and 0.8, respectively. The only difference to the above case is that the bounds appear to be less pronounced and more robusts to leverage, particularly the upper bound.

4.4 Modified Newey and West Estimator

In this Section a slight modification of the Newey and West estimator is presented. This modification is based on an additive bias correction for the particular case of independent and homoskedastic errors. Multiplicative corrections are also possible but only for the diagonal entries of $X'\widehat{\Omega}X$. With independent errors the autocorrelation structure of the residuals can only be eliminated by an additive term.

It is well known that even when errors are homoskedastic and independent, conventional least-squares residuals never have these properties. In particular $E(\widehat{\varepsilon}_i^2) = \sigma^2(1-h_{ii})$ and $E(\widehat{\varepsilon}_i\widehat{\varepsilon}_j) = -\sigma^2 h_{ij}$. Therefore instead of $X'\widehat{\Omega}X$ given in expression (7) the following estimator of $X'\Omega X$ is suggested,

$$X'\widehat{\Omega}^*X = \sum_{i=1}^{n}(\widehat{\varepsilon}_i^2 + \widehat{\sigma}^2 h_{ii})x_i'x_i + \sum_{j=1}^{m}\sum_{i=j+1}^{n}\kappa_j\,(\widehat{\varepsilon}_i\widehat{\varepsilon}_{i-j} + \widehat{\sigma}^2 h_{i,i-j})(x_i'x_{i-j} + x_{i-j}'x_i)\ , \quad (19)$$

where $\widehat{\sigma}^2 = \sum_{i=1}^{n}\widehat{\varepsilon}_i^2/(n-k)$. With homoskedastic and independent errors, $E(\widehat{\Sigma}_w^*) = \Sigma_w = \sigma^2(X'X)^{-1}$. Moreover this estimator (as well the Newey and West estimator) is asymptotically unbiased since $h_{ii} \to 0$ as $n \to \infty$.

To prove the positive semi-definiteness of the modified Newey and West estimator, consider (19) expressed in matrix form as,

$$X'\widehat{\Omega}^*X = X'\widehat{\Omega}X + \widehat{\sigma}^2 X'QX$$

with the elements of Q defined as the symmetric of (11).

Theorem 2: The modified Newey and West estimator is positive semi-definite.

Proof. For a proof it is sufficient to apply theorem 1 of Newey and West (1987) to the second term of the right hand side of the above expression. By doing that $X'QX$

is positive semi-definite and the result of the theorem follows.□

A bias comparison between the modified Newey and West, Newey and West and LS estimators are pictured in figure 10 and 11, using the Mackinnon and White data, for AR(1) and MA(1) errors, respectively. As it can be seen by these pictures, the proportionate bias of the modified Newey and West Estimator associated with $\widehat{\beta}_3$ performs better than the other estimators for the values of ρ and θ in the interval $[0, 0.9]$. However, this conclusion is not clear if the proportionate bias is computed in a direction of a covariate with an important leverage as can be seen from pictures 10a and 11a.

5. THE FINITE-SAMPLE DISTRIBUTIONS OF HETEROSKEDASTICITY AND AUTOCORRELATION ROBUST WALD STATISTIC

The subject of this Section is a natural extension of the results presented above. In the econometric framework, it is well known that hypothesis tests require a consistent estimator of scale. However when this estimator presents an important bias, as in the case of the Newey and West estimator and the sample size is not large enough, it is important to know how this fact can affect inferences based on the asymptotic distribution. For this reason this Section is concerned with the performance of the first order asymptotic normal approximation to the finite-sample distribution of the Wald test statistic, when the variance estimator of $\widehat{\beta}$ is HAC type.

Mackinnon and White (1985) provided Monte Carlo estimates of the exact sizes of nominal 5% and 1% two-sided tests under the null hypothesis, using different HCCM estimators of $\widehat{\beta}$. However the estimates obtained by Monte Carlo simulation depend

on the regression design. Thus, the conclusions derived from this procedure can be very misleading, as shown by Chesher and Austin (1991). These authors derived the exact finite-sample distributions of heteroskedasticity robust Wald statistics of scalar linear hypotheses in the normal linear model. Moreover they showed a straightforward relation between the regression design and the quality of these approximations.

In this Section their methodology is extended to the case in which the variance estimator of $\widehat{\beta}$ is HAC type. In particular, the quality of the asymptotic normal approximation to the finite sample distribution of the test is assessed in the case of independent and AR(1) errors for different configurations of the design of the X matrix.

Considering the linear model presented in Section 2, the Wald test statistic for the hypothesis H_0: $w'\beta = w_0$ is given by,

$$t = \frac{w'\widehat{\beta} - w_0}{(w'\widehat{\Sigma}w)^{1/2}} \tag{20}$$

In the following, consider that the null hypothesis is true. Thus, if X obey to some suitable conditions as the Grenander conditions [see for example Judge et al. (1985)], t converges in law to the standard normal distribution. Deviations of the finite sample distribution of t from its asymptotic distribution will be the measure of the quality of this test statistic.

Assuming $w'\beta = w_0$, the numerator of t^2 can be written as,

$$(w'\widehat{\beta} - w_0)^2 = u'Au$$

where $A = \Omega^{1/2} z z' \Omega^{1/2}$ and z was defined in Section 4. For the denominator one has as well,

$$w' \Sigma w = z' \widehat{\Omega} z = u' B u$$

where $B = \sum_{i=1}^{n} z_i \Omega^{1/2} m_i m_i' \Omega^{1/2} z_i + 2 \sum_{j=1}^{m} \sum_{i=j+1}^{n} \kappa_{i-j} z_j \Omega^{1/2} m_j m_i' \Omega^{1/2} z_i$. Both, numerator and denominator of t^2, are quadratic forms in normal variables and the distribution of t^2 can be calculated using the procedure given by Imhof (1961). Due to the fact that A and B are symmetric matrices, it follows that $A - c^2 B$ is also symmetric, being possible the spectral decomposition, $A - c^2 B = L \Delta L'$, with $L'L = I$. As a consequence, when $w' \widehat{\beta} = w_0$ one has,

$$P(t^2 < c^2) = P[u'(A - c^2 B)u < 0] = P[\sum_{i=1}^{n} r_i^2 \delta_i < 0] \ , \tag{21}$$

where $r_i = u' l_i \sim N(0,1)$, l_i is the ith column of L and δ_i the eigenvalue associated with l_i. Finally, the probability in the left hand side of (21) is computed by the following,

$$P(t^2 < c^2) = P[\sum_{i=1}^{n} r_i^2 \delta_i < 0] = \frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} \frac{\sin \theta(v)}{v \rho(v)} dv \ , \tag{22}$$

where,

$$\theta(v) = 0.5 \sum_{i=1}^{n} \arctan(\delta_i v)$$

$$\rho(v) = \prod_{i=1}^{n} (1 + \delta_i^2 v^2)^{1/4}$$

Because the null distribution of t is symmetric around zero, its determination from the distribution of t^2 is possible, having in mind the relation,

$$P(t^2 < c^2) = 2P(t < c) - 1$$

for any scalar $c > 0$.

Considering the linear model (1) and the Mackinnon and White data, Figures[22] 12 to 15 show upper halves of the exact finite-sample distribution functions of the Wald tests, using the Newey and West estimator with a fixed lag truncation, when the errors are independent and AR(1). In this case $\rho = 0.2$, $\rho = 0.5$ and $\rho = 0.8$, respectively. In each of these Figures one considers the hypotheses $\beta_2 = 0$ on the left and $\beta_3 = 0$ on the right.

From these pictures, one concludes that the exact distribution is less well approximated by the standard normal distribution as the correlation coefficient increases. However the quality of this approximation seems to depend on the particular hypotheses considered. The distribution function of the Wald test moves far to the right of the standard normal in the $\beta_3 = 0$ case and to the left in the $\beta_2 = 0$ case. As seen before the main reason for this difference in the shape of the exact distributions is due to the presence of the 48^{th} leverage point. Deleting this point [figure not presented] both distributions present the same shape.

A more accurate way to evaluate the dependence of the exact distribution on the regression design is by controlling the leverage associated with one particular observation, following the same methodology described in Section 3. Figures 16 to 18 show

[22]Each point of the exact distribution was computed using expressions (22) and (23).

upper halves of the exact distribution for different values of the autocorrelation of the errors. In each of these pictures the 48^{th} point is equal to $(1, \delta)$, with $\delta \in \Re$ replaced by some appropriated value, corresponding to $h_{48,48}$ equal to 0.13, 0.34, 0.56 and 0.81. With independent errors [see figure 16], as the leverage associated with the 48^{th} point increases, the exact distribution moves far to the right of the standard normal, presenting long tails. However, with AR(1) errors, particularly for high values of ρ, this effect should be evaluated with care, due to the non monotonicity between $h_{48,48}$ and the shape of the exact distribution. For low and high values of $h_{48,48}$ the exact distribution is to the right of the standard normal. For moderate values of $h_{48,48}$ the exact distribution moves to the left of the standard normal, presenting a greater peak. A possible reason for this behaviour can be due to the upward bias of the Newey and West estimator, as seen in subsection 4.3.

Figure 1a: Proportionate Bias of NW estimator (β_2)

AR(1) Errors

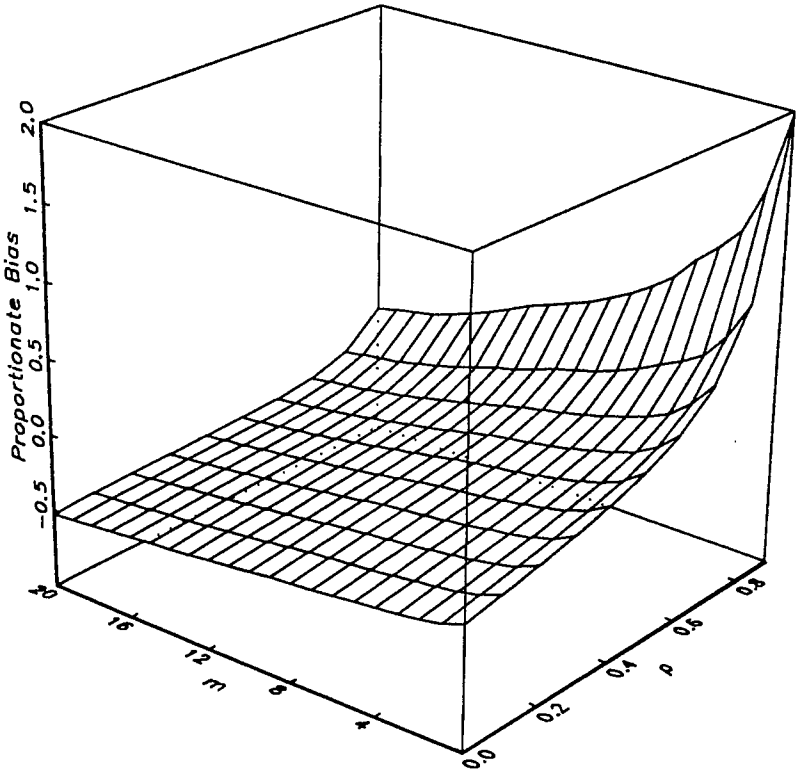


Figure 1b: Proportionate Bias of NW estimator (β_3)

AR(1) Errors

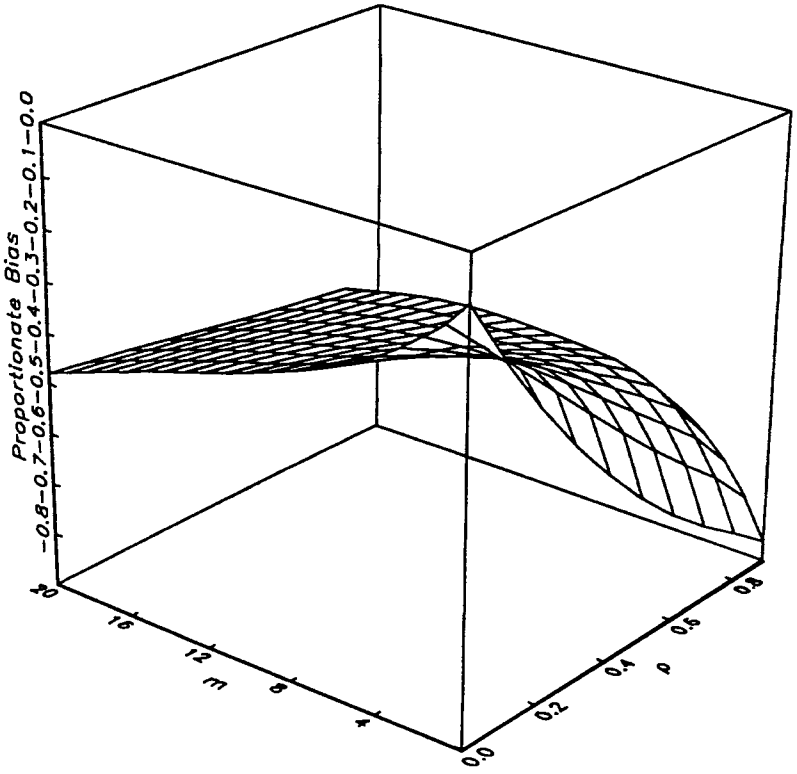


Figure 2a: Proportionate Bias of NW estimator (β_2)

MA(1) Errors

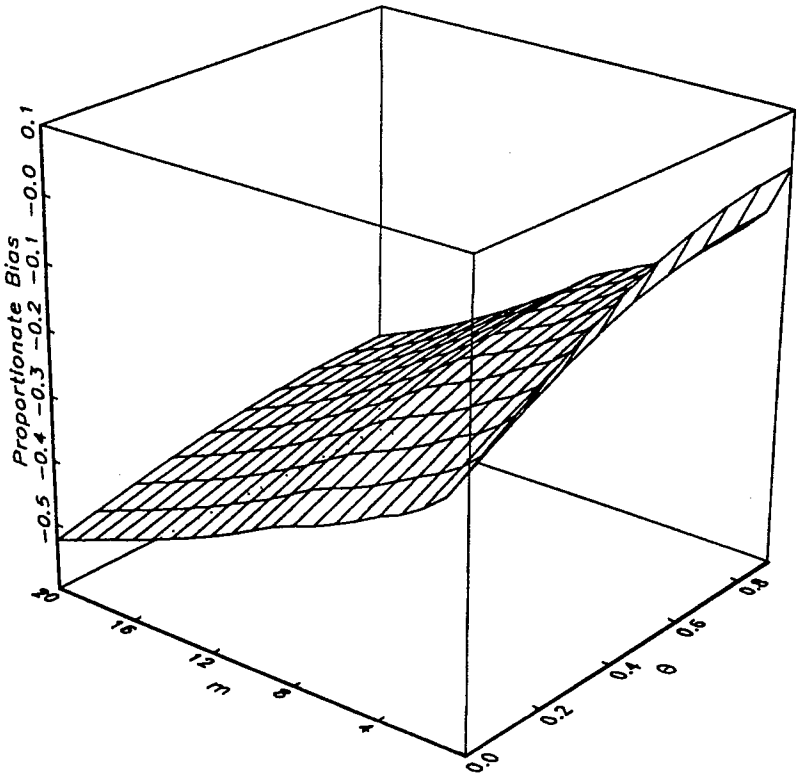


Figure 2b: Proportionate Bias of NW estimator (β_3)

MA(1) Errors

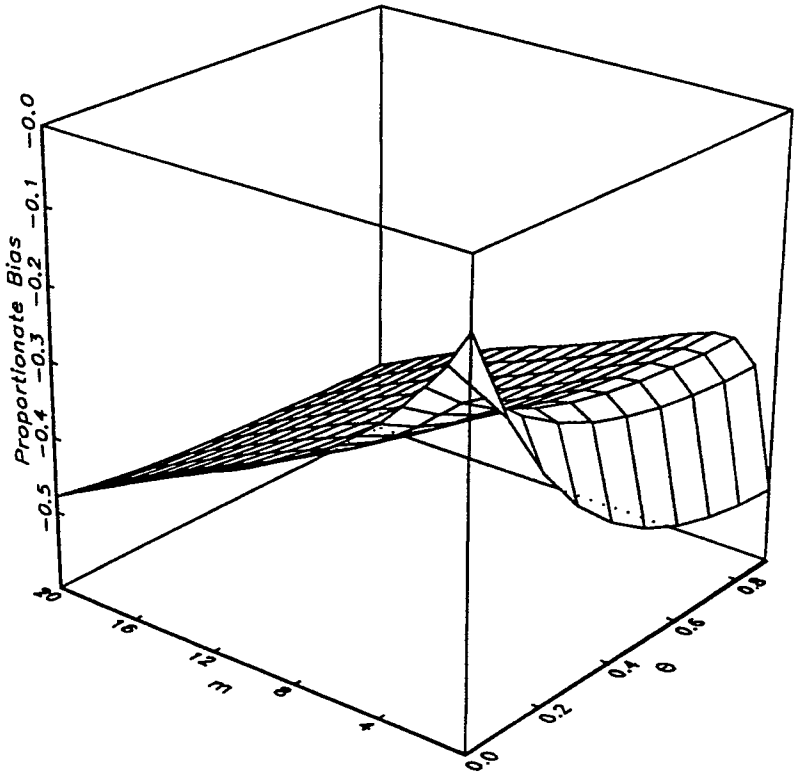


Figure 3: Proportionate Bias of Newey and West Estimator

$\rho=0$, m=1, N=50

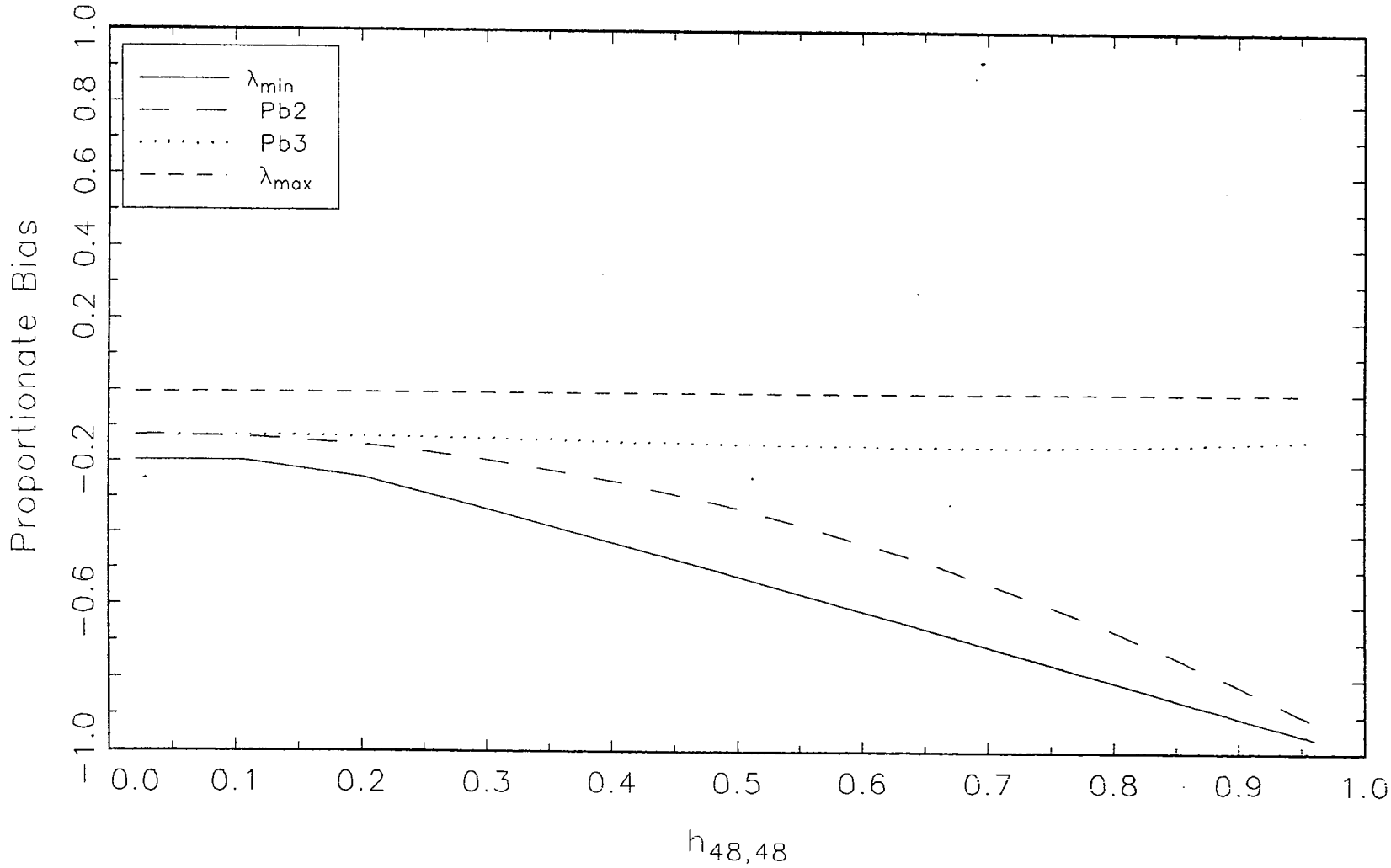


Figure 4: Proportionate Bias of Newey and West Estimator

One regressor, $x_1=\delta$, $x_i=(-1)^{i+1}$ $\rho=0$, m=1, N=50

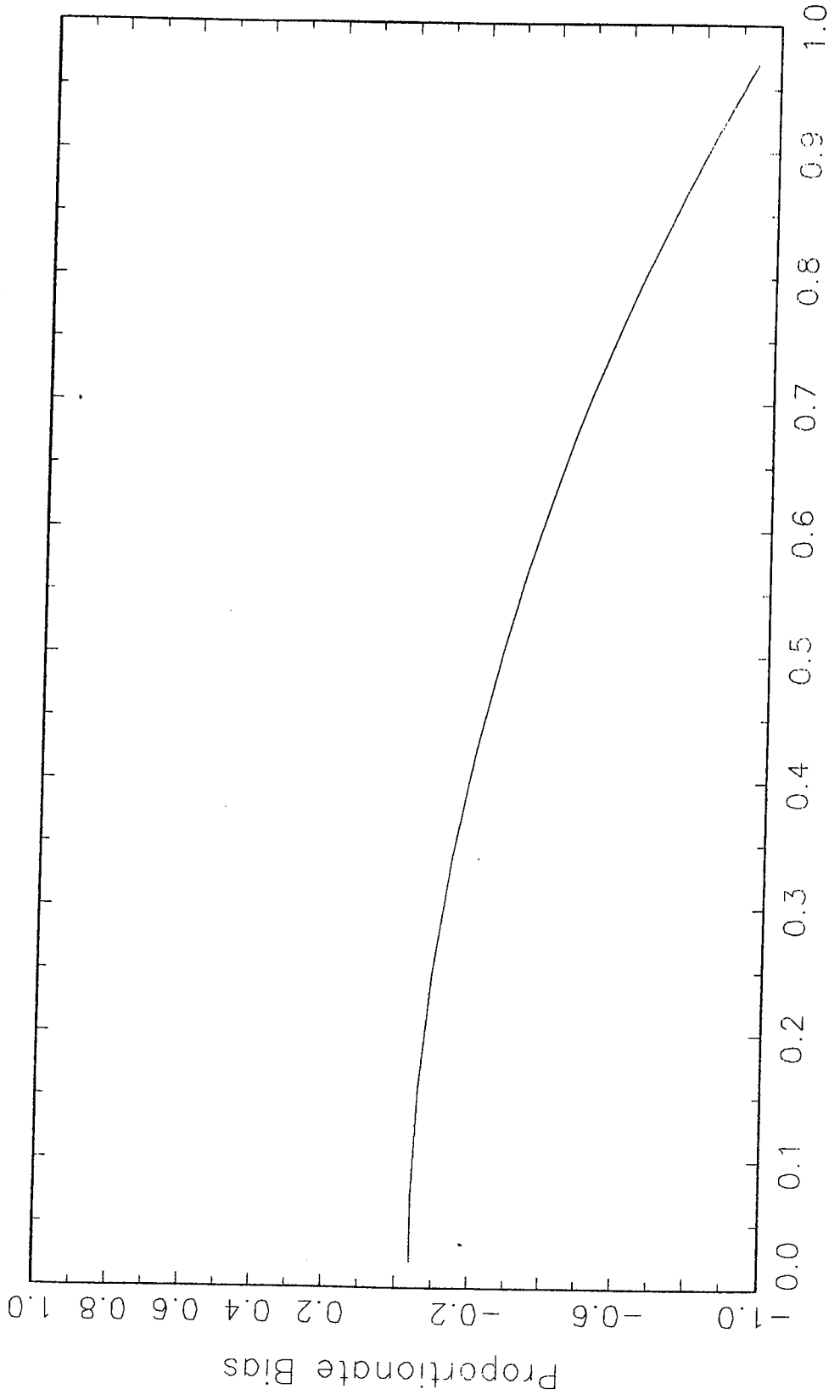


Figure 5a: Pr. Bias of NW and LS Estimator (β_2)

Mackinnon and White Data, AR(1) Errors

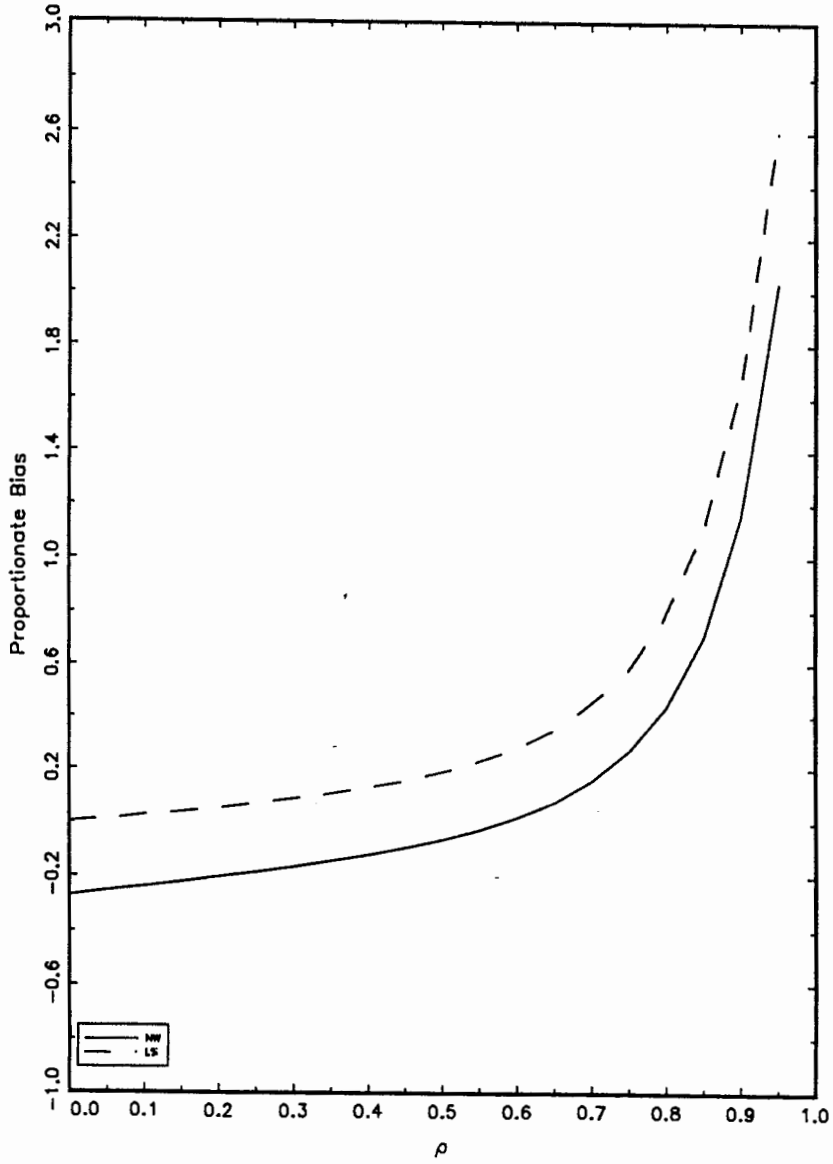


Figure 5b: Pr. Bias of NW and LS Estimator (β_3)

Mackinnon and White Data, AR(1) Errors

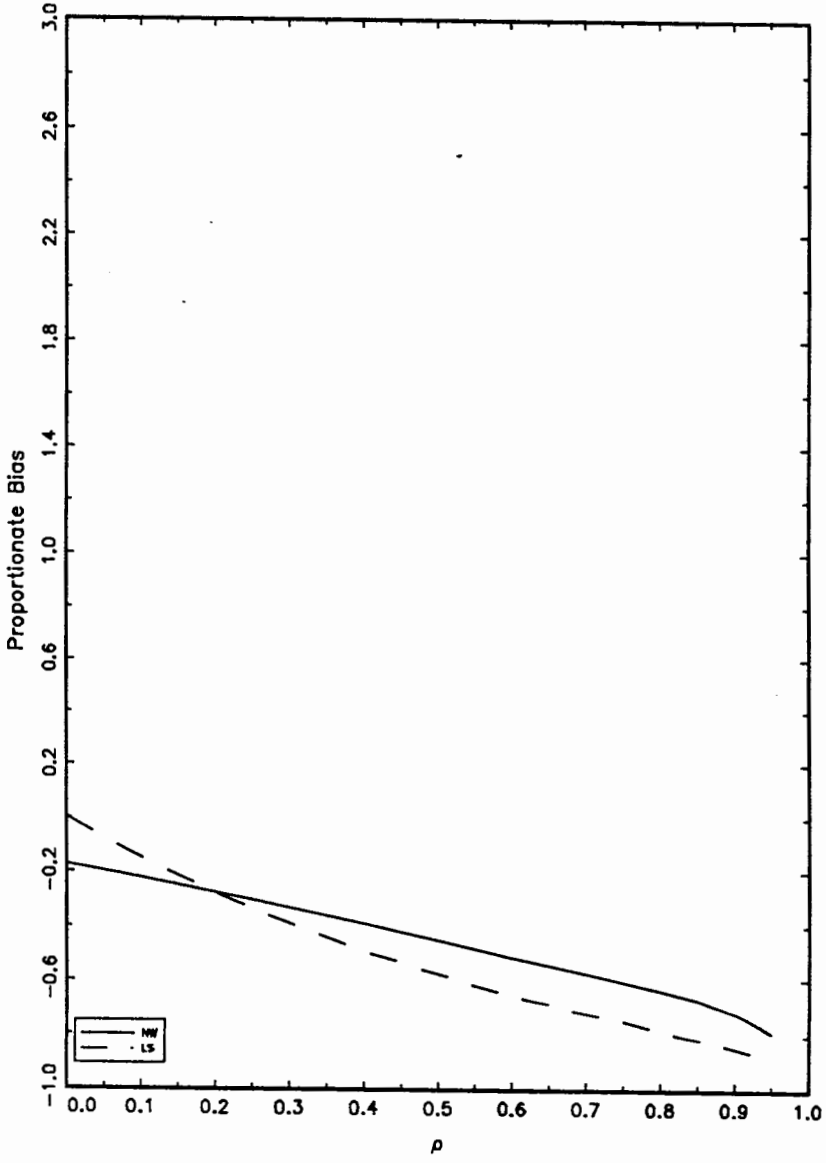


Figure 6a: Pr. Bias of NW and LS Estimator (β_2)

Mackinnon and White Data, 48th point removed, AR(1) Errors

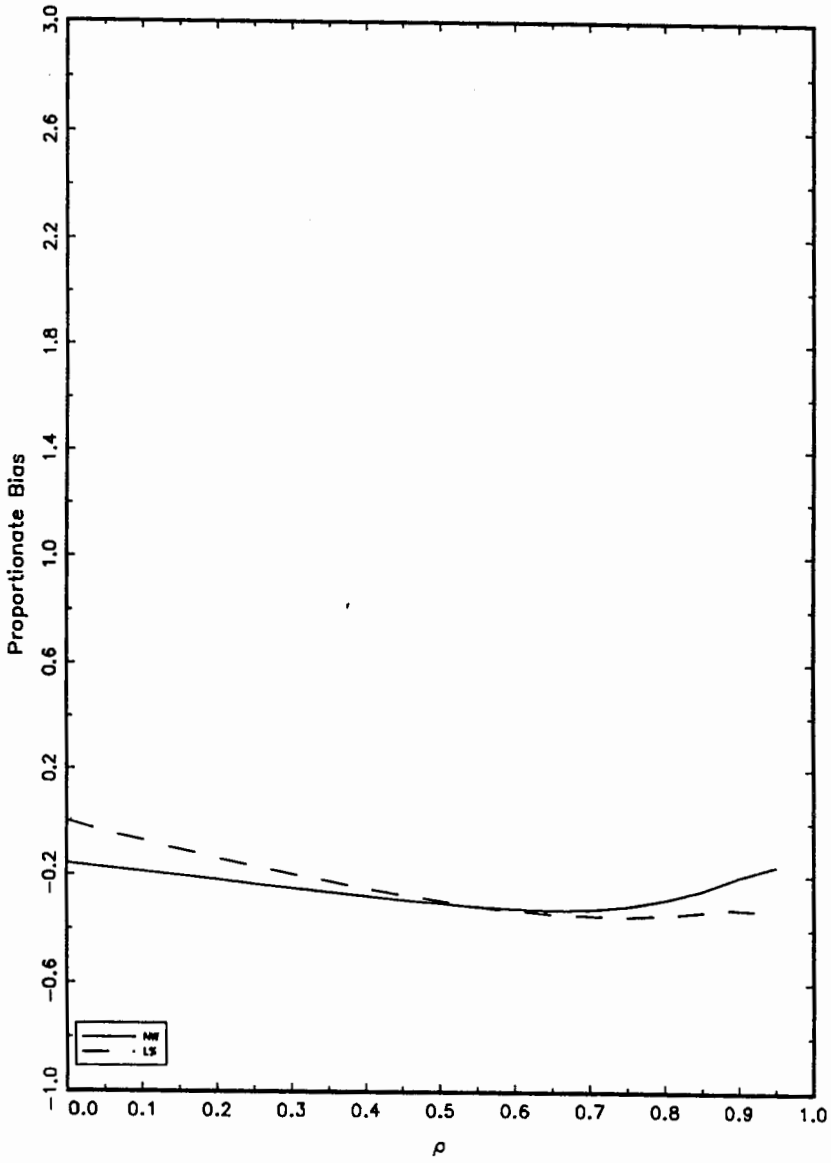


Figure 6b: Pr. Bias of NW and LS Estimator (β_3)

Mackinnon and White Data, 48th point removed, AR(1) Errors

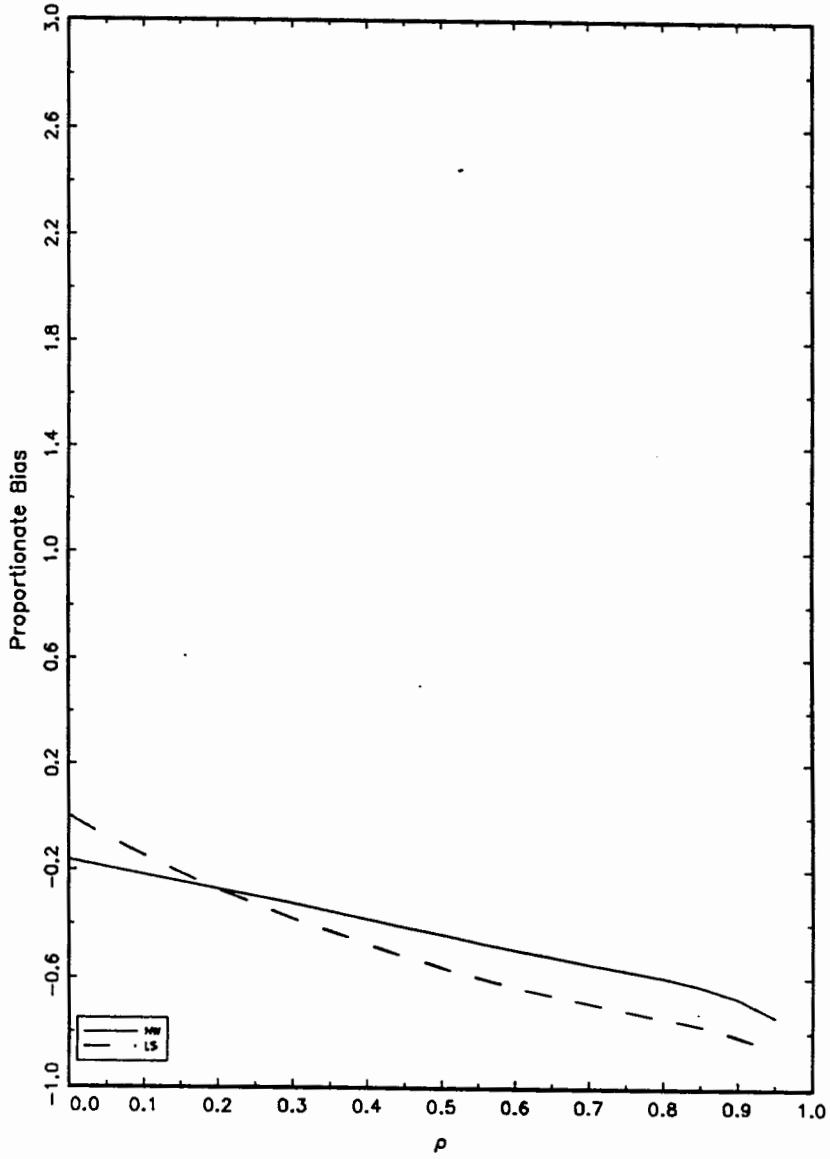


Figure 7: Bounds for the Proportionate Bias of NW Estimator

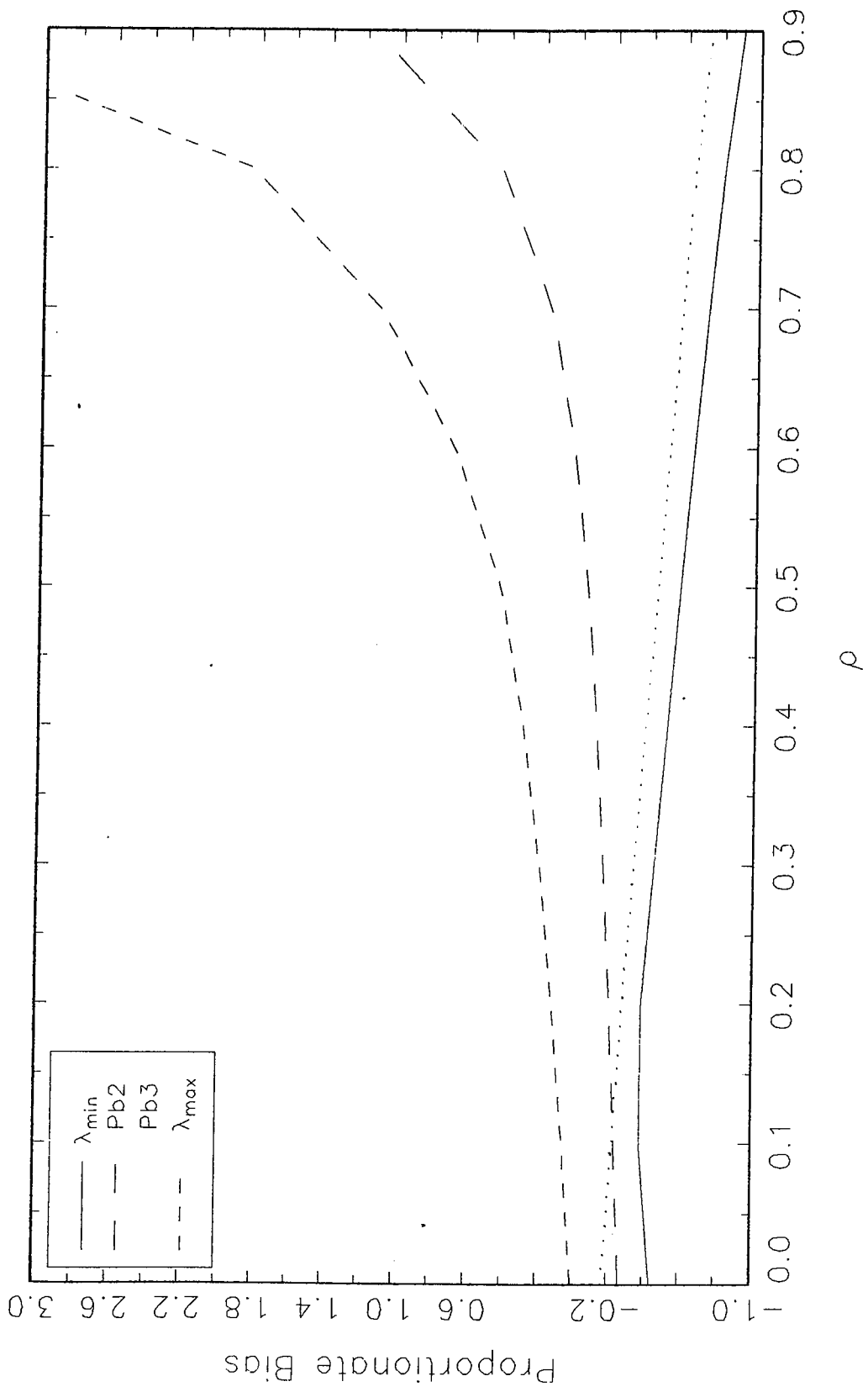


Figure 8a: Proportionate Bias of Newey and West Estimator

ρ=0.2, m=2, N=50

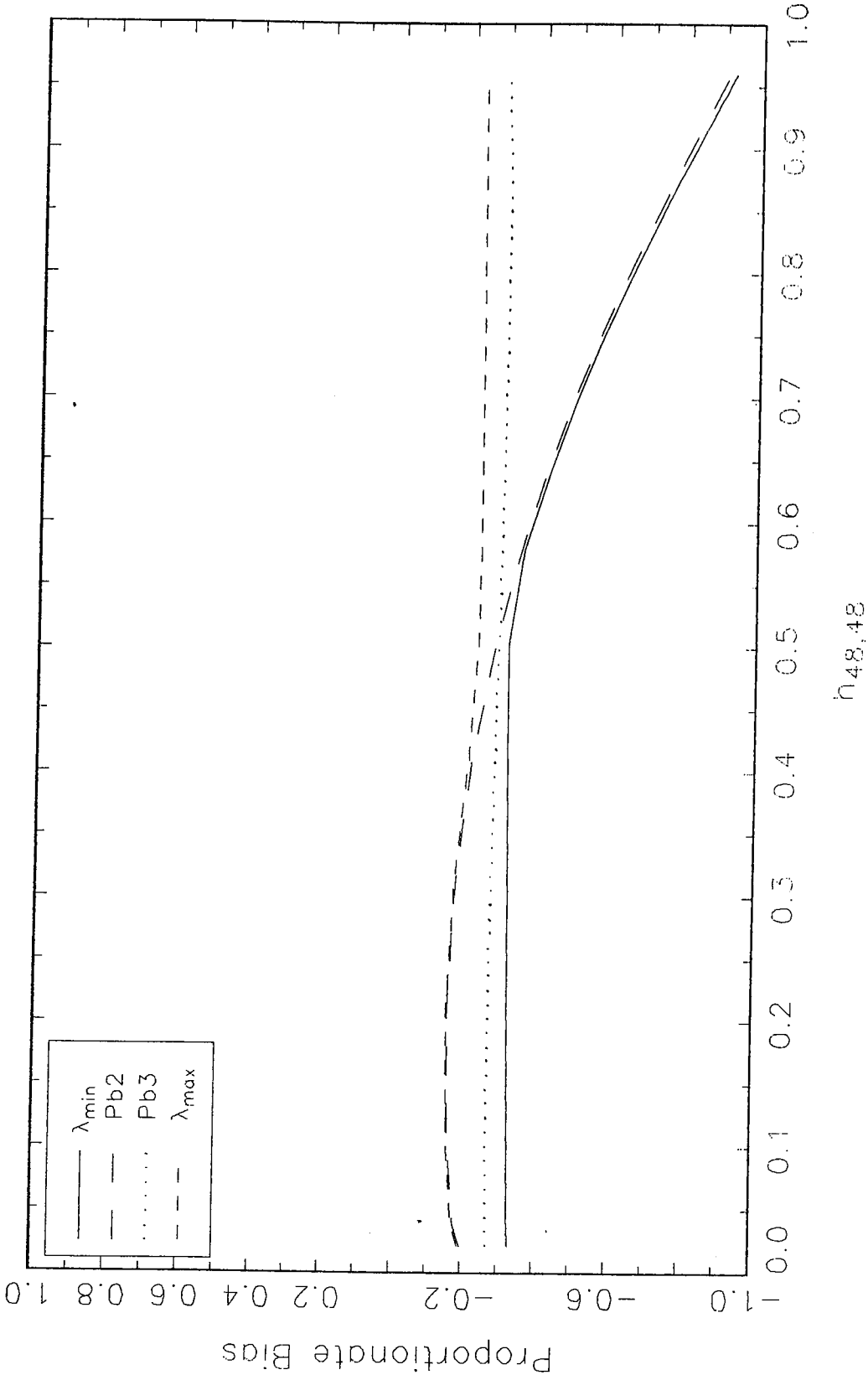


Figure 8b: Proportionate Bias of Newey and West Estimator

ρ=0.5, m=2, N=50

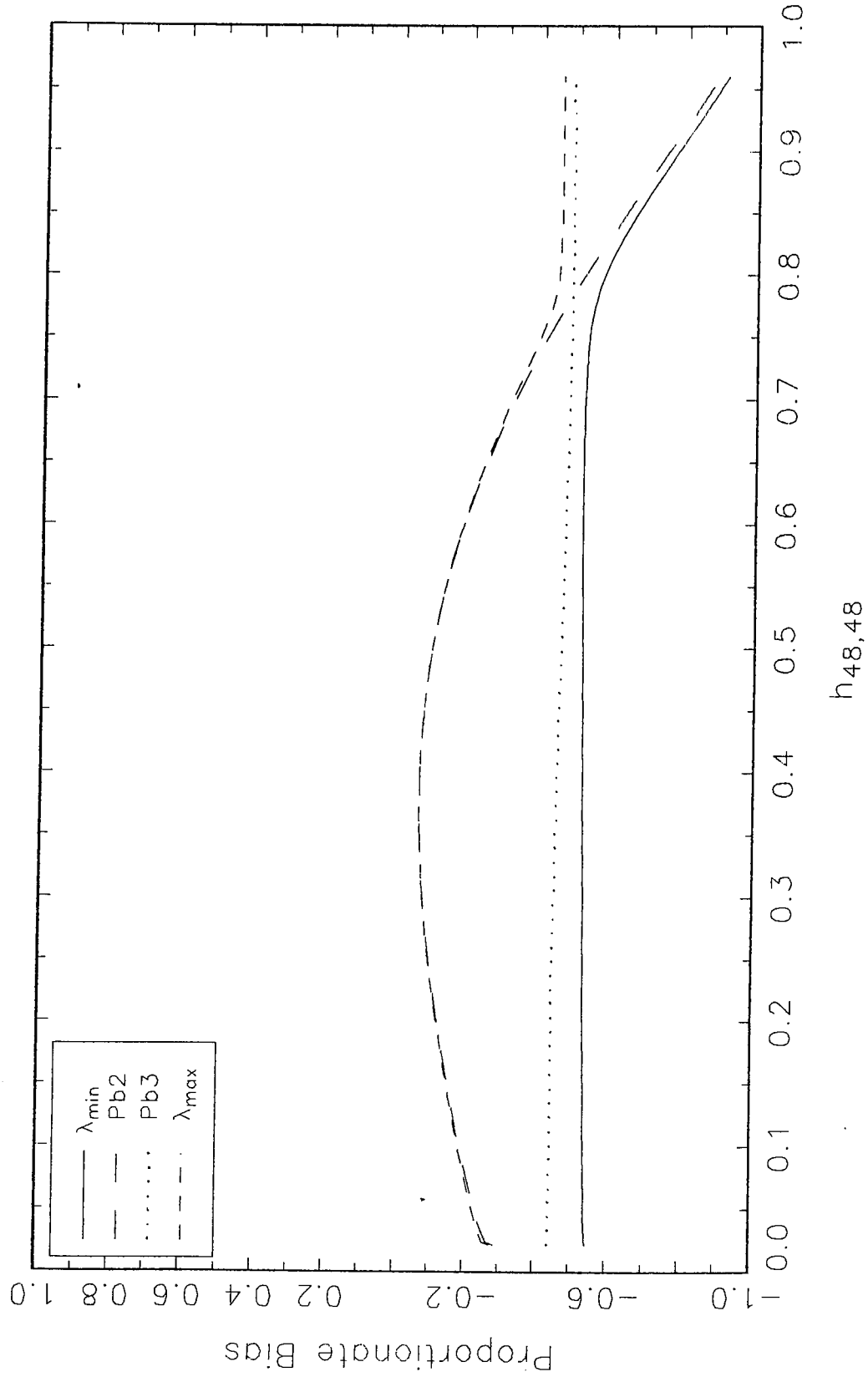


Figure 8c: Proportionate Bias of Newey and West Estimator

ρ=0.8, m=2, N=50

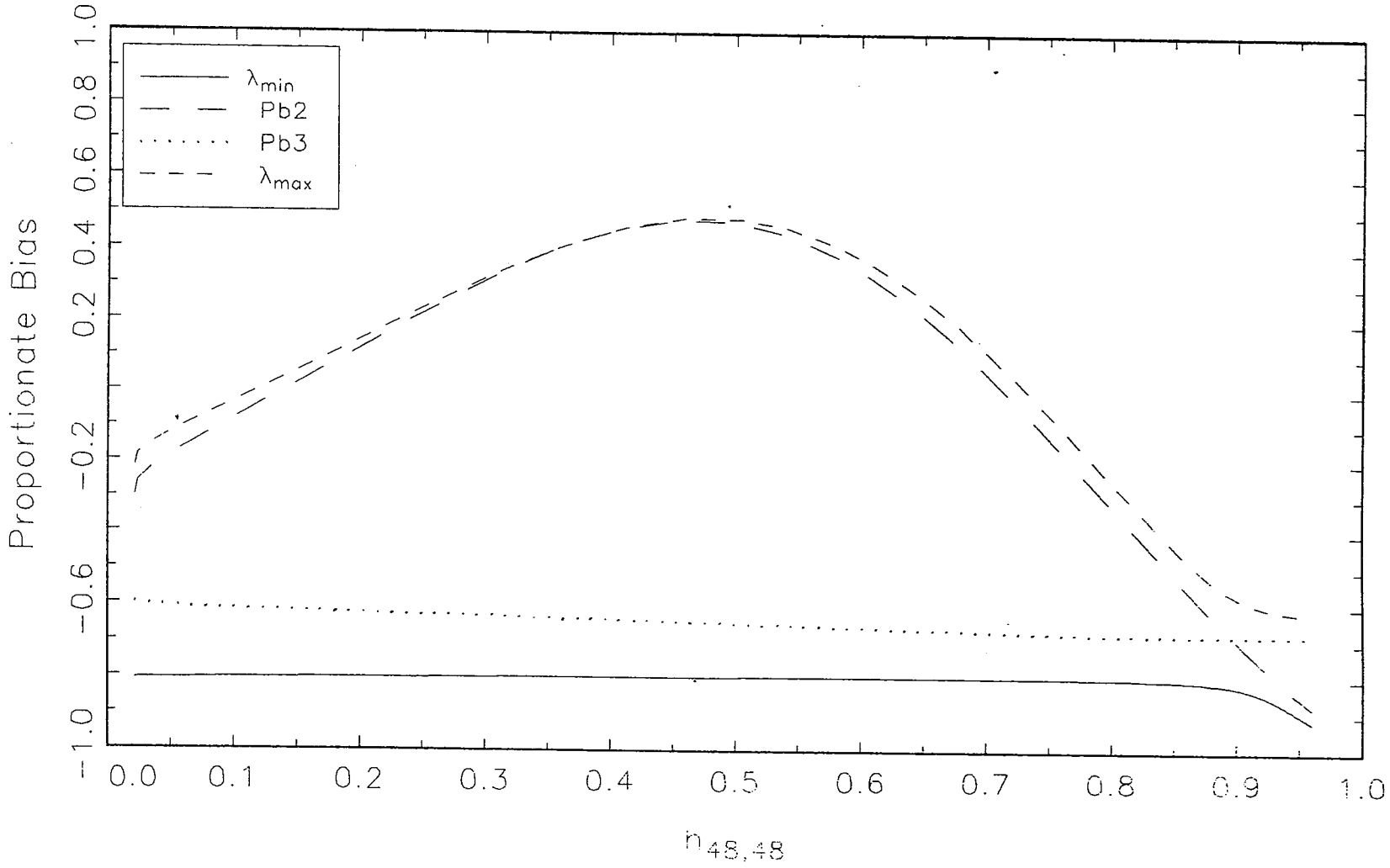


Figure 9a: Proportionate Bias of Newey and West Estimator

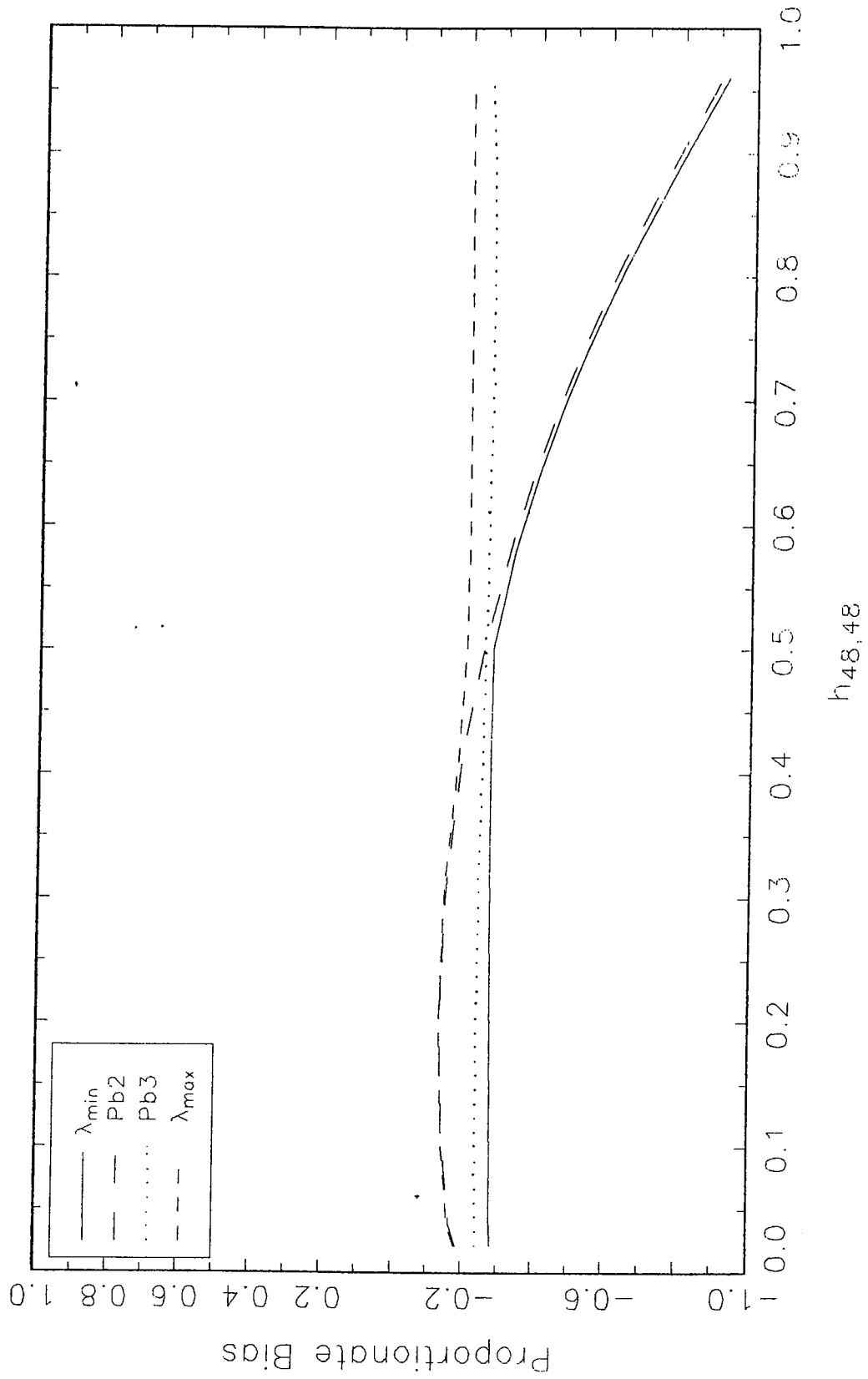


Figure 9b: Proportionate Bias of Newey and West Estimator

MA(1) Errors with Θ=0.5, m=1, N=50

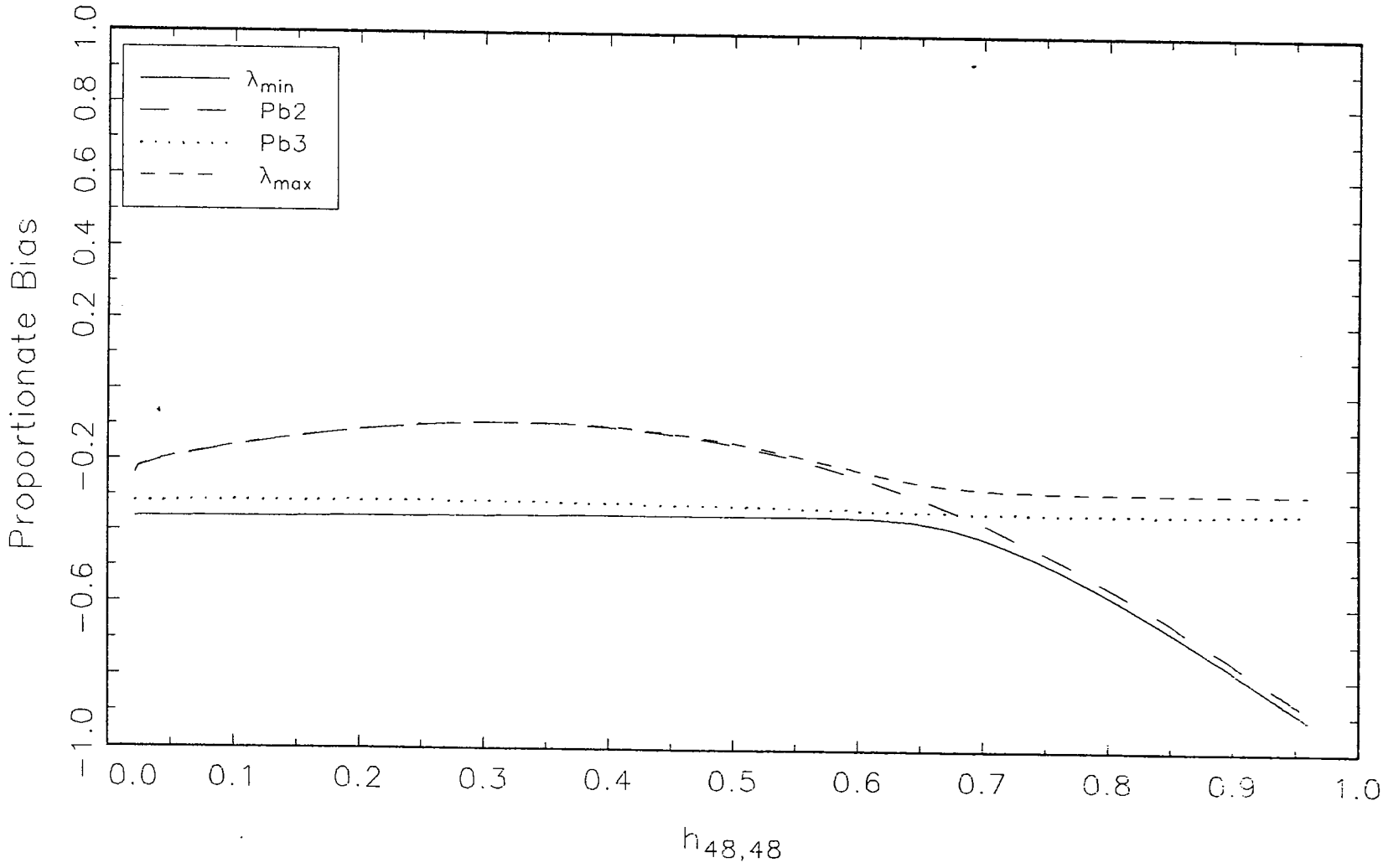


Figure 9c: Proportionate Bias of Newey and West Estimator

MA(1) Errors with $\Theta=0.8$, $m=1$, $N=50$

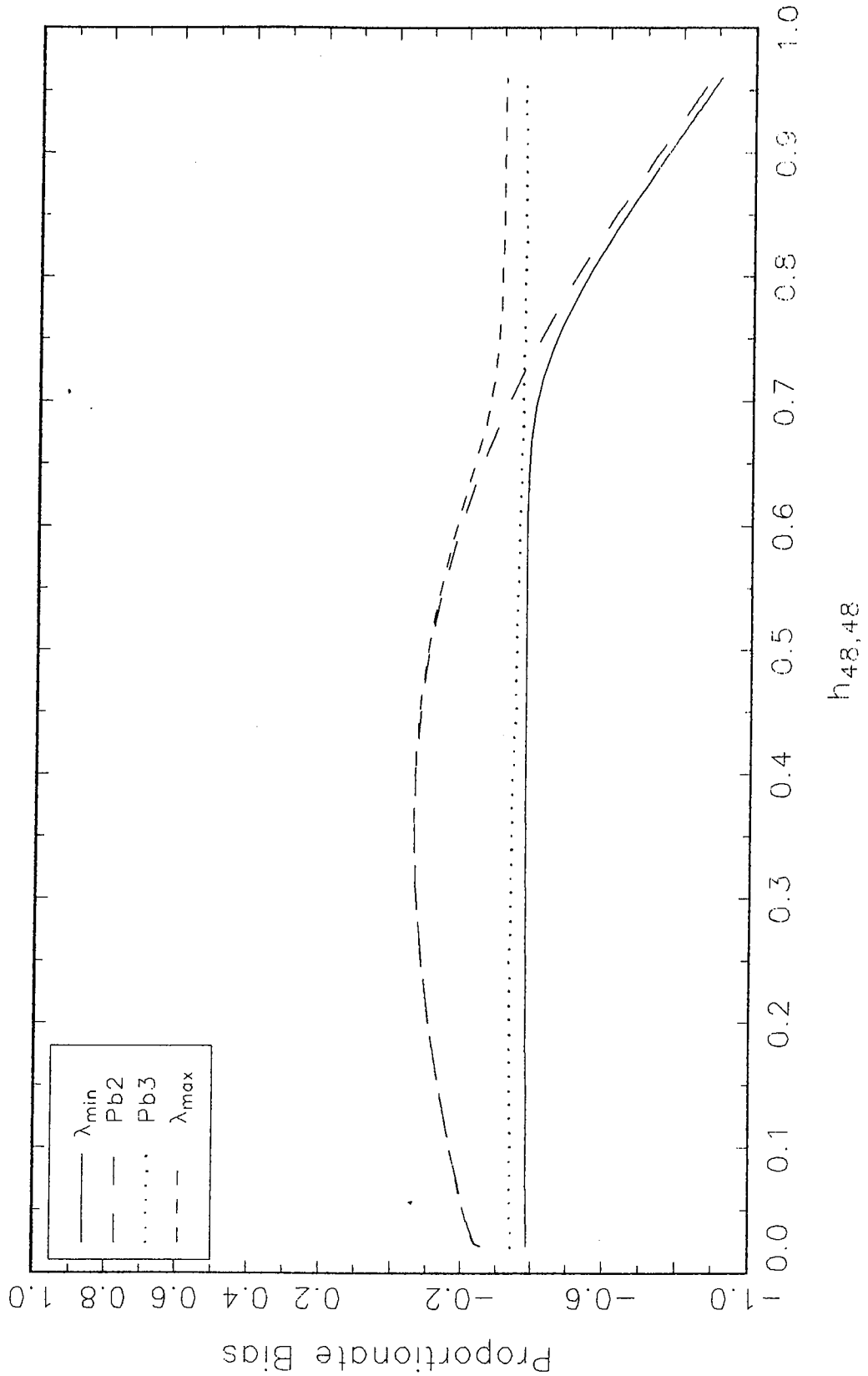


Figure 10a: Pr. Bias of NW, M_NW and LS Estimator (β_2)

Mackinnon and White Data, AR(1) Errors

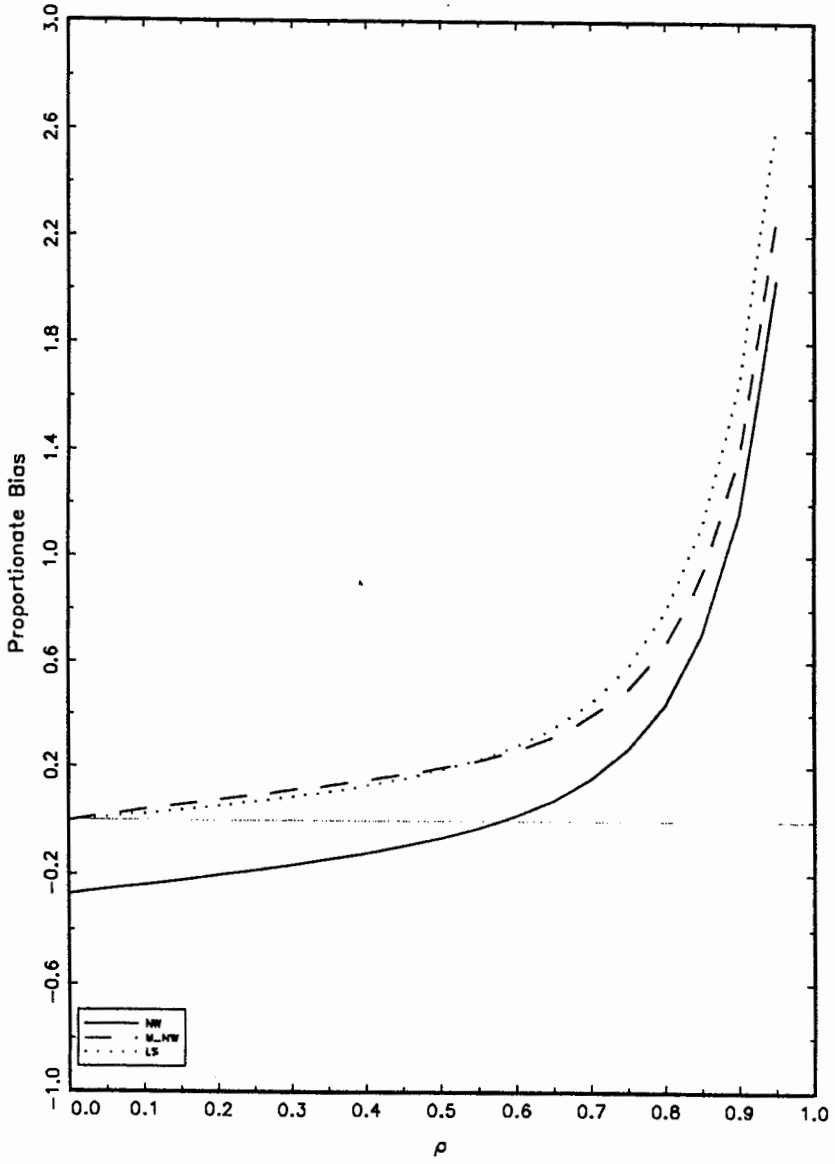


Figure 10b: Pr. Bias of NW, M_NW and LS Estimator (β_3)

Mackinnon and White Data, AR(1) Errors

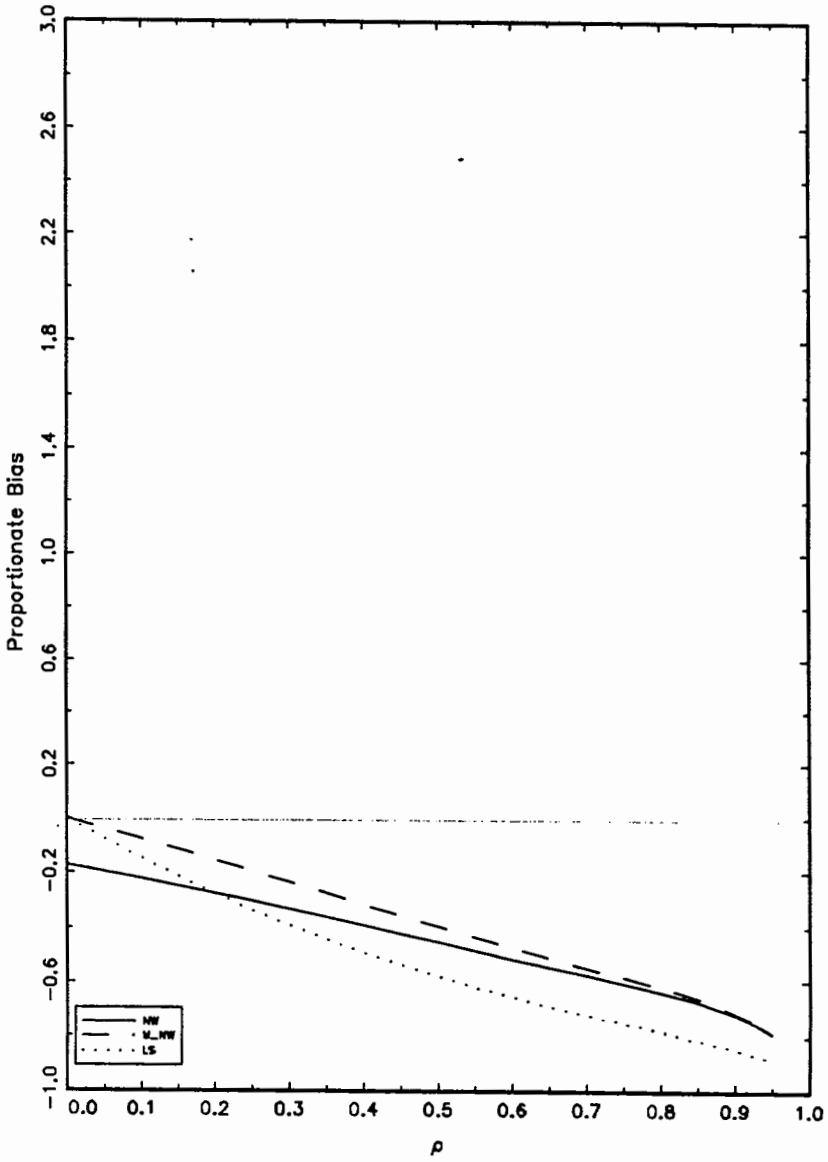


Figure 11a: Pr. Bias of NW, M_NW and LS Estimator (β_2)

Mackinnon and White Data, MA(1) Errors

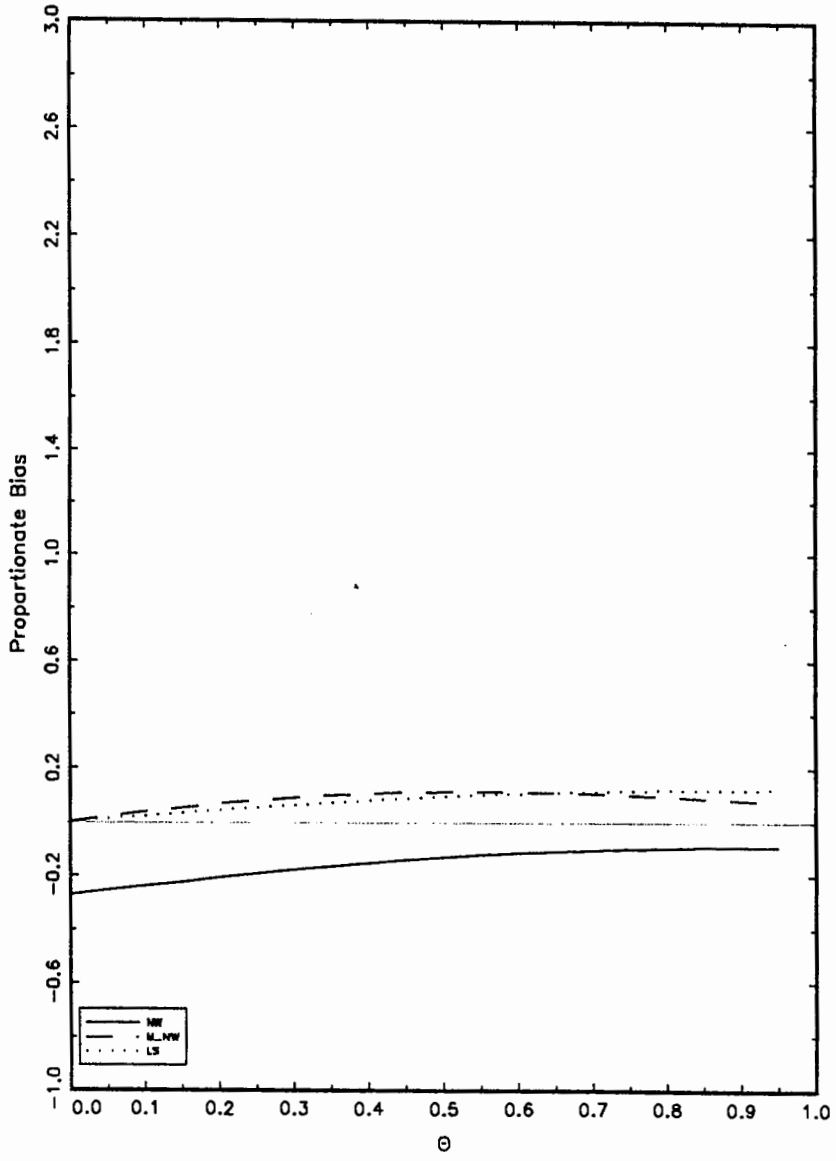


Figure 11b: Pr. Bias of NW, M_NW and LS Estimator (β_3)

Mackinnon and White Data, MA(1) Errors

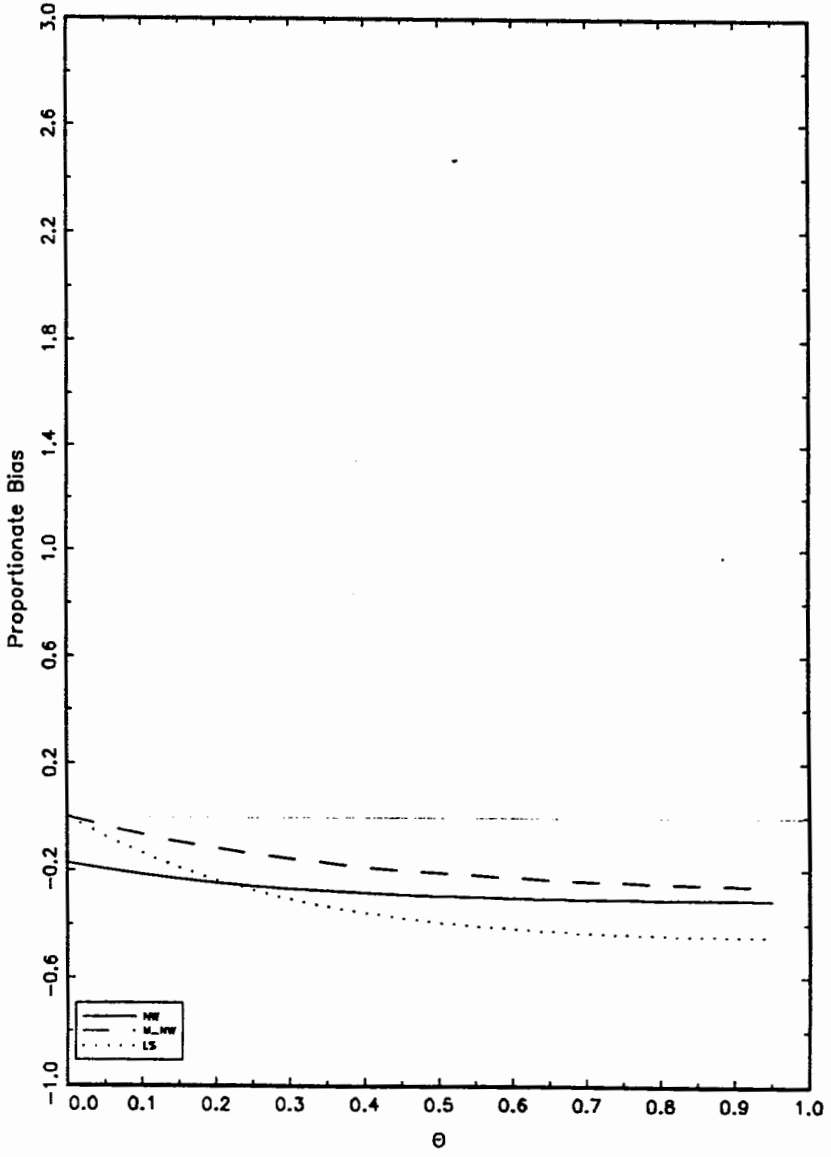


Figure 12a: Exact Distribution Function and Normal Aproximation

AR(1) Errors (ρ=0.0), m=2, N=50

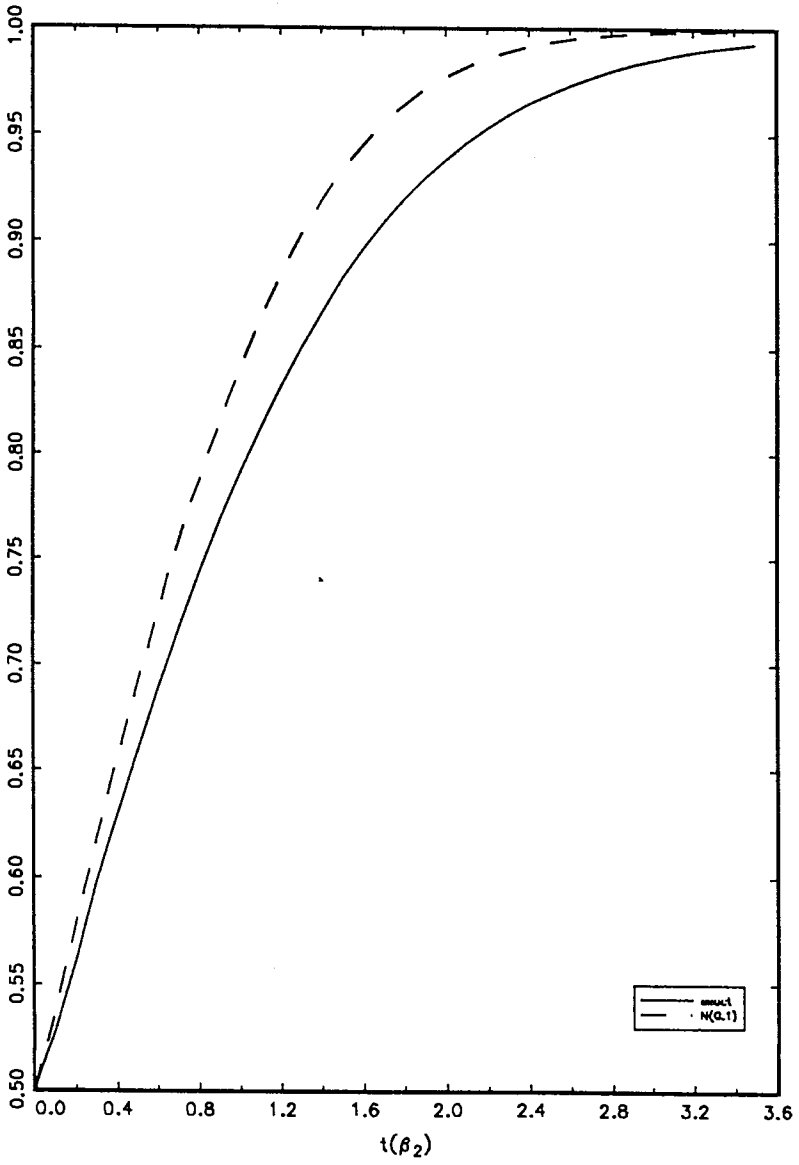


Figure 12b: Exact Distribution Function and Normal Aproximation

AR(1) Errors (ρ=0.0), m=2, N=50

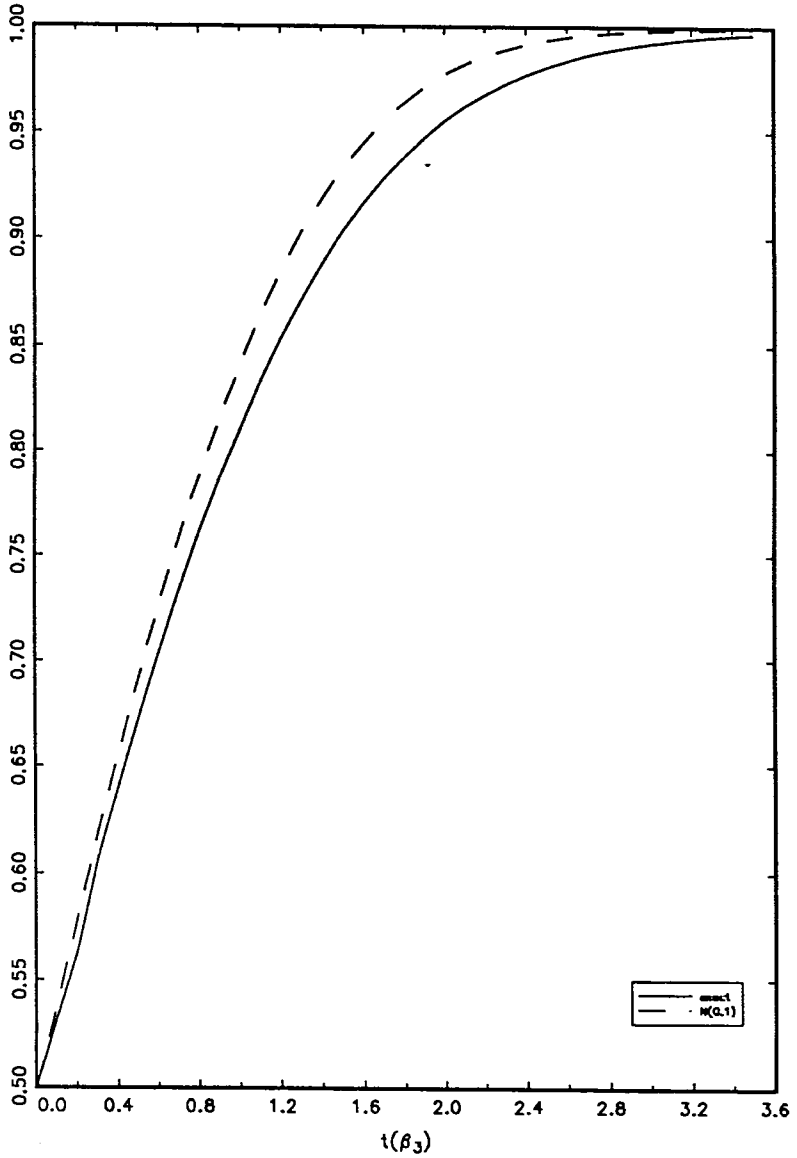


Figure 16: Exact Distribution Function

Leverage Associated with the 48^{th} point

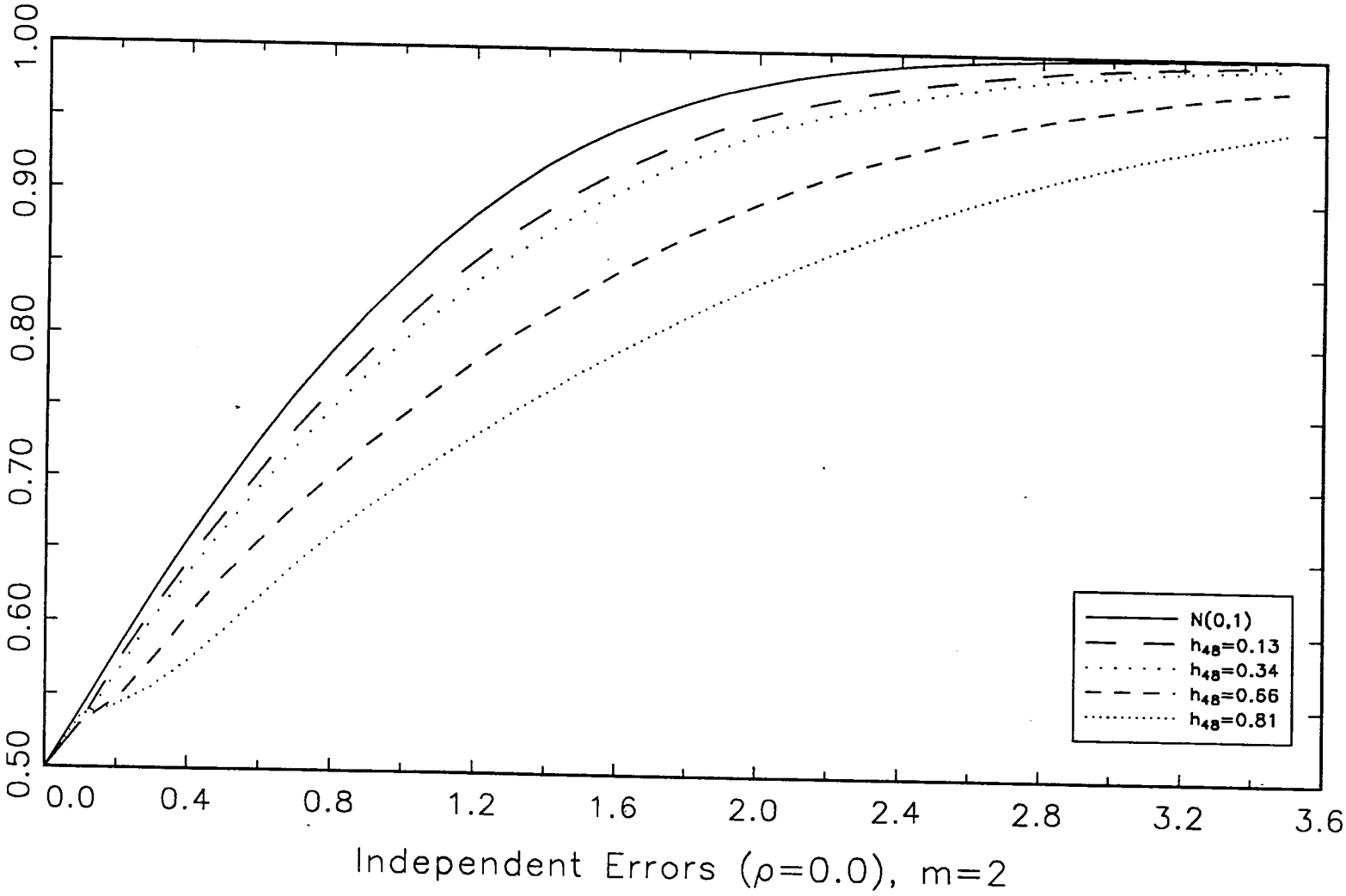


Figure 17: Exact Distribution Function

Leverage Associated with the 48^{th} point

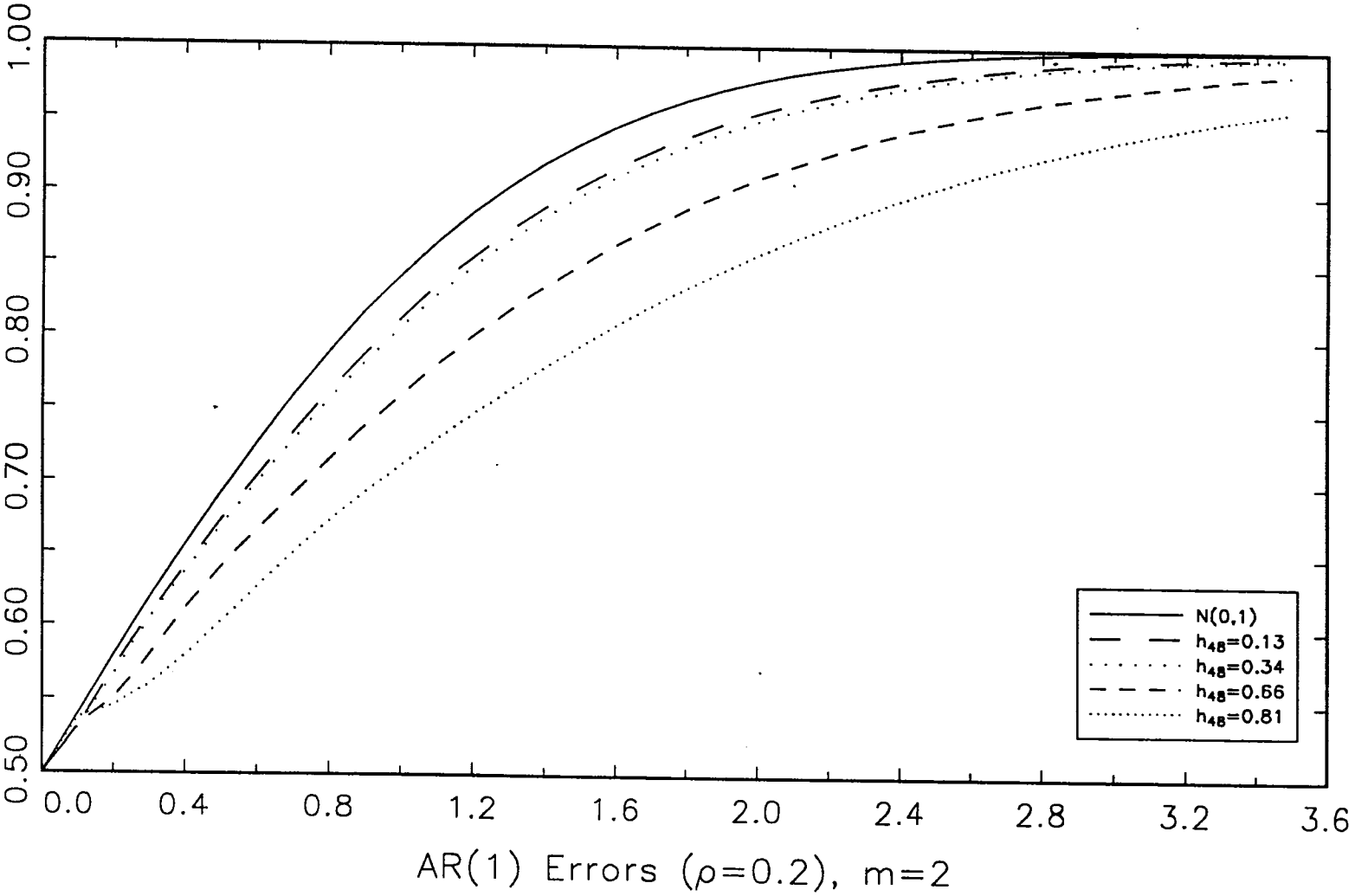


Figure 18: Exact Distribution Function

Leverage Associated with the 48th point

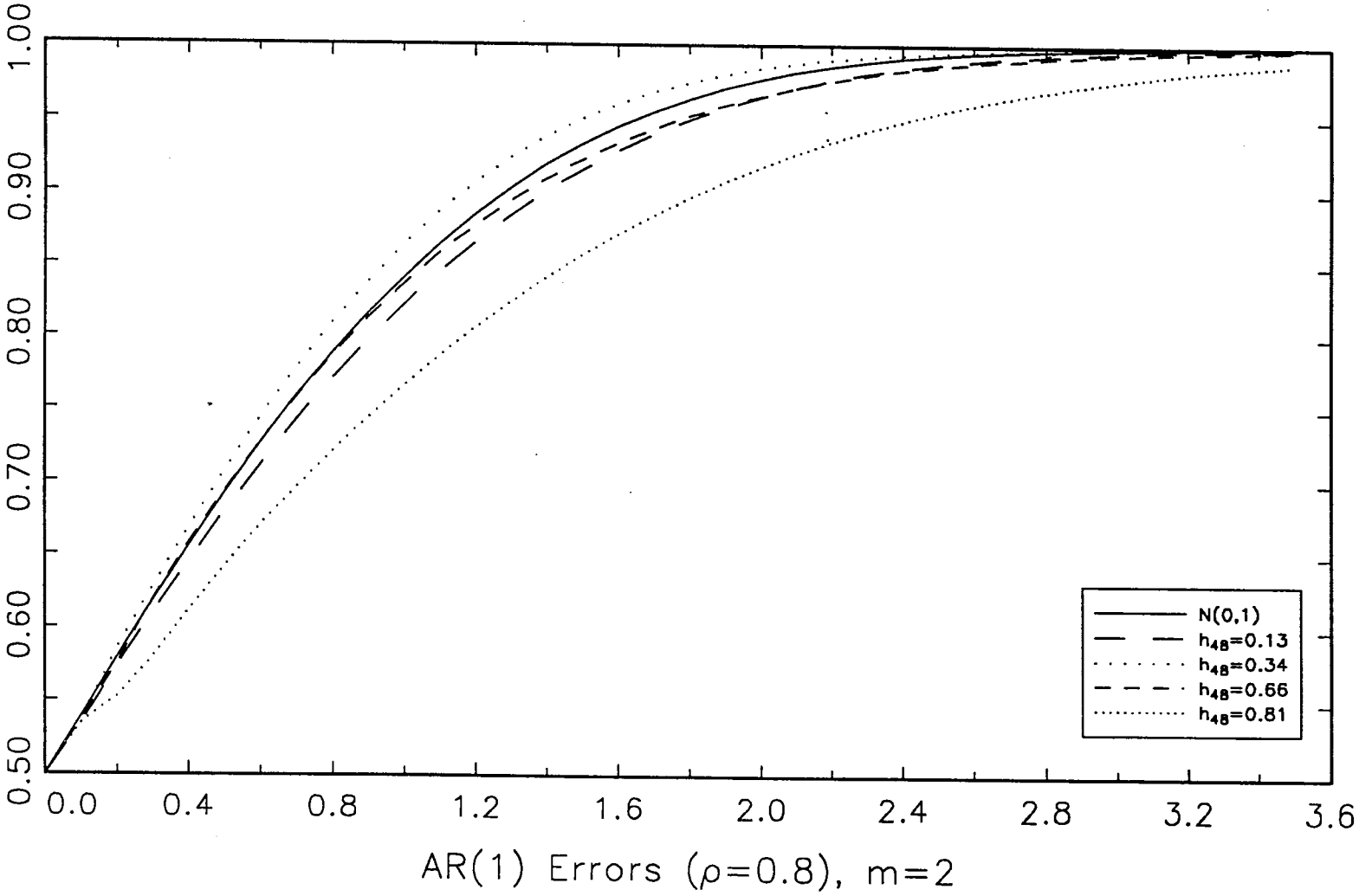


Figure 19a: Scatter Plot: X_2 versus X_3

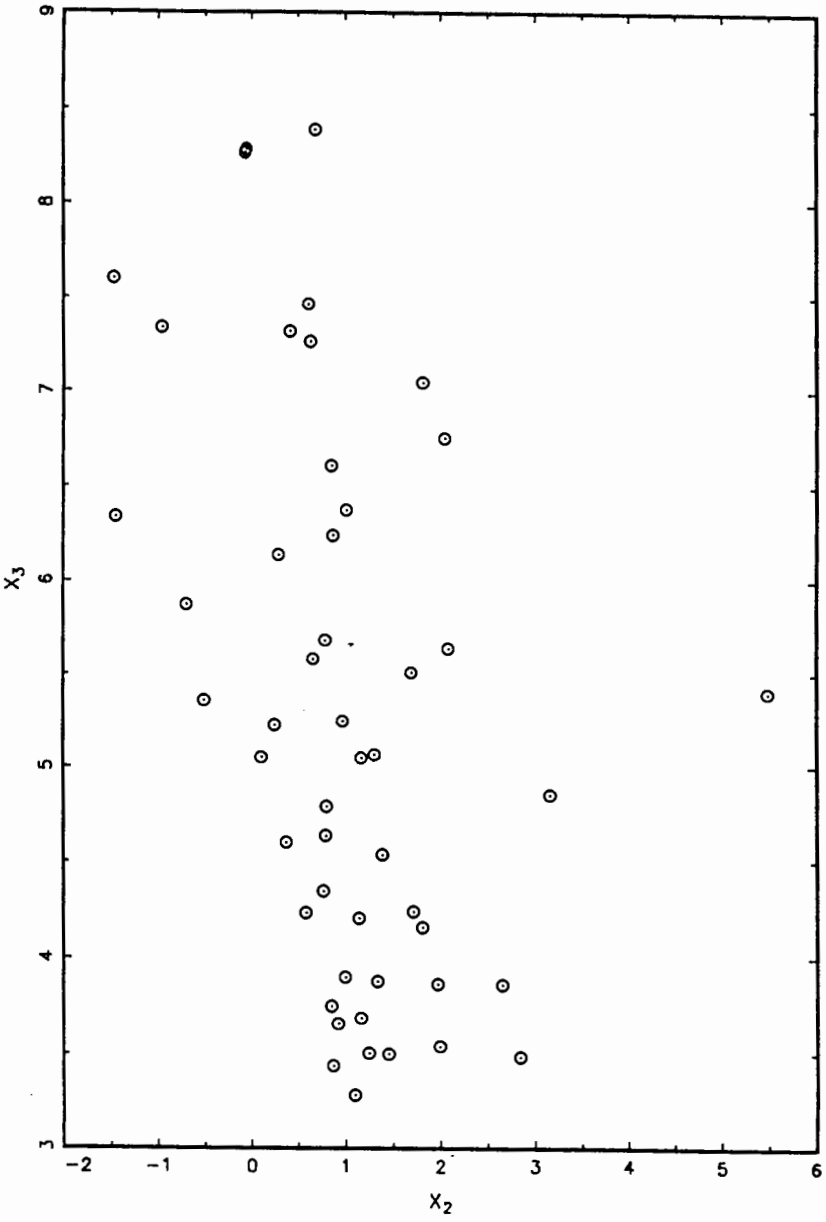


Figure 19b: X_2 versus h_{II}

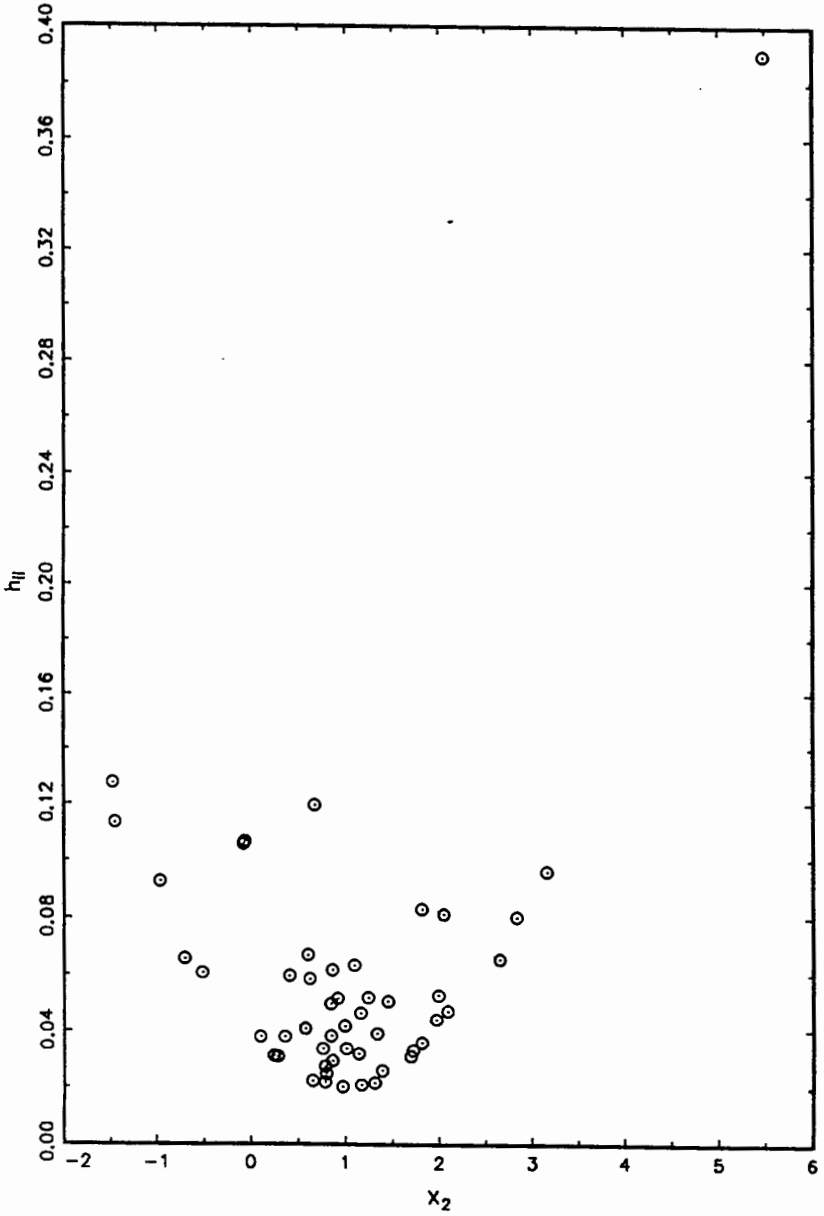


Table 3

Mackinnon and White Data

	X2	X3		X2	X3
1	1.095	3.281	26	0.411	7.318
2	1.451	3.499	27	0.628	7.263
3	1.991	3.538	28	2.048	6.752
4	2.839	3.481	29	1.010	6.375
5	1.242	3.504	30	-0.513	5.358
6	1.159	3.685	31	1.970	3.863
7	0.994	3.900	32	1.141	4.206
8	1.335	3.879	33	0.103	5.050
9	2.651	3.860	34	0.576	4.234
10	1.812	4.159	35	0.866	3.435
11	0.788	4.631	36	0.846	3.748
12	0.359	4.597	37	1.715	4.241
13	1.168	5.048	38	3.163	4.851
14	0.970	5.246	39	2.088	5.640
15	1.387	4.534	40	0.853	6.608
16	0.917	3.657	41	0.684	8.388
17	0.760	4.345	42	0.610	7.462
18	0.799	4.787	43	-1.476	7.600
19	1.306	5.065	44	-0.070	8.268
20	1.694	5.510	45	-0.058	8.286
21	0.242	5.226	46	-0.965	7.336
22	0.654	5.581	47	-0.740	5.873
23	0.282	6.138	48	5.486	5.401
24	0.867	6.240	49	-1.456	6.337
25	1.814	7.047	50	0.785	5.684

Source: Mackinnon and White (1985)

X2 - Rate of growth of real U.S. disposable income (1963-3 to 1974-4).

X3 - U.S. treasury bill rate (1963-3 to 1974-4).

6. CONCLUDING REMARKS

In this Chapter it is suggested that HAC estimators can be severely biased in finite samples, where the importance of this bias depends on the correlation structure of the errors and on the design of X. Downward bias can be severe if ρ is high and X is *quasi-well* balanced[23] or if ρ is small but X contains points with a high leverage. When ρ is high and the design suffers from the presence of leverage points, upward bias is possible. Some of these conclusions were drawn by using some examples, for the particular case of the Newey and West estimator.

The computation of HAC type estimators is related with the derivation of the optimal value for the lag truncation. One of the best proposals in the literature is due to Andrews (1991). However his method is not easily applicable and as shown in Section 2, suffers from some drawbacks. Anyway these conclusions should not be generalized without further investigation.

In the case of independent errors and considering a well balanced design, Newey and West estimator is downward biased. This bias becomes more important in the presence of leverage points, which tend to its minimum value as h_{ii} increases towards one. Additionally, it was also seen that this sensitivity to leverage depends on the direction considered, particularly when the variance of $w'\widehat{\beta}$ is computed in the direction of a covariate that suffers from a leverage in its direction.

In the case of non-independent errors the Newey and West estimator can be upward biased when there is a leverage point associated with a high value of ρ. However this relation is not monotone as one has seen in the particular case of Mackinnon and White data. Moreover the effects of leverage points in the case of MA(1) errors seem

[23] By this one means a design without leverage points.

to be less pronounced than in the case of AR(1) errors.

The modified Newey and West estimator performs better than the Newey and West estimator, in particular for moderate autocorrelations in the errors and if the data does not have leverage points. Moreover, this estimator is asymptotically equivalent to the Newey and West estimator.

A suggestion for future research is the study of the performance of HAC type estimators, considering different HCCM estimators. As seen in Section 2, HAC estimators can be given as a sum of White's HCCM estimator and an Autocorrelation Consistent Covariance Matrix estimator.

The use of the Newey and West estimator in the Wald tests can produce very misleading inferences. In a design without leverage points the exact distribution is not well approximated by the standard normal and the accuracy of this approximation becomes worse when the autocorrelation of the errors increase toward 1. As a consequence the null hypotheses is rejected very often.

One important conclusion derived from this study is that HAC estimators should be used with some precautions. In some cases, particularly for small value of the error autocorrelation, as shown in Figures 10-11, its bias can be even worse than of the traditional LS standard errors. In the particular case of the Newey and West estimator, a possible solution to this problem is to use the modified Newey and West estimator. On the other hand if the value of the error autocorrelation is moderate or high and if the data does not suffer from the presence of leverage points, the bias in the Newey and West estimator lead to very misleading inferences [one reminds Figures 14-15]. To avoid these sort of problems one suggests as a first step the computation of the autocorrelation of the residuals. Unless this value is small the Newey

and West estimator should not be used.

7. REFERENCES

Andrews, Donald W. K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica*, 59, 817-858.

Andrews, Donald W. K. and Monahan, C. (1992). An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator, *Econometrica*, 60, 953-966.

Belsley, A. D.; Kuh, E. and Welsch, R. (1980). *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*, John Wiley and Sons, New York.

Chatterjee, S. and Hadi, A. (1988). *Sensitivity Analysis in Linear Regression*, John Wiley & Sons, New York.

Chesher, A. and Jewitt, I. (1987). The Bias of a Heteroskedasticity Consistent Covariance Matrix Estimator, *Econometrica*, 55, 1217-1222.

Chesher, A. and Austin, G. (1991). The Finite-sample Distributions of Heteroskedasticity Robust Wald Statistics, *Journal of Econometrics*, 47, 153-173.

Cook, D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, New York.

Cumby, R.; Huizinga, J. and Obstfeld, M. (1983). Two-step Two-stage Least Squares Estimation in Models With Rational Expectations, *Journal of Econometrics*, 21, 333-355.

Davidson, R. and Mackinnon, J. (1993). *Estimation and Inference in Econometrics*, Oxford University Press, New York.

Eicker, F. (1963). Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions. *Annals of Mathematical Statistics*, 34, 447-456.

Fox, A. J. (1972). Outliers in Time Series, *Journal of the Royal Statistical Association*, Ser B, 34, 350-363.

Gallant, A. R. (1987). *Nonlinear Statistical Models*, John Wiley & Sons.

Hansen, B. E. (1992). Consistent Covariance Matrix Estimation for Dependent Heterogeneous Processes, *Econometrica*, 60, 967-972.

Hansen, L. P. (1982). Large Samples Properties of Generalized Method of Moment Estimators, *Econometrica*, 50, 1029-1054.

Hansen, L. P. and Singleton, K. (1982). Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models, *Econometrica*, 50, 1269-1286.

Huber, P. (1973). Robust Regression: Asymptotic, Conjectures and Monte Carlo, *Annals of Statistics*, 1, 799-821.

Huber, P. (1973). *Robust Regression*, John Wiley & Sons, New York.

Keener, R. and Kmenta, J. (1991). Estimation of the Covariance Matrix of the Least-Squares Regression Coefficients whe the Disturbance Covariance Matrix is of Unknown Form, *Econometric Theory*, 7, 22-45.

Mackinnon, J. and White, H. (1985). Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Samples Properties. *Journal of Econometrics*, 29, 305-325.

Magnus, J. R. (1978). The Moments of Products of Quadratic Forms in Normal Variables, *Statistica Neerlandica*, 32, 201-210.

Newey, W. and West, K. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica*, 55, 703-708.

Newey, W. and West, K. (1994). Automatic Lag Selection in Covariance Matrix Estimation, *Review of Economic Studies*, 61, 631-653.

Pollock, D. S. (1979). *The Algebra of Econometrics.* John Wiley & Sons.

Strang, G. (1980). *Linear Algebra and its Applications.* Academic Press, New York.

White, H. (1980). A Heteroskedasticity-consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica*, 48, 817-838.

White, H. (1984). *Asymptotic Theory for Econometricians.* Academic Press, Orlando.

White, H. and Domowitz, I. (1984). Nonlinear Regression with Dependent Observations, *Econometrica*, 52, 143-161.

Wooldridge, J. (1991). On the Application of robust, regression-based diagnostics to models of conditional means and conditional variances, *Journal of Econometrics*, 47, 5-46.

CHAPTER 3 - HETEROSKEDASTICITY AND AUTOCORRELATION CONSISTENT COVARIANCE MATRIX ESTIMATOR WITH IMPROVED FINITE-SAMPLE PROPERTIES: AN APPROACH BASED ON A GROUPED JACKKNIFE ESTIMATOR

1. INTRODUCTION

Heteroskedastic and autocorrelated errors are common problems in a regression framework. Neglecting this problem can severely affect inference due to incorrect computation of standard errors. This has led to a substantial literature concerning robust covariance matrix estimators. Examples include the works of White (1980), MacKinnon and White (1985) and Chesher and Jewitt (1987), in the heteroskedastic case and Hansen (1982), Newey and West (1987) and Andrews (1991), in the heteroskedastic and autocorrelated case. In this Chapter a new approach derived from the subseries values of Carlstein (1986) and the grouped jackknife of Quenouille (1956) is suggested.

In the iid case the Jackknife and Bootstrap have been widely used as techniques to estimate the variance of a general linear statistic. Both of these techniques are based on replicas of a general statistic computed from the data. The advantage of its application is in a bias reduction. However, they lead to incorrect results when dependence in the data is neglected. To avoid this problem one can use subsamples or blocks of observations, constructed in an appropriate manner, in order to preserve

this dependence.

In a recent paper, Carlstein (1986) suggests a new approach by using subseries (or blocks) of values to compute the variance of a general statistic from a stationary sequence. Let $\{Y_i,\ -\infty < i < +\infty\}$ be a strictly stationary α-mixing sequence[1]. For n observed values $y_1, y_2, \ldots, y_n$, Carlstein considered the computation of $\sigma^2 = Var\{T_n\}$, where $T_n = T_n(Y_1, Y_2, \ldots, Y_n)$ is a general statistic. To deal with it he suggested dividing the original sample into g non-overlapping blocks of length l such that each one preserves the original dependence of the data. Using this device he moves from the original problem of n dependent entities to $g < n$ *independent*[2] entities $Z_i = Y_{il+1}, Y_{il+2}, \ldots, Y_{il+l},\ i = 0, 1, \ldots, g-1$. The proposed estimator for σ^2 is [see Carlstein (1986), page 1173]

$$\widehat{\sigma}^2 = \frac{l}{g} \sum_{i=0}^{g-1} (s_l^{il} - \overline{s})^2 \ ,$$

where s_l^{il} is the value of the statistic $T_l^{il}(Z_i)$ computed in each block $i = 0, 1, \ldots, g-1$ and $\overline{s}$ is the mean. With these remarks in mind, Carlstein's estimator is nothing more than the usual sample variance amongst the standardized block values.

In the particular case of a sample mean, the subseries value technique of Carlstein is identical to the grouped jackknife [see Quenouille (1956) for a definition of grouped jackknife], i.e., deletion of blocks is the same as selecting blocks. Moreover for more

[1]The idea of α-mixing is associated with situations in which events are independent asymptotically. For a sequence $X_1, X_2, \ldots$ of random variables, let α_n be a number such that

$$| P(A \cap B) - P(A)P(B) | \leq \alpha_n$$

for $A \in \sigma(X_1, \ldots, X_k)$, $B \in \sigma(X_{k+n}, X_{k+n+1}, \ldots)$, and $k \geq 1$, $n \geq 1$ where $\sigma(\cdot)$ is a σ-field. If $\alpha_n \to 0$ as $n \to \infty$ the sequence $\{X_n\}$ is said to be α-mixing. See for example White (1984), page 44-45.

[2]To be more precise this should be viewed approximately in finite samples. However, due to the definition of α-mixing, they are independent asymptotically.

general statistics as the LS estimator, deletion is better than selection [see Künsch (1989), page 1218].

Extensions of Carlstein paper are given in the literature. Künsch (1989) allows for overlapping blocks and downweighting blocks of consecutive observations, instead of deletion. Politis and Romano (1994), using bootstrap methods, have proposed a variant where the length of each consecutive block is a random variable.

In what follows I extend the Carlstein method to a regression model. To fix ideas and for simplifying purposes consider the linear model,

$$y = X\beta + \varepsilon, \tag{1.1}$$

where y is an $n \times 1$ vector of observations, X is a $n \times k$ matrix of full column rank k, β is a $k \times 1$ vector of unknown parameters and ε is a $n \times 1$ vector of errors with $E(\varepsilon \mid X) = 0$ and $E(\varepsilon\varepsilon' \mid X) = \Omega$ positive definite of order $n \times n$. Additionally it is assumed that, conditional on X, the errors are second-order stationary, i.e., the disturbance covariances die out fast enough. The estimation of this model for each non-overlapping block can cause some problems due to a lack of degrees of freedom. For this reason the deletion approach seems to be more reasonable.

The remainder of the Chapter is organised as follows. In Section 2 the notation and the derivation of the covariance matrix estimator is presented. Section 3 deals with the first moment of the grouped jackknife estimator. This will be used in Section 4 where some results concerning the finite-sample performance are shown. A comparison with the Newey and West (1987) estimator is also made.

2. NOTATION AND ESTIMATOR

In this Section the non-overlapping deleted-l jackknife estimator of the covariance of $\widehat{\beta}$ is derived. Consider the set of indices,

$$I_i = \{il_n + 1, il_n + 2, \ldots, il_n + l_n\}, \qquad 0 \leq i \leq g_n - 1 \tag{2.1}$$

where l_n is the number of deleted observations from the $n \times (k+1)$ matrix $[y \; X]$ with $\#(I_i) = l_n$ and $g_n = n/l_n$ is the number of non-overlapping subsamples from the original sample of n elements. Let $M = I - X(X'X)^{-1}X'$, where I is the identity matrix of order n. In this notation the LS estimator after dropping the observations indexed by the set I_i will be,

$$\widehat{\beta}_{(I_i)} = \widehat{\beta} - (X'X)^{-1}X'_{I_i}M^{-1}_{I_iI_i}\widehat{\varepsilon}_{I_i}, \tag{2.2}$$

where[3] $M_{I_iI_i}$ is a square matrix of order l_n with the rows and columns of M indexed by the elements in the set I_i. The difference $\widehat{\beta}_{(I_i)} - \widehat{\beta}$ if properly scaled, can be viewed as measure of the joint influence of the observations index by I_i on $\widehat{\beta}$ [see Cook and Weisberg (1982), page 136]. Let e_{I_i} be a selection matrix of order $l_n \times n$ such that,

$$e_{I_i}X = X_{I_i}\ , \quad e_{I_i}\widehat{\varepsilon} = \widehat{\varepsilon}_{I_i}\ , \quad e_{I_i}Me'_{I_i} = M_{I_iI_i}\ , \quad e_{I_i}M = M_{I_i}\ . \tag{2.3}$$

Then (2.2) can be rewritten as,

$$\widehat{\beta}_{(I_i)} = \widehat{\beta} - (X'X)^{-1}X'e'_{I_i}M^{-1}_{I_iI_i}e_{I_i}\widehat{\varepsilon}, \tag{2.4}$$

and the grouped jackknife estimator of β as,

[3] If $M_{I_iI_i}$ is singular consider $M_{I_iI_i} = 0$ as suggested by Cook and Weiseberg (1982).

$$\widetilde{\beta} = \frac{1}{g_n} \sum_{i=0}^{g_n-1} \widehat{\beta}_{(I_i)}. \tag{2.5}$$

After some algebraical procedures [see appendix A1] it is straightforward to prove that the grouped jackknife estimator of the covariance matrix of $\widetilde{\beta}$ is given by the following expression,

$$\widetilde{\Sigma}_n = \frac{g_n - 1}{g_n}(X'X)^{-1}X'G_n\widehat{\Omega}_n G_n X(X'X)^{-1}, \tag{2.6}$$

where G_n and $\widehat{\Omega}_n$ are square matrices of order n defined as follows,

$$G_n = \begin{bmatrix} M_{I_1 I_1}^{-1} & O & \cdots & O \\ O & M_{I_2 I_2}^{-1} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & M_{I_{g_n-1} I_{g_n-1}}^{-1} \end{bmatrix}, \tag{2.7}$$

$$\widehat{\Omega}_n = D(\widehat{\varepsilon}_{I_i}\widehat{\varepsilon}'_{I_i}) - \frac{1}{g_n}\widehat{\varepsilon}\widehat{\varepsilon}' \tag{2.8}$$

and

$$D(\widehat{\varepsilon}_{I_i}\widehat{\varepsilon}'_{I_i}) = \sum_{i=0}^{g_n-1} e'_{I_i}\widehat{\varepsilon}_{I_i}\widehat{\varepsilon}'_{I_i}e_{I_i} = \begin{bmatrix} \widehat{\varepsilon}_{I_1}\widehat{\varepsilon}'_{I_1} & O & \cdots & O \\ O & \widehat{\varepsilon}_{I_2}\widehat{\varepsilon}'_{I_2} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & \widehat{\varepsilon}_{I_{g_n-1}}\widehat{\varepsilon}'_{I_{g_n-1}} \end{bmatrix} \tag{2.9}$$

with O an square null matrix of order l_n. These matrices are block diagonal where the dimension of each block equals l_n. As can be seen from expressions (2.7) to (2.9),

the estimator given in (2.6) is a function of blocks of the residuals and therefore Carlstein's results will apply. In particular (2.6) can be viewed as a mean of covariances computed over each block of length l_n. The consistency of this estimator can be proved under the following assumptions.

Assumption 1. Let h_{ii}, $i = 1, 2, \ldots, n$ be the diagonal elements of the hat matrix $H = X(X'X)^{-1}X'$. It is assumed that $\max(h_{ii}) \to 0$, as $n \to \infty$.

Assumption 2. The number of elements in each block is a function of the sample size such that $l_n = o(n)$.

Theorem 1 *(L_2 consistency). Define the covariance matrix of $\widehat{\beta}$ as*

$$n\Sigma_n = \left(\frac{X'X}{n}\right)^{-1} \frac{X'E(\varepsilon\varepsilon')X}{n} \left(\frac{X'X}{n}\right)^{-1}$$

where $\lim_{n\to\infty}[X'E(\varepsilon\varepsilon')X]/n = \Phi$. Given assumptions 1 and 2, if the error component of the model (1.1) is an α-mixing sequence and if $[X'_{I_i}M^{-1}_{I_iI_i}\widehat{\varepsilon}_{I_i}]^4$ is uniformly integrable then $n\widetilde{\Sigma}_n \underset{L2}{\longrightarrow} \Sigma$ as $n \to \infty$, where $\widetilde{\Sigma}_n$ is the estimator of Σ_n defined in (2.6).

Proof. : see Appendix A2. □

Note that as $n \to \infty$ one achieves an infinite number of subsamples with an infinite number of elements matching the original autocorrelation structure of the population. Moreover, due to the α-mixing assumption these subsamples are asymptotically independent.

An interesting result concerning this estimator is that it generalizes the deleted-1 jackknife estimator of variance, discussed in the paper of MacKinnon and White (1985). For the independent case, $l_n = 1$, one has

$$
\begin{aligned}
M_{I_i I_i}^{-1} &= (1 - h_{ii})^{-1} \\
\widehat{\varepsilon}_{I_i} &= \widehat{\varepsilon}_i \\
G_n &= diag\{(1 - h_{ii})^{-1}\} \\
D(\widehat{\varepsilon}_{I_i}\widehat{\varepsilon}'_{I_i}) &= diag\{\widehat{\varepsilon}_i^2\}
\end{aligned}
$$

3. FIRST MOMENT OF THE GROUPED JACKKNIFE ESTIMATOR OF THE VARIANCE

In this Section one derives the first moment of (2.6). The goal is to present in the next Section a study concerning the finite-sample performance of the estimator being presented.

Let ω be any chosen combination of the parameters in $\widehat{\beta}$. Therefore, $\widetilde{Var(\omega'\widehat{\beta})} = \widetilde{\Sigma}_\omega$ is a scalar and thus can be expressed in vectorial form as,

$$\widetilde{\Sigma}_\omega = (z'_\omega \otimes z_\omega) vec(G_n \widehat{\Omega}_n G_n)$$

where $z'_\omega = \omega'(X'X)^{-1}X'$ is a $k \times 1$ vector, $\otimes$ is the kroneker product and $vec(G_n\widehat{\Omega}_n G_n)$ is a $n^2 \times 1$ vector formed by stacking the columns of $G_n\widehat{\Omega}_n G_n$.

The first moment of the grouped jackknife estimator of the variance is given by

$$E(\widetilde{\Sigma}_\omega) = (z'_\omega \otimes z_\omega) vec[\Xi] \tag{3.1}$$

where Ξ is a $n \times n$ matrix defined as $\Xi = G_n E(\hat{\Omega}_n) G_n$ with

$$E(\tilde{\Omega}_n) = \begin{bmatrix} M_{I_1}\Omega M'_{I_1} & O & \cdots & O \\ O & M_{I_2}\Omega M'_{I_2} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & M_{I_{g_n-1}}\Omega M'_{I_{g_n-1}} \end{bmatrix} - \frac{1}{g_n} M\Omega M, \qquad (3.2)$$

4. APPLICATION

This Section deals with the finite-sample performance of the variance estimator $\widetilde{\Sigma}_w$, for some linear combination w of $\widehat{\beta}$. With purposes of notational simplification let w be a vector that select elements of $\widehat{\beta}$, $\widehat{\beta}_j$, $j = 1, 2, \ldots, k$, such that $\widetilde{\Sigma}_w = \widetilde{Var}(\widehat{\beta}_j)$. As a measure of its performance one will consider the proportionate bias

$$pb(\widetilde{\Sigma}_w) = \frac{E(\widetilde{\Sigma}_w) - \Sigma_w}{\Sigma_w} .$$

For a relative measure of its performance, $pb(\widetilde{\Sigma}_w)$ is confronted with the proportionate bias of the Newey and West (1987) estimator[4] because this has been suggested as the robust estimator to be used in the non-iid case. This estimator can be expressed by

$$\hat{\Sigma}_w = w'(X'X)^{-1}X'BX(X'X)^{-1}w$$

where $B = [b_{ij}]$ is a band matrix with generic element

[4]See Passos (1994) for a study concerning the finite-sample properties of the Newey-West estimator.

$$b_{ij} = \begin{cases} \widehat{\varepsilon}_i^2 & , \ j = i \\ \kappa_{ij}\widehat{\varepsilon}_i\widehat{\varepsilon}_j & , \ j \neq i, \ \mid j-i \mid \leq m \\ 0 & , \ \mid j-i \mid > m \end{cases}$$

where

$$\kappa_{ij} = 1 - \frac{\mid j-i \mid}{m+1}$$

is the Bartlett kernel and $m = l - 1$ is a bandwidth. In this notation, when for example $m = 0$ $(l = 1)$, $\widehat{\Sigma}_w$ is the White's (1980) estimator and $\widetilde{\Sigma}_w$ is the deleted-1 jackknife.

In what follows one considers two different situations: model without covariates, $y_i = \mu + \varepsilon_i$, where $\mu = X'\beta$ is a constant and $i = 1, 2, \ldots, 50$ (Case 1) and models with two covariates, $y_i = \beta_o + x_{i1}\beta_1 + x_{i2}\beta_2 + \varepsilon_i$ with designs specified as follows: Case 2 - $\beta_o = 0$ and $(x_{i1}\ x_{i2})$, $i = 1, 2, \ldots, 100$ define a well balanced design in a circle [see Chesher and Austin (1991)] ; Case 3 - $x_{1i} \sim N(0, 1)$ and $x_{2i} \sim \chi^2(5)$, $i = 1, 2, \ldots, 100$. For all cases the error component ε is an $AR(1)$ process,

$$\varepsilon_i = \rho\varepsilon_{i-1} + u_i$$

where $u_i \overset{iid}{\sim} N(0, 1)$, $i = 1, 2, \ldots, n$ and $\rho \in [0,\ 0.8]$.

Case 1: In this case $\widehat{\mu} = (1/n)\sum_{i=1}^{n} y_i$ is the sample mean and $\widetilde{\Sigma} = \widetilde{\Sigma}_w$ is a scalar. For each $\rho \in [0,\ 0.8]$ Tables 1.1 and 1.2 (see Appendix) and Figures 1.1 and 1.2 below show the values of the proportionate bias computed at $l_n \in [1,\ 25]$, associated with

$\widetilde{\Sigma}$ and $\widehat{\Sigma}$, respectively. The first conclusion to draw from these results is that the grouped jackknife performs significantly better than the Newey and West estimator whatever the value of l_n is. However for high values of ρ, $pb(\widetilde{\Sigma})$ seems to attain its minimum for values of l_n higher than the correspondent minimum of $pb(\widehat{\Sigma})$. The reason is due probably to a reduction of the number of blocks as l_n increases (note that $n = l_n g_n$) which can result in less accurate estimates. The second conclusion, is that the bias of $\widetilde{\Sigma}$ is almost zero for a good choice of l_n. In third place, $\widetilde{\Sigma}$ appears to stabilize around its first moment after some value of l_n is reached. This is a very desirable property meaning that one minimizes the loss of an incorrect value of l_n by increasing its value. For example, if $\rho = 0$ any value of l_n in the range $(1,\ 25)$ proportionate an estimate with an insignificant bias. However less bias is usually associated with a bigger variance and we should be careful without further research in this field. Finally, the proportionate bias appears to be slightly inferior when n/l_n is an integer then in a neighbourhood of it. This can be due to the presence of one block (the last one) with less elements then the others and therefore it can be improved through some appropriate correction factor.

Figure 1.1: Proportionate Bias – Grouped Jackknife
Without Covariates and AR(1) Errors

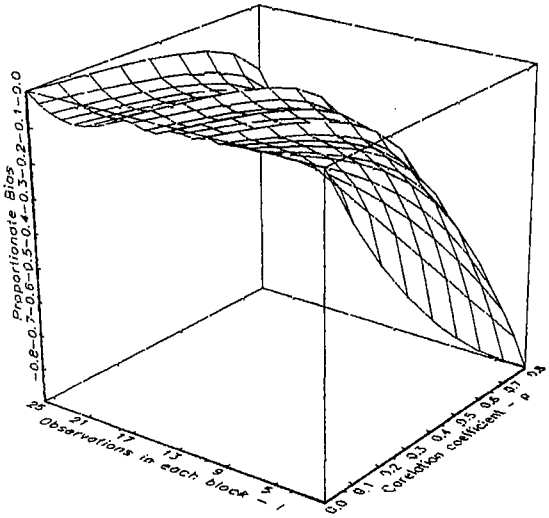
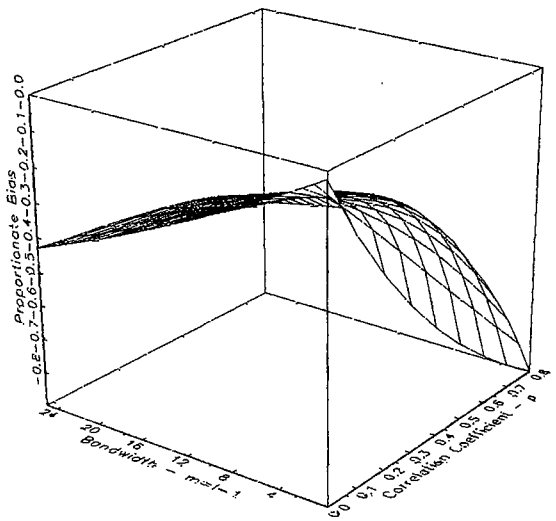


Figure 1.2: Proportionate Bias – Newey and West estimator
Without Covariates and AR(1) Errors



Case 2: Due to the symmetry inherent to this design only the results associate with the first covariate are shown. The introduction of covariates does not change the conclusions reached above as can be seen from figures 2.1 and 2.2 [see also Tables 2.1 and 3.1 in Appendix]. The jackknife estimator appears to be less biased than the Newey and West estimator. When ρ varies from 0 to 0.8 the proportionate bias is less then 0.1, in absolute value, for a wide strip of l values. On the other hand and once more, the wave effect due to the non-integer division, n/l_n, is evident, particularly for higher values of l.

Figure 2.1: Proportionate Bias – Grouped Jackknife (β_1)
Two Covariates and AR(1) Errors

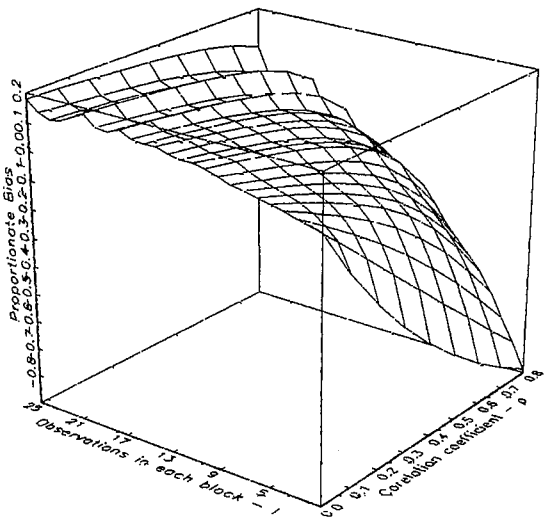
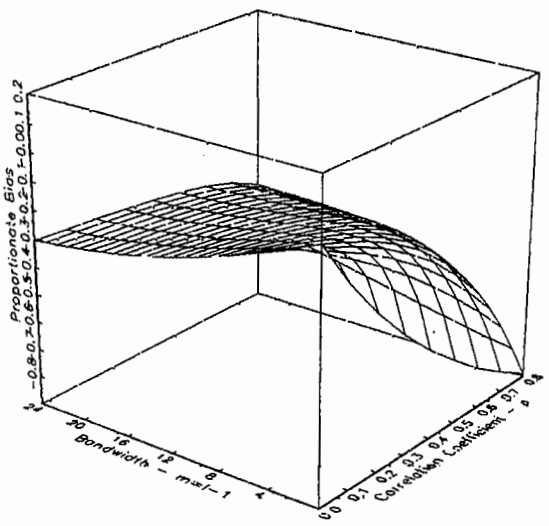


Figure 2.2: Proportionate Bias of Newey and West estimator (β_1)
Two covariates and AR(1) Errors



Case 3: In the two cases presented the designs have some good properties in order to access the performance of the grouped-jackknife estimator of the variance, viewed only as a function of l_n. The idea was to isolate the effect of the design in the performance of the estimator. In these cases the estimator seems to perform relatively well and better than the Newey and West estimator, so far as first moments are concerned. However, Passos (1994) showed that the performance of the Newey and West estimator is sensitive to the design of the covariates, particularly when this design contains leverage points [see also Chesher and Jewitt (1987) in the case of heteroskedasticity consistent covariance matrix estimators]. In this case a more realistic design that allows for the presence of leverage points is considered.

Once more, from the results presented, the grouped jackknife estimator of the variance appears to perform better than the Newey and West estimator. For any value of ρ and for an appropriate value of l_n the proportionate bias is less than 0.1 in absolute value [see Figures 3.1 to 3.4 above and tables 3.1 to 3.4 in appendix]. Considering for example $\rho = 0.3$, the proportionate bias computed in the direction

of $\widehat{\beta}_3$, is less than 0.1, in absolute value, for all $l_n \neq 25$ in the range $\{1, \ldots, 25\}$. For $\rho = 0.6$ one has the same conclusion for all $l_n \neq \{11, 12, 14, 18, 19, 24, 25\}$.

The wave effect is in this case much more pronounced. Another source of explanation, that is more likely in a non-balanced design, is due to the presence of joint influential observations [see Cook and Weisberg (1982), chapter 3]. Note that $\widehat{\beta}_{(I_i)}$ enters in the definition of the covariance matrix estimator of $\widehat{\beta}$. Moreover, as pointed out in Section 2, $\widehat{\beta}_{(I_i)} - \widehat{\beta}$ is a measure of the joint influence of the observations index by I_i. Thus, large values of $\widehat{\beta}_{(I_i)}$, with respect to some metric, can explain this wave effect. To avoid this problem two solutions are possible: a) given l_n, the g values of $\widehat{\beta}_{(I_i)}$, for $i = 0, 1, \ldots, g$, can be trimmed in some appropriated metric [see Cook and Weisberg (1982)], redefining (2.6) as a trimmed estimator; b) since this wave effect depends on the design, they can be easily identified as peaks in a plot of the variance estimate against l_n. Thus, values of l_n with associated peaks should not be considered as candidates. If for example we have some *a priori* information about the structure of the errors that lead to some value of l_n and if this value correspond to a peak simply consider another value in this neighbourhood.

Figure 3.1: Proportionate Bias – Grouped Jackknife (β_1)
Non Balanced Design – Two Covariates and AR(1) Errors

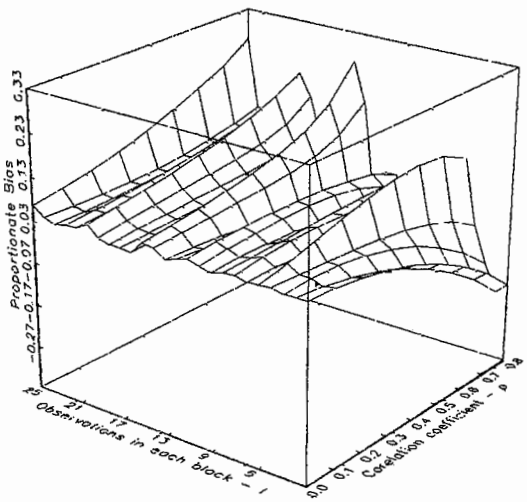


Figure 3.2: Proportionate Bias – Grouped Jackknife (β_2)
Non Balanced Design – Two Covariates and AR(1) Errors

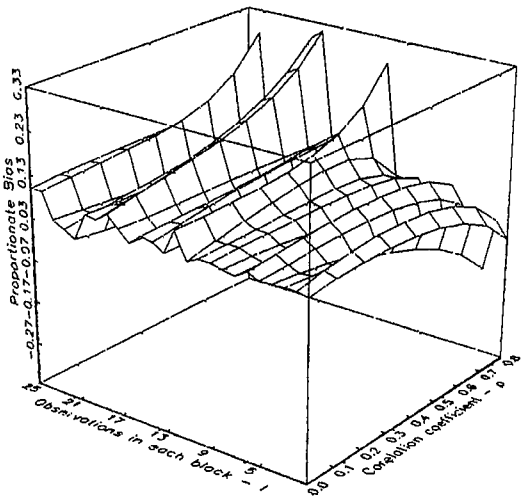


Figure 3.3: Proportionate Bias of Newey and West Estimator (β_1)
Non Balanced Design – Two Covariates and AR(1) Errors

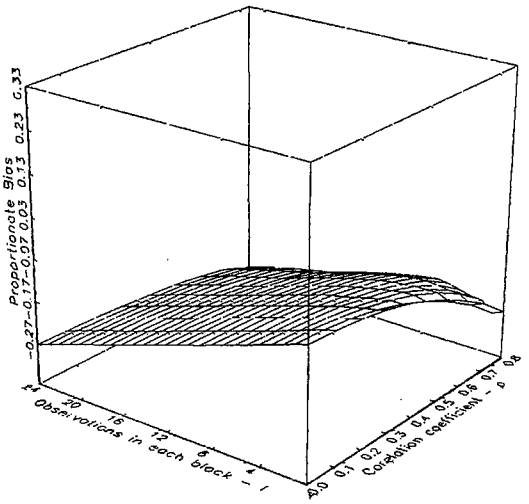
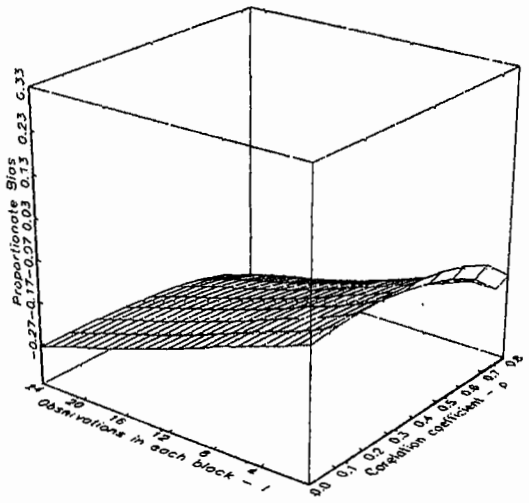


Figure 3.4: Proportionate Bias of Newey and West Estimator (β_7)
Non Balanced Design – Two Covariates and AR(1) Errors



5. FURTHER COMMENTS

In a linear regression framework, the deleted-l jackknife with non-overlapping blocks (or the grouped jackknife) is used in this Chapter as a technique to compute the covariance matrix of $\widehat{\beta}$, robust to departures of the iid errors if the covariances die fast enough as the sample size increases.

It is also pointed out two interesting features concerning this estimator: it can be viewed as a generalization of the deleted-1 jackknife ($l_n = 1$) to the non-independent case and the finite-sample study of Section 4 reveals a desirable stability of the estimator around its first moment, for a wide range of l_n values. This last property is relatively important having in mind the nonexistence of straightforward procedure to determine the value of l_n.

A possible direction for further research is an extension of this procedure to over-

lapping blocks in the way considered by Künsch (1989).

6. APPENDIX

[A1] Derivation of the covariance matrix estimator of $\widehat{\beta}$:

The derivation of the estimator presented in (2.6) is straightforward, requiring only some simple but tedious algebra. To achieve expression (2.2) one will make use of the following Lemma [see Cook and Weisberg (1982), page 210]:

Lemma: If A and D are nonsingular matrices of order k and m, respectively, B of order $k \times m$ and C of order $k \times m$, the inverse of the sum $(A + BDC')$ is given as follows:

$$(A + BDC')^{-1} = A^{-1} - A^{-1}B(D^{-1} + C'A^{-1}B)^{-1}C'A^{-1}$$

From this Lemma and the fact that $\widehat{\beta}_{(I_i)}$ can be expressed as

$$\begin{aligned} \widehat{\beta}_{(I_i)} &= (X'_{(I_i)}X_{(I_i)})^{-1}X'_{(I_i)}y_{(I_i)} \\ &= (X'X - X'_{I_i}X_{I_i})^{-1}(X'y - X'_{I_i}y_{I_i}) \end{aligned}$$

result (2.2) follows immediately [see also Cook and Weiseberg (1982), page 136]. Moreover, one can rewrite expression (2.5) as

$$\widetilde{\beta} = \widehat{\beta} - \frac{1}{g_n}(X'X)^{-1}X' \sum_{i=0}^{g_n-1} \left(e'_{I_i}M_{I_iI_i}e_{I_i}\right)\widehat{\varepsilon}$$

where $\sum_{i=0}^{g_n-1} \left(e'_{I_i}M_{I_iI_i}e_{I_i}\right) = G_n$ and e_{I_i} and G_n are defined as in (2.3) and (2.7), respectively. Having in mind these results one has

$$\widetilde{\Sigma} = \frac{g_n-1}{g_n}\sum_{i=0}^{g_n-1}(\widehat{\beta}_{(I_i)}-\widetilde{\beta})(\widehat{\beta}_{(I_i)}-\widetilde{\beta})'$$

$$= \frac{g_n-1}{g_n}(X'X)^{-1}X'\left[\sum_{i=0}^{g_n-1} e'_{I_i}M^{-1}_{I_iI_i}e_{I_i}\widehat{\varepsilon}\widehat{\varepsilon}'e'_{I_i}M^{-1}_{I_iI_i}e'_{I_i} - (1/g_n)G_n\widehat{\varepsilon}\widehat{\varepsilon}'G_n\right]X(X'X)^{-1}$$

where the first term inside brackets is a sum of block diagonal matrices of the form

$$\begin{bmatrix} O & \cdots & O & \cdots & O \\ \vdots & & \vdots & & \vdots \\ O & \cdots & M^{-1}_{l_il_i}\widehat{\varepsilon}_{I_i}\widehat{\varepsilon}'_{l_i}M^{-1}_{l_il_i} & \cdots & O \\ \vdots & & \vdots & & \vdots \\ O & \cdots & O & \cdots & O \end{bmatrix}$$

that simplifies to $G_nD(\widehat{\varepsilon}_{I_i}\widehat{\varepsilon}'_{I_i})G_n$ with $D(\widehat{\varepsilon}_{I_i}\widehat{\varepsilon}'_{I_i})$ given by (2.9).

[A2] Proof of Theorem 1: Rewriting expression (2.6) as,

$$\widetilde{\Sigma}_n = \frac{g_n-1}{g_n}\frac{1}{n}\left(\frac{X'X}{n}\right)^{-1}\frac{X'G_n\widehat{\Omega}_nG_nX}{n}\left(\frac{X'X}{n}\right)^{-1}$$

and assuming that $\lim_{n\to\infty}(X'X/n) = Q$ is a finite positive definite matrix, it is well known that the consistency of $n\widetilde{\Sigma}_n$ depends on the asymptotic behaviour of the matrix $\widetilde{\Phi}_n = (X'G_n\widehat{\Omega}_nG_nX)/n$. For simplifying purposes and without loss of generality consider the one regressor case with $X = x$. From (2.8) one has

$$\widetilde{\Phi}_n = \frac{1}{n}x'_nG_nD(\widehat{\varepsilon}_{I_i}\widehat{\varepsilon}'_{I_i})G_nx_n - \frac{1}{g_n}\frac{1}{n}x'_nG_n\widehat{\varepsilon}\widehat{\varepsilon}'G_nx_n$$

or, in terms of summation

$$\begin{aligned}
\widetilde{\Phi}_n &= \frac{1}{n}\sum_{i=0}^{g_n-1} x'_{I_i} M^{-1}_{I_iI_i}\widehat{\varepsilon}_{I_i}\widehat{\varepsilon}'_{I_i} M^{-1}_{I_iI_i} x_{I_i} - \frac{1}{n}\frac{1}{g_n}\sum_{i=0}^{g_n-1} x'_{I_i} M^{-1}_{I_iI_i}\widehat{\varepsilon}_{I_i}\widehat{\varepsilon}'_{I_i} M^{-1}_{I_iI_i} x_{I_i} - \\
&\quad -\frac{2}{n}\frac{1}{g_n}\sum_{j=0}^{g_n-2}\sum_{i=j+1}^{g_n-1} x'_{I_i} M^{-1}_{I_iI_i}\widehat{\varepsilon}_{I_i}\widehat{\varepsilon}'_{I_{i-j}} M^{-1}_{I_{i-j}I_{i-j}} x_{I_{i-j}} \\
&= \frac{1}{n}\sum_{i=0}^{g_n-1}\left(x'_{I_i} M^{-1}_{I_iI_i}\widehat{\varepsilon}_{I_i} - \frac{1}{g_n}\sum_{j=0}^{g_n-1} x'_{I_i} M^{-1}_{I_iI_i}\widehat{\varepsilon}_{I_i}\right)^2 \\
&= \frac{l_n}{g_n}\sum_{i=0}^{g_n-1}\left(s_{I_i} - \overline{s}\right)^2
\end{aligned}$$

where $s_{I_i} = (1/l_n)x'_{I_i} M^{-1}_{I_iI_i}\widehat{\varepsilon}_{I_i}$ and $\overline{s}$ is the sample mean of the s_{I_i}'s. This expression is similar to Carlstein estimator of the variance and therefore the L_2-consistency of $\widetilde{\Phi}_n$ follows from Theorem 2 of Carlstein [see Carstein (1986), pages 1174-75].

TABLE 1.1: Case 1 - Proportionate Bias - Grouped Jackknife Estimator

l_n \ ρ	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
1	0.000	-0.182	-0.334	-0.464	-0.575	-0.671	-0.755	-0.830	-0.896
2	-0.000	-0.100	-0.201	-0.303	-0.405	-0.507	-0.609	-0.711	-0.813
3	-0.001	-0.070	-0.142	-0.221	-0.307	-0.401	-0.505	-0.620	-0.744
4	-0.002	-0.054	-0.110	-0.173	-0.244	-0.328	-0.428	-0.545	-0.684
5	-0.000	-0.041	-0.084	-0.133	-0.192	-0.264	-0.356	-0.472	-0.620
6	-0.003	-0.040	-0.079	-0.122	-0.174	-0.239	-0.323	-0.434	-0.584
7	-0.002	-0.035	-0.068	-0.105	-0.150	-0.206	-0.282	-0.387	-0.537
8	-0.005	-0.034	-0.064	-0.098	-0.138	-0.189	-0.259	-0.358	-0.506
9	-0.009	-0.033	-0.059	-0.089	-0.125	-0.172	-0.237	-0.333	-0.479
10	0.000	-0.020	-0.042	-0.067	-0.097	-0.137	-0.193	-0.280	-0.422
11	-0.014	-0.034	-0.056	-0.081	-0.111	-0.151	-0.206	-0.291	-0.430
12	-0.009	-0.029	-0.051	-0.075	-0.103	-0.139	-0.189	-0.266	-0.397
13	-0.011	-0.028	-0.045	-0.065	-0.089	-0.121	-0.167	-0.240	-0.369
14	-0.024	-0.040	-0.058	-0.077	-0.102	-0.134	-0.179	-0.251	-0.376
15	-0.024	-0.040	-0.058	-0.078	-0.102	-0.134	-0.178	-0.246	-0.365
16	-0.013	-0.030	-0.047	-0.066	-0.089	-0.117	-0.156	-0.217	-0.328
17	-0.010	-0.022	-0.035	-0.049	-0.068	-0.091	-0.126	-0.183	-0.291
18	-0.033	-0.045	-0.058	-0.072	-0.090	-0.114	-0.149	-0.205	-0.312
19	-0.048	-0.060	-0.073	-0.088	-0.106	-0.130	-0.164	-0.220	-0.325
20	-0.057	-0.069	-0.082	-0.097	-0.115	-0.139	-0.174	-0.229	-0.332
21	-0.059	-0.071	-0.084	-0.100	-0.118	-0.142	-0.177	-0.231	-0.330
22	-0.055	-0.067	-0.081	-0.096	-0.114	-0.139	-0.172	-0.225	-0.319
23	-0.044	-0.056	-0.070	-0.086	-0.104	-0.127	-0.159	-0.208	-0.295
24	-0.026	-0.039	-0.052	-0.067	-0.084	-0.104	-0.132	-0.175	-0.256
25	-0.000	-0.008	-0.017	-0.027	-0.039	-0.055	-0.078	-0.116	-0.194

TABLE 1.2: Case 1 - Proportionate Bias - Newey and West Estimator

ρ / m_n	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
0	-0.020	-0.198	-0.348	-0.474	-0.583	-0.678	-0.760	-0.833	-0.898
1	-0.040	-0.138	-0.236	-0.334	-0.432	-0.530	-0.627	-0.725	-0.822
2	-0.059	-0.125	-0.194	-0.268	-0.349	-0.439	-0.536	-0.644	-0.760
3	-0.078	-0.128	-0.180	-0.238	-0.305	-0.382	-0.474	-0.582	-0.710
4	-0.097	-0.137	-0.179	-0.226	-0.281	-0.348	-0.432	-0.537	-0.669
5	-0.115	-0.149	-0.184	-0.224	-0.270	-0.329	-0.404	-0.503	-0.636
6	-0.134	-0.162	-0.193	-0.227	-0.267	-0.318	-0.386	-0.479	-0.609
7	-0.152	-0.177	-0.204	-0.233	-0.269	-0.315	-0.376	-0.462	-0.589
8	-0.169	-0.192	-0.216	-0.242	-0.274	-0.315	-0.371	-0.451	-0.573
9	-0.187	-0.207	-0.229	-0.253	-0.282	-0.319	-0.369	-0.444	-0.561
10	-0.204	-0.223	-0.242	-0.265	-0.291	-0.325	-0.371	-0.441	-0.552
11	-0.221	-0.238	-0.256	-0.277	-0.301	-0.332	-0.375	-0.440	-0.546
12	-0.238	-0.254	-0.271	-0.289	-0.312	-0.341	-0.381	-0.442	-0.543
13	-0.254	-0.269	-0.285	-0.303	-0.324	-0.351	-0.388	-0.445	-0.541
14	-0.270	-0.284	-0.299	-0.316	-0.336	-0.361	-0.396	-0.450	-0.542
15	-0.286	-0.299	-0.313	-0.329	-0.348	-0.372	-0.405	-0.456	-0.543
16	-0.302	-0.314	-0.328	-0.342	-0.360	-0.383	-0.414	-0.463	-0.546
17	-0.317	-0.329	-0.342	-0.356	-0.373	-0.394	-0.424	-0.470	-0.550
18	-0.332	-0.343	-0.356	-0.369	-0.385	-0.406	-0.435	-0.478	-0.555
19	-0.347	-0.358	-0.369	-0.382	-0.398	-0.418	-0.445	-0.487	-0.561
20	-0.361	-0.372	-0.383	-0.396	-0.410	-0.429	-0.456	-0.496	-0.567
21	-0.376	-0.386	-0.396	-0.409	-0.423	-0.441	-0.466	-0.505	-0.573
22	-0.390	-0.399	-0.410	-0.421	-0.435	-0.453	-0.477	-0.514	-0.580
23	-0.403	-0.413	-0.423	-0.434	-0.447	-0.464	-0.488	-0.524	-0.587
24	-0.417	-0.426	-0.436	-0.447	-0.459	-0.476	-0.499	-0.533	-0.595
25	-0.430	-0.439	-0.448	-0.459	-0.471	-0.487	-0.509	-0.543	-0.602

TABLE 2.1: Case 2 - Proportionate Bias - Grouped Jackknife Estimator ($\hat{\beta}_1$)

ρ / l_n	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
1	0.010	-0.173	-0.327	-0.457	-0.568	-0.665	-0.749	-0.822	-0.886
2	0.021	-0.081	-0.184	-0.287	-0.390	-0.492	-0.594	-0.695	-0.793
3	0.031	-0.041	-0.116	-0.196	-0.283	-0.379	-0.483	-0.595	-0.713
4	0.043	-0.010	-0.066	-0.128	-0.200	-0.285	-0.385	-0.501	-0.634
5	0.054	0.011	-0.034	-0.085	-0.145	-0.219	-0.311	-0.426	-0.565
6	0.062	0.025	-0.015	-0.059	-0.113	-0.179	-0.264	-0.375	-0.516
7	0.072	0.038	0.003	-0.037	-0.085	-0.144	-0.221	-0.325	-0.464
8	0.080	0.051	0.020	-0.015	-0.057	-0.111	-0.182	-0.281	-0.419
9	0.095	0.068	0.039	0.008	-0.029	-0.076	-0.139	-0.229	-0.361
10	0.109	0.088	0.065	0.039	0.007	-0.034	-0.091	-0.175	-0.302
11	0.115	0.092	0.068	0.042	0.011	-0.027	-0.080	-0.158	-0.279
12	0.115	0.094	0.072	0.047	0.016	-0.022	-0.075	-0.151	-0.269
13	0.123	0.106	0.087	0.065	0.039	0.006	-0.041	-0.113	-0.226
14	0.136	0.118	0.098	0.076	0.050	0.017	-0.026	-0.090	-0.193
15	0.134	0.119	0.103	0.084	0.062	0.033	-0.008	-0.071	-0.173
16	0.140	0.124	0.107	0.087	0.064	0.034	-0.007	-0.067	-0.161
17	0.159	0.146	0.133	0.118	0.100	0.077	0.043	-0.008	-0.094
18	0.140	0.127	0.114	0.098	0.079	0.054	0.020	-0.033	-0.120
19	0.146	0.132	0.118	0.101	0.082	0.056	0.022	-0.029	-0.109
20	0.182	0.172	0.162	0.151	0.137	0.119	0.093	0.054	-0.013
21	0.158	0.148	0.138	0.126	0.112	0.093	0.067	0.026	-0.042
22	0.138	0.128	0.117	0.105	0.090	0.071	0.044	0.002	-0.067
23	0.131	0.120	0.109	0.097	0.082	0.062	0.035	-0.006	-0.071
24	0.144	0.134	0.123	0.110	0.095	0.076	0.052	0.016	-0.037
25	0.179	0.173	0.167	0.160	0.151	0.139	0.123	0.099	0.060

TABLE 2.2: Case 2 - Proportionate Bias - Newey and West Estimator ($\widehat{\beta}_1$)

ρ	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
m_n									
0	-0.020	-0.198	-0.347	-0.473	-0.581	-0.675	-0.756	-0.828	-0.889
1	-0.040	-0.137	-0.235	-0.332	-0.429	-0.526	-0.622	-0.716	-0.807
2	-0.059	-0.124	-0.193	-0.267	-0.347	-0.434	-0.529	-0.632	-0.740
3	-0.077	-0.127	-0.179	-0.236	-0.302	-0.377	-0.466	-0.569	-0.684
4	-0.095	-0.135	-0.177	-0.223	-0.278	-0.343	-0.423	-0.521	-0.640
5	-0.113	-0.146	-0.181	-0.220	-0.266	-0.322	-0.394	-0.486	-0.603
6	-0.130	-0.158	-0.188	-0.222	-0.261	-0.311	-0.375	-0.461	-0.574
7	-0.146	-0.171	-0.197	-0.227	-0.261	-0.305	-0.363	-0.442	-0.550
8	-0.161	-0.183	-0.207	-0.233	-0.264	-0.303	-0.356	-0.429	-0.532
9	-0.176	-0.196	-0.217	-0.241	-0.269	-0.304	-0.352	-0.419	-0.517
10	-0.189	-0.208	-0.227	-0.249	-0.275	-0.307	-0.351	-0.413	-0.505
11	-0.202	-0.219	-0.237	-0.257	-0.281	-0.311	-0.351	-0.409	-0.496
12	-0.215	-0.230	-0.247	-0.265	-0.287	-0.315	-0.352	-0.407	-0.489
13	-0.226	-0.241	-0.256	-0.273	-0.294	-0.320	-0.355	-0.405	-0.483
14	-0.237	-0.250	-0.265	-0.281	-0.300	-0.325	-0.357	-0.405	-0.479
15	-0.247	-0.259	-0.273	-0.288	-0.306	-0.329	-0.360	-0.405	-0.475
16	-0.256	-0.268	-0.281	-0.295	-0.312	-0.334	-0.363	-0.406	-0.473
17	-0.264	-0.276	-0.288	-0.302	-0.318	-0.338	-0.366	-0.407	-0.470
18	-0.272	-0.283	-0.294	-0.307	-0.323	-0.342	-0.369	-0.408	-0.469
19	-0.279	-0.289	-0.300	-0.313	-0.328	-0.346	-0.372	-0.409	-0.467
20	-0.285	-0.295	-0.306	-0.318	-0.332	-0.350	-0.374	-0.410	-0.466
21	-0.291	-0.300	-0.311	-0.322	-0.336	-0.353	-0.377	-0.411	-0.465
22	-0.296	-0.305	-0.315	-0.326	-0.339	-0.356	-0.379	-0.412	-0.464
23	-0.301	-0.310	-0.319	-0.330	-0.343	-0.359	-0.381	-0.413	-0.463
24	-0.305	-0.314	-0.323	-0.334	-0.346	-0.361	-0.382	-0.414	-0.462

TABLE 3.1: Case 3 - Proportionate Bias - Grouped Jackknife Estimator ($\widehat{\beta}_1$)

l_n \ ρ	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
1	0.047	0.043	0.033	0.018	-0.004	-0.034	-0.072	-0.119	-0.180
2	0.046	0.038	0.025	0.006	-0.020	-0.053	-0.093	-0.143	-0.205
3	0.045	0.047	0.043	0.033	0.016	-0.009	-0.043	-0.087	-0.145
4	0.051	0.047	0.043	0.038	0.031	0.020	0.004	-0.020	-0.059
5	0.048	0.063	0.080	0.099	0.116	0.129	0.134	0.126	0.096
6	0.047	0.047	0.042	0.035	0.026	0.015	0.004	-0.008	-0.026
7	0.053	0.060	0.065	0.066	0.063	0.058	0.051	0.042	0.029
8	0.050	0.051	0.054	0.057	0.060	0.062	0.061	0.054	0.032
9	0.042	0.044	0.043	0.039	0.030	0.014	-0.011	-0.050	-0.114
10	0.048	0.054	0.059	0.062	0.062	0.057	0.044	0.020	-0.029
11	0.056	0.062	0.065	0.067	0.066	0.062	0.054	0.042	0.013
12	0.048	0.039	0.029	0.019	0.009	0.000	-0.004	-0.001	0.012
13	0.038	0.034	0.029	0.021	0.012	0.002	-0.010	-0.026	-0.049
14	0.056	0.059	0.063	0.068	0.073	0.079	0.091	0.116	0.170
15	0.058	0.066	0.075	0.087	0.100	0.117	0.142	0.184	0.260
16	0.030	0.036	0.042	0.047	0.051	0.053	0.051	0.044	0.023
17	0.032	0.035	0.037	0.039	0.039	0.039	0.039	0.041	0.048
18	0.032	0.037	0.040	0.044	0.047	0.051	0.056	0.061	0.061
19	0.039	0.047	0.055	0.062	0.069	0.076	0.086	0.106	0.150
20	0.061	0.070	0.080	0.092	0.106	0.122	0.143	0.170	0.202
21	0.045	0.050	0.053	0.053	0.052	0.051	0.051	0.057	0.077
22	0.044	0.049	0.054	0.061	0.069	0.080	0.093	0.107	0.112
23	0.038	0.039	0.041	0.044	0.047	0.052	0.060	0.071	0.081
24	0.046	0.044	0.043	0.043	0.046	0.052	0.062	0.076	0.092
25	0.067	0.077	0.091	0.108	0.129	0.154	0.181	0.212	0.256

TABLE 3.2: Case 3 - Proportionate Bias - Grouped Jackknife Estimator ($\widehat{\beta}_2$)

l_n \ ρ	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
1	0.048	0.057	0.061	0.060	0.055	0.043	0.022	-0.013	-0.075
2	0.053	0.066	0.076	0.081	0.080	0.073	0.056	0.024	-0.037
3	0.053	0.046	0.036	0.021	0.002	-0.022	-0.053	-0.097	-0.162
4	0.051	0.044	0.034	0.020	0.003	-0.018	-0.046	-0.086	-0.146
5	0.050	0.054	0.058	0.060	0.059	0.055	0.044	0.022	-0.021
6	0.062	0.069	0.075	0.079	0.081	0.079	0.072	0.055	0.019
7	0.045	0.042	0.036	0.026	0.012	-0.008	-0.035	-0.077	-0.141
8	0.050	0.048	0.043	0.036	0.027	0.017	0.007	-0.005	-0.024
9	0.056	0.059	0.060	0.061	0.061	0.059	0.054	0.043	0.015
10	0.056	0.058	0.061	0.064	0.067	0.067	0.062	0.044	0.002
11	0.042	0.058	0.073	0.086	0.097	0.104	0.101	0.077	0.009
12	0.104	0.105	0.108	0.112	0.119	0.131	0.153	0.196	0.272
13	0.045	0.043	0.041	0.035	0.025	0.008	-0.022	-0.072	-0.153
14	0.062	0.073	0.083	0.093	0.100	0.105	0.107	0.110	0.121
15	0.067	0.069	0.072	0.076	0.081	0.087	0.095	0.100	0.096
16	0.043	0.040	0.034	0.025	0.014	-0.002	-0.022	-0.046	-0.072
17	0.068	0.065	0.064	0.064	0.064	0.064	0.063	0.060	0.056
18	0.094	0.106	0.119	0.133	0.150	0.172	0.202	0.246	0.312
19	0.073	0.084	0.097	0.113	0.132	0.156	0.187	0.231	0.291
20	0.077	0.074	0.074	0.076	0.080	0.087	0.098	0.116	0.140
21	0.013	0.004	-0.005	-0.014	-0.022	-0.028	-0.033	-0.038	-0.047
22	0.028	0.035	0.043	0.052	0.061	0.070	0.071	0.056	0.005
23	0.043	0.037	0.031	0.023	0.014	0.004	-0.005	-0.011	-0.017
24	0.100	0.105	0.110	0.115	0.121	0.132	0.153	0.196	0.272
25	0.100	0.104	0.106	0.108	0.109	0.110	0.113	0.123	0.142

TABLE 3.3: Case 3 - Proportionate Bias - Newey and West Estimator ($\widehat{\beta}_1$)

ρ	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
m_n									
1	-0.053	-0.057	-0.065	-0.079	-0.098	-0.125	-0.159	-0.201	-0.255
2	-0.061	-0.065	-0.074	-0.088	-0.109	-0.136	-0.170	-0.213	-0.268
3	-0.071	-0.074	-0.081	-0.093	-0.108	-0.130	-0.158	-0.194	-0.242
4	-0.080	-0.084	-0.090	-0.100	-0.114	-0.133	-0.159	-0.193	-0.240
5	-0.090	-0.093	-0.098	-0.107	-0.119	-0.136	-0.159	-0.189	-0.232
6	-0.100	-0.103	-0.107	-0.115	-0.126	-0.141	-0.161	-0.187	-0.225
7	-0.110	-0.112	-0.116	-0.123	-0.133	-0.147	-0.165	-0.188	-0.222
8	-0.119	-0.121	-0.124	-0.131	-0.140	-0.153	-0.169	-0.192	-0.226
9	-0.128	-0.130	-0.133	-0.139	-0.147	-0.159	-0.175	-0.197	-0.231
10	-0.138	-0.139	-0.142	-0.148	-0.156	-0.167	-0.182	-0.203	-0.237
11	-0.147	-0.148	-0.151	-0.156	-0.164	-0.175	-0.189	-0.210	-0.243
12	-0.156	-0.157	-0.160	-0.165	-0.173	-0.183	-0.197	-0.217	-0.249
13	-0.165	-0.166	-0.169	-0.174	-0.182	-0.192	-0.205	-0.225	-0.256
14	-0.174	-0.175	-0.178	-0.183	-0.190	-0.200	-0.213	-0.233	-0.263
15	-0.183	-0.185	-0.187	-0.192	-0.199	-0.209	-0.222	-0.240	-0.271
16	-0.192	-0.194	-0.197	-0.201	-0.208	-0.217	-0.230	-0.249	-0.279
17	-0.201	-0.203	-0.206	-0.210	-0.217	-0.226	-0.239	-0.257	-0.287
18	-0.210	-0.212	-0.215	-0.219	-0.226	-0.235	-0.247	-0.264	-0.293
19	-0.219	-0.221	-0.223	-0.228	-0.234	-0.243	-0.254	-0.271	-0.299
20	-0.227	-0.229	-0.232	-0.236	-0.242	-0.250	-0.261	-0.276	-0.304
21	-0.236	-0.237	-0.240	-0.244	-0.250	-0.257	-0.268	-0.283	-0.310
22	-0.245	-0.246	-0.249	-0.253	-0.258	-0.265	-0.275	-0.290	-0.316
23	-0.254	-0.255	-0.258	-0.262	-0.267	-0.274	-0.284	-0.298	-0.324
24	-0.263	-0.264	-0.267	-0.271	-0.276	-0.283	-0.293	-0.307	-0.332
25	-0.272	-0.273	-0.276	-0.280	-0.285	-0.292	-0.301	-0.315	-0.340

TABLE 3.4: Case 3 - Proportionate Bias - Newey and West Estimator ($\widehat{\beta}_2$)

m_n \ ρ	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
1	-0.054	-0.047	-0.043	-0.043	-0.047	-0.058	-0.076	-0.107	-0.163
2	-0.064	-0.062	-0.063	-0.068	-0.078	-0.093	-0.115	-0.149	-0.206
3	-0.075	-0.075	-0.077	-0.082	-0.091	-0.104	-0.124	-0.155	-0.206
4	-0.085	-0.086	-0.088	-0.093	-0.101	-0.113	-0.131	-0.158	-0.204
5	-0.094	-0.096	-0.099	-0.104	-0.112	-0.123	-0.140	-0.167	-0.212
6	-0.103	-0.105	-0.108	-0.114	-0.121	-0.133	-0.150	-0.178	-0.224
7	-0.113	-0.115	-0.118	-0.123	-0.131	-0.143	-0.160	-0.187	-0.233
8	-0.122	-0.124	-0.128	-0.133	-0.141	-0.152	-0.169	-0.197	-0.244
9	-0.133	-0.135	-0.139	-0.144	-0.152	-0.164	-0.181	-0.208	-0.254
10	-0.144	-0.146	-0.149	-0.155	-0.162	-0.174	-0.191	-0.218	-0.263
11	-0.154	-0.156	-0.160	-0.165	-0.172	-0.184	-0.201	-0.227	-0.271
12	-0.164	-0.166	-0.169	-0.174	-0.181	-0.193	-0.209	-0.236	-0.278
13	-0.174	-0.176	-0.179	-0.183	-0.191	-0.201	-0.218	-0.243	-0.285
14	-0.183	-0.185	-0.188	-0.192	-0.199	-0.210	-0.226	-0.251	-0.292
15	-0.192	-0.194	-0.197	-0.201	-0.208	-0.218	-0.234	-0.259	-0.299
16	-0.201	-0.203	-0.206	-0.210	-0.216	-0.226	-0.242	-0.266	-0.305
17	-0.210	-0.212	-0.214	-0.218	-0.224	-0.234	-0.249	-0.273	-0.310
18	-0.219	-0.220	-0.223	-0.226	-0.232	-0.242	-0.256	-0.279	-0.316
19	-0.228	-0.229	-0.231	-0.234	-0.240	-0.249	-0.263	-0.286	-0.322
20	-0.236	-0.237	-0.239	-0.242	-0.247	-0.256	-0.270	-0.292	-0.327
21	-0.244	-0.245	-0.247	-0.249	-0.254	-0.263	-0.276	-0.298	-0.333
22	-0.252	-0.253	-0.255	-0.257	-0.262	-0.270	-0.283	-0.304	-0.339
23	-0.261	-0.262	-0.263	-0.266	-0.270	-0.278	-0.291	-0.312	-0.347
24	-0.269	-0.270	-0.272	-0.274	-0.279	-0.287	-0.300	-0.321	-0.355
25	-0.278	-0.279	-0.280	-0.283	-0.288	-0.295	-0.309	-0.329	-0.364

7. REFERENCES

Andrews, D. W. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica*, 59, 817-858.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence, *The Annals of Statistics*, Vol. 14, N. 3, 1171-1179.

Chesher, A. and Jewitt, I. (1987). The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica*, 55, 1217-1222.

Chesher, A. and Austin, G. (1991). The finite-sample distributions of heteroskedasticity robust Wald statistics. *Journal of Econometrics*, 47, 153-173.

Cook and Weisberg (1982). *Residuals and Influence in Regression*, New York, Chapman & Hall.

Hansen, L. P. (1982). Large samples properties of Generalized Method of Moment Estimators. *Econometrica*, 50, 1029-1054.

Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17, 1217-1241.

MacKinnon, J. and White, H. (1985). Some Heteroskedasticity - Consistente Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics*, 29, 305-325.

Newey, W. and West, K. (1987). A Simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55, 703-708.

Passos, J. (1994). *Finite-sample performance of heteroskedasticity and autocorrelation consistent covariance matrix estimator*, University of Bristol - Department of Economics, Discussion Paper no 94/385.

Politis, N. and Romano, J. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89, 1303-1313.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-60.

White, Halbert (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817-838.

White, Halbert (1984). *Asymptotic theory for econometricians*, Orlando, Academic Press.

CHAPTER 4 - ADAPTING FOR HETEROSKEDASTICITY OF UNKNOWN FORM: AN APPROACH VIA ADE

1. INTRODUCTION

Heteroskedasticity is an econometric problem associated with the non-constancy of the error variance and arises in numerous applications as in the analysis of cross-section data. Neglecting this problem can lead to invalid inferences about the model due to erroneous standard errors. One common solution has been given in the literature through robust standard errors [see White (1980)]. Another solution is a two-step estimator through generalized least squares estimation.

Much of the recent literature has been concerned to efficient estimation and testing under heteroskedasticity. An example is the efficient estimation under weak distributional assumptions. It is also well known that one of the most common restrictions is the assumption that some location measure of the conditional distribution has a known functional form [for a good survey of this kind of models see Powell (1992)]. In such a case, Chamberlain (1987) has shown that the maximum efficiency is attainable *via* weighted least squares (WLS) estimation with weighted proportional to the inverse of the conditional variance of the response variable. However this estimator depends on a sequence of weights that in general are unknown and methods to deal with this problem have been proposed in the literature.

One possibility is based on a parametric correction, when the pattern of the error variance is a function of a known skedastic function. In this case a large econometric literature appeared in the last two decades, studying possible parametric forms for the error variance as well as its properties [see for example Greene (1993), chapter 4]. In a model with one regressor a simple plot against the residuals can give some guidance to identify this parametric form. However in a general context this approach seems to be infeasible.

Another possibility can be based in the particular structure of the model's family considered. For example, in a Poisson regression model the weight is known to be equal to the conditional mean function. Moreover, in the case of quasi-maximum-likelihood estimation of a conditional mean, Gourieroux, Monfort and Trognon (1984) have shown that using a likelihood in the linear exponential family results in a implicit choice for the weighting function.

In a general context, Carrol (1982) and Robinson (1987) have proposed efficient estimation using nonparametric methods to handle the weight sequence. Recently Andrews (1994) generalized this approach, proving $\sqrt{n}$-consistency and asymptotic normality of estimators that minimize a criterion function that depends on some nonparametric preliminary estimator[1]. He also provides an application to semiparametric WLS estimators of partially parametric regression models. Having in mind these results, the approaches of Carrol and Robinson as well as the method proposed in this chapter can be viewed as particular cases.

In this chapter one presents an estimator alternative to Carrol and Robinson for a general model under the assumption that the conditional expectation has a known

[1]Andrews (1994) page 43, define these estimators as MINPIN estimators: *'(...) estimators that MINimize a criterion function that may depend on a Preliminary Infinite dimensional Nuisance parameter estimator'.*

functional form. Kernel and average derivative estimation (henceforth ADE) will be applied to estimate the conditional variance. This method presents some advantages: a) it provides a natural test for heteroskedasticity, simply by assessing the statistical significance of the ADE; b) it allow us a straightforward identification of the covariates responsible for the presence of heteroskedasticity and c) it generalizes Carrol's approach.

The Chapter is organized as follows. In Section 2 is presented the estimation method of the weight sequence. The assumptions of Andrews (1994) are considered for the asymptotic normality of WLS estimator. These assumptions are discussed for the particular case considered in this chapter. Section 3 presents two alternatives tests for heteroskedasticity: a Wald test and a score based test. Furthermore a test for the validity of the conditional variance assumption is also presented. The reason is that the efficiency of the WLS is strongly dependent on the validity of this assumption. The test is based on a similar one proposed by Wooldridge (1990) in a parametric context. Section 4 presents a Monte Carlo study. The main purpose is to access the finite sample performance of the estimator presented and to confront these results with the method suggested in Carrol (1982). Finally Section 5 presents some concluding remarks.

2. THE ESTIMATOR

In this Section one presents a two-step efficient estimator for models where the conditional mean restriction has a known functional form. To fix ideas consider an independent sample $\{(y_i, x_i),\ i = 1, 2, \ldots, n\}$ from an absolutely continuous $(k+1)$-variate distribution function with joint density $f(y, x) = f(y \mid x)f(x)$, where $y \in \Re$ is the dependent variable and $x \in \Re^k$ a vector of k-variate regressors.

ASSUMPTION 1: The true regression function has the following expression,

$$E(y \mid x) = g(x, \theta_o), \tag{2.1}$$

where $\theta_o \in \Theta \subset \Re^k$ is an unknown $k \times 1$ vector to be estimated and $g(x, \theta)$ is a known real-valued function, measurable on $\Re^k$ for each $\theta \in \Theta$ and continuous on Θ.

Under this conditional expectation assumption it is well known [see Chamberlain (1987)] that the maximum attainable efficiency is given by the weighted (nonlinear) least squares estimator (henceforth WNLS), with weight equal to the inverse of $Var(y \mid x) = \sigma^2(x)$. If in particular $\sigma^2(x)$ were known and equal to the true value $\sigma_o^2(x)$ the WNLS estimator of θ, $\widetilde{\theta}_n$, could be easily computed as[2]

$$\min_{\theta \in \Theta} \; Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{[y_i - g(x_i, \theta)]^2}{\sigma_o^2(x_i)} \; , \tag{2.2}$$

with first-order condition

$$\begin{aligned} \overline{d}_n(\theta) &= \tfrac{1}{n} \sum_{i=1}^{n} d(z_i, \theta, \sigma_o^2(x_i)) \\ &= \tfrac{1}{n} \sum_{i=1}^{n} \nabla_\theta g(x_i, \theta)' \, \frac{y_i - g(x_i, \theta)}{\sigma_o^2(x_i)} \;\; = \;\; 0 \; , \end{aligned}$$

where $\nabla_\theta g(x_i, \theta)$ is a vector of derivatives and $z_i = (y_i, x_i)$. Expanding $\overline{d}_n(\theta)$ around θ_o and solving for $\widetilde{\theta}_n - \theta_o$ gives

$$\sqrt{n}(\widetilde{\theta}_n - \theta_o) = - \left[n^{-1} \sum_{i=1}^{n} \nabla_\theta d(z_i, \overline{\theta}, \sigma_o^2(x_i)) \right]^{-1} \sqrt{n} \left(n^{-1} \sum_{i=1}^{n} d(z_i, \theta_o, \sigma_o^2(x_i)) \right) \; , \tag{2.3}$$

where $\overline{\theta}$ lies on the line segment joining $\widetilde{\theta}_n$ and θ_o. The asymptotic normality of

[2]Given Assumption 1, the existence of $\widehat{\theta}_n$ is ensured by Lemma 2 of Jennrich (1969).

$\widetilde{\theta}_n$ can be easily settled through a uniform weak law of large numbers [see Andrews(1992,1994)] together with a CLT for a sequence of rv's, such as the Lindeberg-Feller CLT.

Theorem 1: Let $\mathcal{N}_o \subset \Theta$ be a neighbourhood of $\theta_o$where θ_o is an interior point of the parameter space Θ. Define $S = \lim_{n\to\infty} Var[n^{-1/2}\sum_1^n d(z_i, \theta_o, \sigma_o^2(x_i))]$ and $D = \lim_{n\to\infty}(1/n)\sum_1^n \nabla_\theta E d(z_i, \theta_o, \sigma_o^2(x_i))$. Consider the following assumptions,

(a) $\widetilde{\theta}_n \xrightarrow{p} \theta_o \in \Theta \subset \Re^k$.

(b) $n^{-1/2}\sum_1^n d(z_i, \theta_o, \sigma_o^2(x_i)) \xrightarrow{d} N(0, S)$.

(c) $d(z_i, \theta, \sigma_o^2(x_i))$ is continuously differentiable in θ on $\mathcal{N}_o$, $\forall i \geq 1$ and the sequences $\{d(z_i, \theta, \sigma_o^2(x_i)) : i \geq 1\}$ and $\{\nabla_\theta d(z_i, \theta, \sigma_o^2(x_i)) : i \geq 1\}$ satisfies uniform weak law of large numbers over $\mathcal{N}_o$. Moreover it is assumed that $d(\theta) = \lim_{n\to\infty}(1/n)\sum_1^n E d(z_i, \theta, \sigma_o^2(x_i))$ and $D(\theta) = \lim_{n\to\infty}(1/n)\sum_1^n E\nabla_\theta d(z_i, \theta, \sigma_o^2(x_i))$ exist uniformly over $\mathcal{N}_o$ and are continuous at θ_o with respect to some pseudo-metric on $\mathcal{N}_o$ for which $\widetilde{\theta}_n \xrightarrow{p} \theta_o$.

(d) The matrix D is non-singular.

Under the above the sequence $\{\widetilde{\theta}_n\}$ satisfies,

$$\sqrt{n}(\widetilde{\theta}_n - \theta_o) \xrightarrow{d} N(0, D^{-1}SD^{-1'})$$

Proof. Assumption (a) is easily verified by uniform weak law of large numbers [see for example Pötscher and Prucha (1989), Lemma 3.1 or Andrews (1994), Theorem A-1]. Considering the first term of the right hand side of (2.2) and using assumption (a), (c) and Lemma 3 of Jennrich (1969), it is straightforward to show that,

$$n^{-1}\sum_{i=1}^n \nabla_\theta d(z_i, \bar{\theta}, \sigma_o^2(x_i)) = n^{-1}\sum_{i=1}^n \nabla_\theta d(z_i, \theta_o, \sigma_o^2(x_i)) + o_p(1)$$

and thus

$$n^{-1}\sum_{i=1}^{n}\nabla_\theta d(z_i,\overline{\theta},\sigma_o^2(x_i)) \xrightarrow{p} D\ .$$

Given assumption (b) and (d) the result of the theorem follows.□

If $\{z_i : i \geq 1\}$ is independent it is also straightforward to show that $S = D$ and the asymptotic covariance of $\widetilde{\theta}$ simplifies to,

$$Asy\,Var\{n^{1/2}(\widetilde{\theta}_n - \theta_o)\} = \lim_{n\to\infty}\left\{\sum_{i=1}^{n} E\left[\frac{\nabla_\theta g(x_i,\theta_o)\nabla_\theta g(x_i,\theta_o)'}{\sigma_o^2(x_i)}\right]\right\}^{-1}. \qquad (2.4)$$

However in a general context $\sigma_o^2(x_i)$ is unknown and thus an asymptotically efficient WNLS estimator requires a consistent estimator of $\sigma_o^2(x_i)$. Robinson (1987) [see also Stone (1977)] proposed in the linear model $(g(x,\theta) = x'\theta)$ a consistent estimator of $\sigma^2(x)$ through k-Nearest Neighbour (k-NN) estimation[3], given by

$$\widehat{\sigma}^2(x_i) = \sum_{j=1}^{n} w_{ij}\widehat{\varepsilon}_i^2$$

where w_{ij} is a weight dependent on some metric defined on $\Re^k$ and $\widehat{\varepsilon}_i = y_i - x'\widehat{\theta}_{LS}$ with $\widehat{\theta}_{LS}$ a preliminary consistent estimator of θ, e.g., least squares estimator. Carrol's (1982) approach differs from the above in the computation of the weight w_{ij} where $\widehat{\sigma}^2(x_i)$ is the nonparametric Nadaraya-Watson kernel estimator[4]. However this approach can be infeasible. The problem is that the rate of convergence of kernel estimators is slow if the number of the covariates is large [see Härdle (1990), pag. 91]. To handle this problem he tries a dimension reduction technique, assuming that the variance of the error term is an unknown function of the mean response variable,

[3]For an exposition of this method see for example Härdle (1990).

[4]Carrol gives a proof of the asymptotic normality of the WLS estimator only in the univariate case, where $x \in \Re$. In the multivariate case see Hidalgo (1992).

i.e., $\sigma^2[g(x,\theta)] = \sigma^2(x'\theta)$ with θ replaced by $\widehat{\theta}_{LS}$. Considering situations in which $\sigma^2(x'\theta) = x'\theta$, Carrol's assumption is easily checked in models like the Poisson, but not in general and thus it can be very restrictive. In what follows a similar approach is used but with a more general assumption.

ASSUMPTION 2: The conditional variance is in an index form,

$$\begin{aligned} Var(y \mid x) &= \sigma_o^2(x) \\ &= m_o(x'\zeta_o) \end{aligned} \tag{2.5}$$

where $\zeta_o \in \mathcal{Z} \subset \Re^k$ is a vector of unknown parameters not necessarily equal to θ_o and $m_o(\cdot) \in \mathcal{M}_\rho$ is a vector-valued function defined on some Euclidean Space $\mathcal{M}_\rho$ where ρ is a pseudo-metric.

In order that expression (2.5) can describe a variance, the set $\mathcal{M}_\rho$ should be restricted as follows. Let $\mathcal{X} \subset \Re^k$ be the set of all possible realizations of x and $\mathcal{X}^*$ an open and bounded subset of $\mathcal{X}$. For any $\zeta \in \mathcal{Z}$ and $x \in X$ let $\mathcal{X}_\zeta \subset \Re$ be the set of all possible realizations of $x'\zeta$. For some pseudo-metric ρ consider as in Andrews (1994) only those functions $m(\cdot) \in \mathcal{M}_\rho$ such that,

$$\mathcal{M}_\rho = \left\{ m(\cdot) : \| m_o(\cdot) \|_{q,\mathcal{X}_\zeta^*} \leq \infty, \inf_{x'\zeta \in \mathcal{X}^*} | m(x'\zeta) | \geq \xi \right\}$$

for any small $\xi > 0$ and $q > 1/2$. For a given norm ρ these restrictions guarantee the finiteness and the non-negativity of the conditional variance.

Defining the error component of the model as $\varepsilon_i = y_i - g(x_i,\theta)$, the conditional variance (2.5) can be rewritten as

$$E(\varepsilon_i^2 \mid x_i) = m(x_i'\zeta). \tag{2.6}$$

For a preliminary $\sqrt{n}$ - consistent estimator of θ, $\widehat{\theta}_{LS}$, replace ε_i^2 in the above expression to $\widehat{\varepsilon}_i^2$. Expanding $\widehat{\varepsilon}_i^2$ around θ_o one has, $\widehat{\varepsilon}_i^2 = \varepsilon_i^2 + \Delta_i(n)$ where $\Delta_i(n) \to 0$ as n increases and thus for n sufficiently large $E(\widehat{\varepsilon}_i^2) \approx E(\varepsilon_i^2)$. However it remains the problem of the unknown quantity ζ. To estimate ζ one will apply average derivative estimator (henceforth ADE) of Stoker (1986) and Härdle and Stoker (1989) to estimate ζ and then non-parametric kernel estimation to handle the unknown function $m(\cdot)$.

The basic idea underlying ADE, due to Stoker (1986), is to measure the local effects of changing x on $\widehat{\varepsilon}^2$, $\partial m(x'\zeta)/\partial x$, as the average of these effects,

$$\delta = E_x \left[\frac{\partial m(x'\zeta)}{\partial x} \right], \tag{2.7}$$

where the expectation is computed over x. Simplifying the derivative in brackets one has

$$\delta = E_x \left[\frac{\partial m(x'\zeta)}{\partial (x'\zeta)} \right] \zeta = \gamma\zeta, \tag{2.8}$$

so that δ is proportional to ζ. Therefore the average derivative, δ, determines ζ up to a scale and, provided $\gamma \neq 0$, one can replace ζ by δ in (2.6), with $m(\cdot)$ obeying the normalisation $E[\partial m(x'\delta)/\partial(x'\delta)] = 1$.

To estimate δ one of the following estimators can be used: (a) direct estimator; (b) indirect estimator and (c) slope estimator. Asymptotically they are equivalent as shown recently by Stoker (1991). In what follows, the indirect estimator of Härdle and Stoker (1989) is considered.

Applying integration by parts in (2.7) [see Stoker (1986), Theorem 1] one has

$$\delta = E_x \left[l(x)\hat{\varepsilon}^2 \right] \tag{2.9}$$

where $l(x) = -f'(x)/f(x)$ is the score of the marginal density $f(x)$. Therefore one can define the ADE estimator of δ as the sample analog [see Härdle and Stoker (1989)]

$$\hat{\delta}_n = \frac{1}{n} \sum_{i=1}^{n} \hat{l}_h(x_i)\hat{\varepsilon}_i^2 \hat{I}_i \tag{2.10}$$

where $\hat{I}_i = I[\hat{f}_h(x_i) > b]$ is an indicator function that trims values greater than some bound $b = b_n$ such that $b \to 0$ as $n \to \infty$, $h = h_n$ is the bandwidth parameter with $h \to 0$ as $n \to \infty$ and $\hat{f}_h(x_i)$ is the estimated density function at x_i. In particular at any x the estimator of $f(x)$ is defined as

$$\hat{f}_h(x) = \frac{1}{n}\frac{1}{h^k} \sum_{j=1}^{n} k(\frac{x - x_i}{h}) \ ,$$

where $k(\cdot)$ is a kernel function that obeys some regularity conditions [see Härdle and Stoker (1989)].

Given an estimate of δ one forms the quantity $\hat{v} = x'\hat{\delta}_n$ and the relation between $\hat{\varepsilon}$ and $x'\hat{\delta}_n$ will be estimated by nonparametric kernel estimator

$$\widehat{m}(v) = \frac{\frac{1}{nh'} \sum_{j=1}^{n} k(\frac{v - \hat{v}_j}{h'})\hat{\varepsilon}_j^2}{\frac{1}{nh'} \sum_{j=1}^{n} k(\frac{v - \hat{v}_j}{h'})} \tag{2.11}$$

where h' is define as above. Considering

$$w_j = \frac{k(\frac{v - \hat{v}_j}{h'})}{\sum_{j=1}^{n} k(\frac{v - \hat{v}_j}{h'})}$$

one has

$$\widehat{m}(v) = \sum_{j=1}^{n} w_j \widehat{\varepsilon}_j^2 \widehat{I}_i$$

that is similar (in form) to the estimator presented by Carrol (1982) and Robinson (1987).

Assumptions required for the consistency and asymptotic normality of ADE and Nadaraya-Watson kernel estimator can be seen for example in Härdle and Stoker (1989).

From now on one addresses to the study of the asymptotic distribution of the WNLS estimator of θ, $\widehat{\theta}_n$, given as a solution of (2.2) but with the conditional variance of y estimated consistently by (2.11). More precisely what is the effect in the distribution of $\widetilde{\theta}_n$ when in the expansion (2.3) $\sigma_o^2(x)$ is replaced by $\widehat{\sigma}^2(x) = \widehat{m}(x'\widehat{\zeta})$. To answer this question one has to show first that the replacement of ζ to $\widehat{\zeta}$ does not change the asymptotic properties of $\widehat{m}(\cdot)$. If assumption 3 is true and ζ known then the one regressor nonparametric estimator $\widehat{m}(\cdot)$ consistently estimate $m(\cdot)$ [see for example Härdle (1990), proposition 3.1.1]. Using Theorem 3.1 of Härdle and Stoker (1989) $\widehat{\zeta}$ is $\sqrt{n}$ consistent to ζ_o and therefore one can use $\widehat{\zeta}$ in place of ζ without changing the convergence rate of $\widehat{m}(\cdot)$ [see Härdle and Stoker (1989), Theorem 3.3]. Similar result was obteined by Carrol (1982) with $\widehat{\theta}_{LS}$ replacing $\widehat{\zeta}$.

Since $\widehat{m}$ is consistent to m_o in probability it seems reasonable that replacing $\sigma_o^2(x)$ by $\widehat{m}$ in (2.3) provides us a $\sqrt{n}$-consistent and asymptotic normal estimator of θ. A proof of this assertion in a general framework of semiparametric WLS estimators of partially parametric regression models is given in Andrews (1994). To prove the asymptotic normality of $\widehat{\theta}_n$ one needs to introduce some additional assumptions as

well the definition of stochastic equicontinuity [see Andrews (1991,1994)].

Definition [Andrews (1994)]: Define an empirical process

$$\nu_n(m) = n^{-1/2} \sum_{i=1}^{n} \{d(z_i, \theta_o, m) - E[d(z_i, \theta_o, m)]\}$$

where $m \in \mathcal{M}_\rho$. The sequence $\{\nu_n(m)\}_1^\infty$ is stochastically equicontinuous at m_o if all sequences $\{\widehat{m}_n\}_1^\infty$ satisfying $\rho(\widehat{m}_n, m_o) \xrightarrow{p} 0$ lead to $\nu_n(\widehat{m}_n) - \nu_n(m_o) \xrightarrow{p} 0$.

Following Andrews (1994) this definition represents '*a stochastic and asymptotic version of the concept of the continuity of a function*'.

ASSUMPTION 3 [Andrews (1994)]: replacing $\sigma_o^2(x)$ in (2.3) to $\widehat{m} = \widehat{m}(x'\widehat{\delta})$ the following assumptions are sufficient for the asymptotic normality of $\widehat{\theta}_n$:

(a) $\widehat{\theta}_n \xrightarrow{p} \theta_o \in \Theta \subset \Re^k$ and θ_o is an interior point of Θ.

(b) $\widehat{m}$ lies in a pseudo-metric space $\mathcal{M}_\rho$ wp$\rightarrow$ 1 and $\widehat{m} \xrightarrow{p} m_o \in \mathcal{M}_\rho$ with respect to the pseudo-metric ρ, i.e, $\rho(\widehat{m}, m_o) \xrightarrow{p} 0$.

(c) $\sqrt{n}E[n^{-1} \sum_1^n d(z_i, \theta_o, \widehat{m})] \xrightarrow{p} 0$.

(d) $\nu_n(m_o) \xrightarrow{d} N(0, S)$.

(e) the sequence $\{\nu_n(\cdot) : n \geq k\}$ is stochastically equicontinuous at $m_o \in \mathcal{M}$.

(f) $d(z_i, \theta, m)$ is continuously differentiable in θ on $\mathcal{N}_o$, $\forall m \in \mathcal{M}$, $\forall i \geq 1$ and the sequences $\{d(z_i, \theta, m : i \geq 1\}$ and $\{\nabla_\theta d(z_i, \theta, m) : i \geq 1\}$ satisfy uniform weak law of large numbers over $\mathcal{N}_o \times \mathcal{M}$. Moreover it is assumed that $d(\theta, m) = \lim_{n\to\infty}(1/n) \sum_1^n Ed(z_i, \theta, m)$ and $D(\theta, m) = \lim_{n\to\infty}(1/n) \sum_1^n E\nabla_\theta d(z_i, \theta, m)$ exist uniformly over $\mathcal{N}_o \times \mathcal{M}$ and are continuous at (θ_o, m_o) with respect to some pseudo-metric on $\mathcal{N}_o \times \mathcal{M}$ for which $(\widehat{\theta}, \widehat{m}) \xrightarrow{p} (\theta_o, m_o)$.

(g) The matrix D is non-singular.

As pointed out by Andrews (1994) the '*difference between this assumption and assumptions commonly used to establish asymptotic normality of nonlinear parametric estimators is the appearance of (b), (c) and (e)*'. Sufficient conditions for assumption 3(e) can be seen in Andrews(1991,1994). Assumption 3(a) is proved by Andrews (1994), Theorem A-1. Assumption 3(c) is an orthogonality assumption which means that the asymptotic distribution of $\widehat{\theta}$ does not change if $\sigma_o^2(x)$ is replaced to $\widehat{m}$. This assumption requires $n^{1/4}$-consistency of $\widehat{m}$ to m_o. A proof for the case $g(x,\theta) = x'\theta$ can be seen in Carrol (1982), Theorem 5.2. In the non-linear case if $g(x,\theta)$ is sufficiently smooth in a neighbourhood of θ_o the $n^{1/4}$-consistency of $\widehat{m}$ follows. Therefore assumptions (b) and (c) of Andrews can be replaced by the assumption of $n^{1/4}$-consistency of $\widehat{m}$. However the validity of this assumption requires some primitive conditions about the kernel $k(\cdot)$, bandwidth and the first derivative of $g(x,\theta)$ [see assumptions A4-A7 and Theorem 5.2 of Carrol (1987)].

Theorem 2: Given assumption 3,

$$\begin{aligned} \sqrt{n}(\widehat{\theta} - \theta_o) &= \sqrt{n}(\widetilde{\theta} - \theta_o) + o_p(1) \\ &\xrightarrow{d} N(0, \Sigma) \end{aligned} \tag{2.12}$$

where $\Sigma = D^{-1}SD^{-1'}$ and is the asymptotic covariance matrix (2.4) if the sequence $\{z_i : i \geq 1\}$ is independent.

Proof. See Andrews 1994, Theorem 1 and 2.

3. TESTING THE HETEROSKEDASTICITY AND THE VARIANCE ASSUMPTION

Under the conditional variance given by assumption 2 and having in mind the interpretation of the ADE, a test for heteroskedasticity can be based on the hypothesis $H_o : \delta = 0$ against $H_1 : \delta \neq 0$. One possibility is the evaluation of H_o through the Wald statistic [see Härdle and Stoker (1989)]

$$W = n(R\widehat{\delta} - r_o)'(R\widehat{\Sigma}_\delta R')^{-1}(R\widehat{\delta} - r_o) \tag{3.1}$$

where, in this case, $R = I_k$ is the identity matrix of order k, $r_o = 0$ and $\widehat{\Sigma}_\delta$ is the estimator of the covariance matrix of $\widehat{\delta}$ [see for example Härdle and Stoker (1989) for a definition of $\widehat{\Sigma}_\delta$]. If H_o is true then $W \sim \chi^2(k)$.

Another possible test is a score type test. Assuming $m(0) = \sigma^2$ suppose that one has to decide if the data $\{(y_i, x_i) : i = 1, \ldots, n\}$ was generated from $f(y \mid x; \theta, m(x'\delta))$ or $f(y \mid x; \theta, \sigma^2)$. To deal with this one can approximate the density $f(y \mid x; \theta, m(x'\delta))$ by taking Taylor series approximations around $\delta = 0$. Retaining terms to second order in the elements of δ and rearrange one has,

$$\begin{aligned} f(y \mid x; \theta, m(x'\delta)) \;=\; & f(y \mid x; \theta, \sigma^2)\left\{1 + \delta_i x_i m' F'(y \mid x; \theta, \sigma^2) + \right. \\ & \tfrac{1}{2}\delta_i \delta_j x_i x_j \left[m'' F'(y \mid x; \theta, \sigma^2) + \right. \\ & \left. \left. m' m' \left(F''(y \mid x; \theta, \sigma^2) + F'(y \mid x; \theta, \sigma^2) F'(y \mid x; \theta, \sigma^2)\right)\right]\right\} + \\ & o(\delta^2) \end{aligned} \tag{3.2}$$

where summation from 1 to k is to be performed over the indices in subscripts,

$i, j = 1, 2, \ldots, k$, and

$$F'(y \mid x; \theta, \sigma^2) = \frac{\partial \ln f(y \mid x; \theta, m(x'\delta))}{\partial m(x'\delta)} \mid_{\delta=0}$$

$$F''(y \mid x; \theta, \sigma^2) = \frac{\partial^2 \ln f(y \mid x; \theta, m(x'\delta))}{\partial m(x'\delta)^2} \mid_{\delta=0}$$

$$m' = \frac{dm(x'\delta)}{d(x'\delta)} \mid_{\delta=0}$$

$$m'' = \frac{d^2 m(x'\delta)}{d(x'\delta)^2} \mid_{\delta=0}$$

and subscripts indicates elements of vectors.

If H_o is true one expects that the score,

$$\frac{\partial \ln f(y \mid x; \theta, m(x'\delta))}{\partial \delta} \mid_{\delta=0} = x_i m' F'(y \mid x; \theta, \sigma^2) \,, \tag{3.3}$$

is approximately equal to zero. As it can be seen this quantity is similar to the result attained by Breush and Pagan (1979), suggesting that the shape of the function $m(\cdot)$ is irrelevant when testing for heteroskedasticity. Note that the same conclusion can be extended to the Wald test presented above.

If the conditional variance assumption is misspecified the WNLS is still consistent but inefficient due to invalid standard errors and therefore infeasible for inferences purposes. For this reason the question that arises at this point is to assess the validity of the index form (2.5) under the conditional mean restriction (2.1).

In a parametric context Wooldridge (1990, 1991) pointed out that if the conditional variance is correctly specified any function of the covariates should be uncorrelated with $\varepsilon^2 - E(\varepsilon^2 \mid x)$. Having in mind this result the test can be based on the hypothesis

$H_o : E(\varepsilon^2 \mid x) = m(x'\delta)$ against the general alternative $H_1 : E(\varepsilon^2 \mid x) = m^*(x)$ where $m^*(\cdot)$ is any unspecified function of the covariates. The significance of this independence can be assessed through a t-test of the LS estimated coefficient of α_2 in the auxiliary regression,

$$\widehat{\varepsilon}_i^2 - \widehat{m}(x_i'\widehat{\delta}) = \alpha_1 + \alpha_2 \widehat{m}^*(x_i) + v_i \ , \quad i = 1, 2, \ldots, n \tag{3.4}$$

where in the right hand side of (3.4) $\widehat{m}^*(x_i)$ is the nonparametric kernel smoother.

4. APPLICATION

This Section deals with the computation of standard errors under heteroskedasticity, using the ADE and the Carrol's approach. In particular it is presented a Monte Carlo study. The purpose is to access the finite sample performance of the method presented in Section 2 and to show that in some cases, particularly when the pattern of the heteroskedasticity is significantly different from the pattern of the mean response model (e.g., when the subspace generated by them are orthogonal), the approach suggested in this chapter provides better results than Carrol's, due to its flexibility. As a reference for comparison, the standard errors computed from the true weighted sequence are used.

In what follows the model considered is the linear regression model

$$y_i = \alpha + x_i\theta + \varepsilon_i$$

where $i = 1, 2, \ldots, 50$, y_i is the i^{th} observed value, $x_i = [x_{1i} \quad x_{2i}]$ is a fixed 1×2 row vector, α is a unknown parameter, θ is a 2×1 vector of unknown parameters and ε_i is the error term. It is assumed that $\varepsilon_i \sim N(0, \Omega)$ and Ω is a diagonal matrix with elements given by

$$\sigma_i^2 = \exp(x_i\zeta)$$

where ζ and is a 2×1 vector of unknow parameters. For the true parameters of this specification it is considered $\alpha = 1$, $\theta = [0.6 \quad 0.8]'$ and $\zeta = [-0.8 \quad 0.6]$. Because the distributions of many estimators depend crucially on the way the regressors are distributed [see Chesher and Peters (1994)] two different designs are considered: in design 1, $x_{1i}, x_{2i} \sim N(0,1)$; in design 2 $x_{1i} \sim N(0,1)$ and $x_{2i} \sim \chi^2(1)$.

Given the model and the covariates, 1000 Monte Carlo replications are considered to produce the results that follows. All computations were done in GAUSS.

The estimation procedure has two steps: in the first step the residuals are computed from $\widehat{\varepsilon}_i = y_i - \widehat{\alpha}^{LS} - x_i\widehat{\theta}^{LS}$, for each $i = 1, 2, \ldots, 50$ where $\widehat{\alpha}^{LS}$ and $\widehat{\theta}^{LS}$ are the least squares estimates. Given the residuals, ζ and $m(.)$ are estimated by applying the results of Section 2. For the kernel function it is considered the gaussian kernel[5]. In all the examples presented, the bandwidth considered is 1.0 and 0.3 for ADE and the nonparametric kernel estimation of $m(.)$, respectively. In the estimation of $m(.)$ the bandwidth is computed by Generalized Cross Validation (GCV). In the estimation of ζ there is not a simple and useful procedure to compute the bandwidth. However, Härdle and Stoker (1989) pointed out that Monte Carlo experience suggests that reasonable small-sample performance is obtained by setting the bandwidth in the range of one to two standard deviations of the predictors. To avoid unbounded situations, $\widehat{I}_i = I[\widehat{f}_h(x_i) > b]$ in expression (2.10) is considered with $b = 0.06$, meaning that 6% of the observations with the smallest estimated density are dropped (in this case, 3 observations).

[5]Other kernels are possible but the results do not change significantlly [see for example Härdle (1990)].

In the second step the estimated variance is then used as the weighted sequence in the computation of the WLS estimate of θ. Tables[6] 1.1 to 2.2 and Figure 1A to 2F summarize the results achieved in a 1000 Monte Carlo replications.

The estimated values of θ presented in Table 1.1 and 1.2, for all four estimators, are unbiased as expected. The first moment is not affected even if the weighted sequence is incorrectly estimated. Aditionally, these estimators are consistent and asymptotically normal. The finite-sample distribution functions of the standartized estimators are not different from the asymptotic results [see the QQ-plots in Figures 1B-1E and 2B-2E, for the ADE and Carrol approaches]. However, the efficiency of $\widehat{\theta}$ depends crucially on the particular estimator considered. If the functional form of σ^2 is correct, it is well known that the WLS estimator of θ attains (2.4) asymptotically. In the example considered in this Section, the weighted sequence, σ_i^2, is incorrectly estimated using Carrol's procedure (as well as the OLS). As a consequence, Carrol's estimator leads to invalid standard errors for $\widehat{\theta}$ with asymptotic covariance matrix exceeding those given from expression (2.4). In the finite-sample case considered in this Section the std($\widehat{\theta}$) of the WLS-Carrol estimator appears much bigger than the std($\widehat{\theta}$) of the WLS(σ^2) estimator. Among all the estimators considered, the WLS-ADE is the most efficient. This can be seen in Table 1.1 and 2.1.

[6] **WLS(σ^2)** means the values of the WLS estimator of θ, using the true values σ^2; **WLS-ADE** is for the values of the WLS estimator of θ, using the method proposed in this chapter; **WLS-Carrol** is for the estimator proposed in Carrol (1982) and **OLS** is for the values of the least squares estimator of θ. The values presented are computed over a 1000 Monte Carlo replications.

Table 1.1: Design 1 - Comparison of results

	$\widehat{\theta}_1$	$\widehat{\theta}_2$	std($\widehat{\theta}_1$)	std($\widehat{\theta}_2$)
WLS(σ^2)	0.596	0.799	0.163	0.124
WLS-ADE	0.597	0.799	0.183	0.136
WLS-Carrol	0.604	0.798	0.221	0.179
OLS	0.594	0.796	0.241	0.216

Table 1.2: Design 1 - Results from ADE

$\widehat{\zeta}$	-0.791	0.549
std($\widehat{\zeta}$)	(0.162)	(0.216)

Table 2.1: Design 2 - Comparison of results

	$\widehat{\theta}_1$	$\widehat{\theta}_2$	std($\widehat{\theta}_1$)	std($\widehat{\theta}_2$)
WLS(σ^2)	0.597	0.797	0.208	0.251
WLS-ADE	0.597	0.791	0.228	0.301
WLS-Carrol	0.591	0.804	0.267	0.365
OLS	0.589	0.790	0.311	0.385

Table 2.2: Design 2 - Results from ADE

$\widehat{\zeta}$	-0.739	0.557
std($\widehat{\zeta}$)	(0.208)	(0.318)

Due to its flexibility, the ADE approach seems to produce good estimates of the variance of the error term of the linear model and these estimates are better than those given from Carrol's method if the first moment is used, in this case, the mean of the estimated σ^2 in 1000 Monte Carlo replications [see Figures 1A and 1C and Figures 2A and 2C, for the designs 1 and 2, respectively]. The ADE of ζ appears slightly biased as can be seen from Table 3.1 below. This is probably a reflex of the sample size that is relatively small. However, this bias does not have an important effect in the bias of $\widehat{\sigma}_i^2$.

Table 3.1: Proportionate bias of $\widehat{\zeta}$, $\left[Pb(\widehat{\zeta}) = E(\widehat{\zeta})/\zeta - 1\right]$

	$Pb(\widehat{\zeta}_1)$	$Pb(\widehat{\zeta}_2)$
Design1	-0.0113	-0.0850
Design 2	-0.0763	-0.0717

Figure 1A: Error Variance Estimation (ADE and Nonparametric)
$E(\varepsilon^2|x)=Exp(x\zeta)$

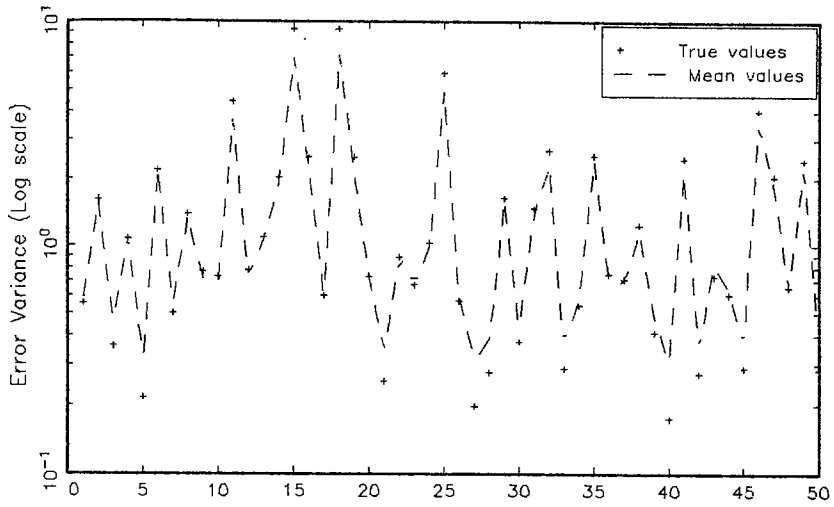


Figure 1B: QQ Plot – $\hat{\beta}_1$
Weight Least Squares – Design 1

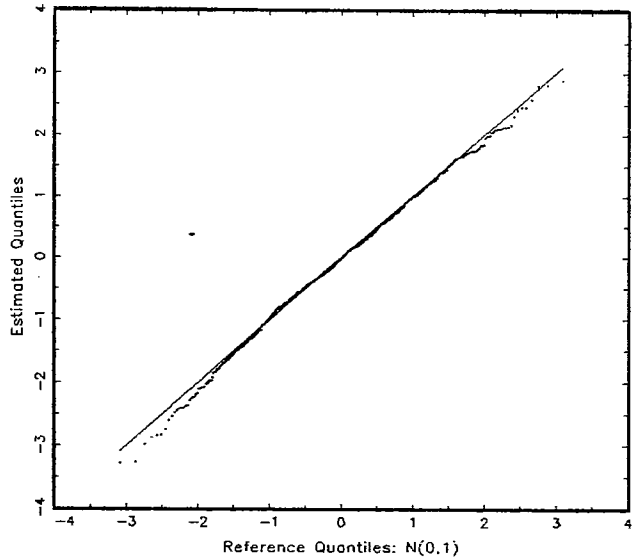


Figure 1C: QQ Plot – $\hat{\beta}_2$
Weight Least Squares – Design 1

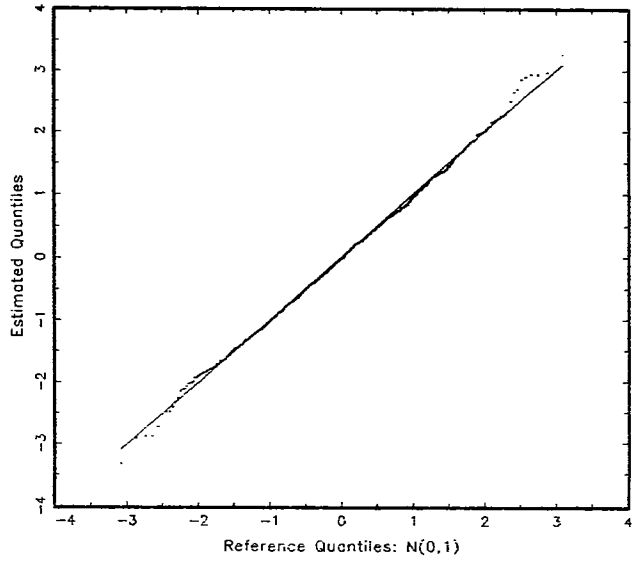


Figure 1D: Error Variance Estimation – Carrol (1982)
$E(\varepsilon^2|x)=Exp(x'\zeta)$

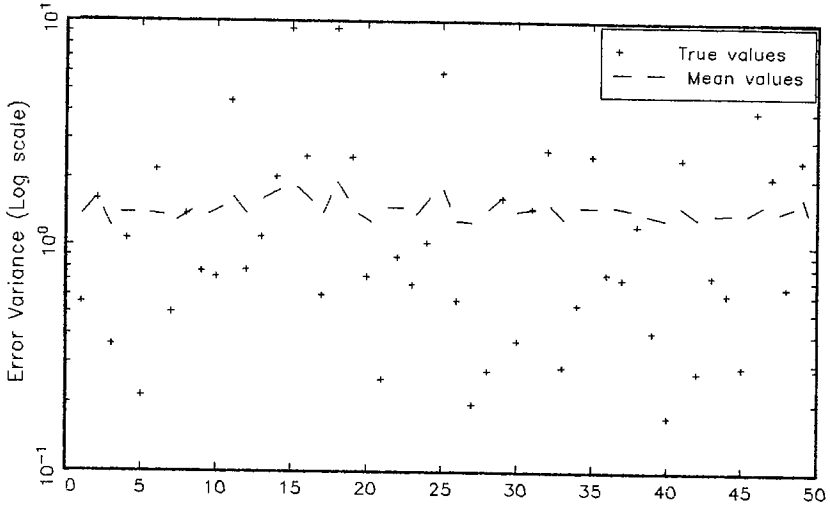


Figure 1E: QQ Plot – $\hat{\beta}_1$
Weight Least Squares – Design 1

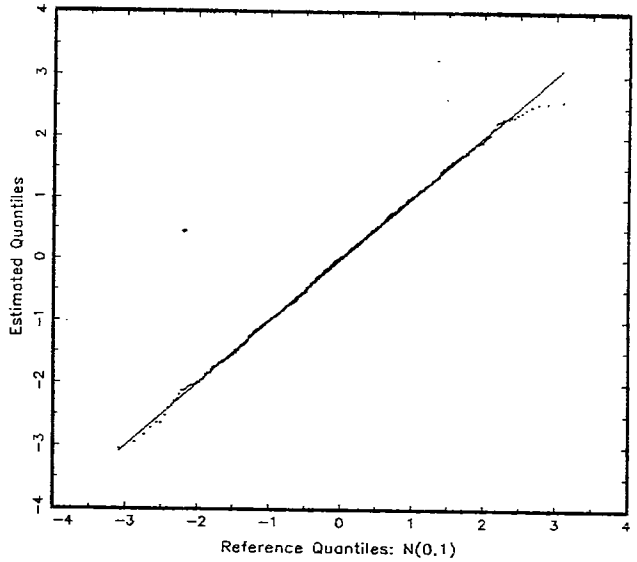


Figure 1F: QQ Plot – $\hat{\beta}_2$
Weight Least Squares – Design 1

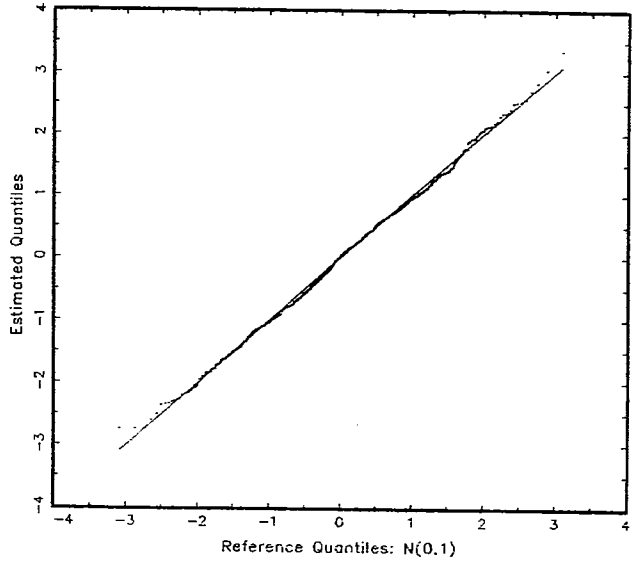


Figure 2A: Error Variance Estimation (ADE and Nonparametric)
$E(\varepsilon^2|x)=Exp(x\zeta)$

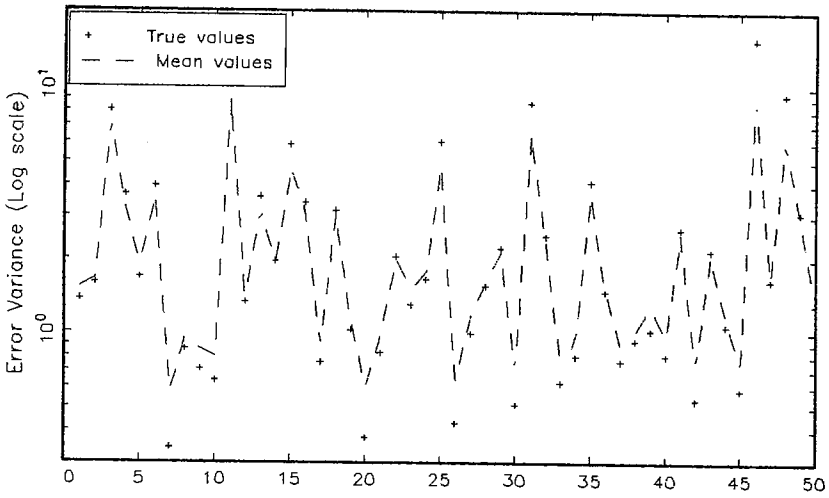


Figure 2B: QQ Plot – $\hat{\beta}_1$
Weight Least Squares – Design 2

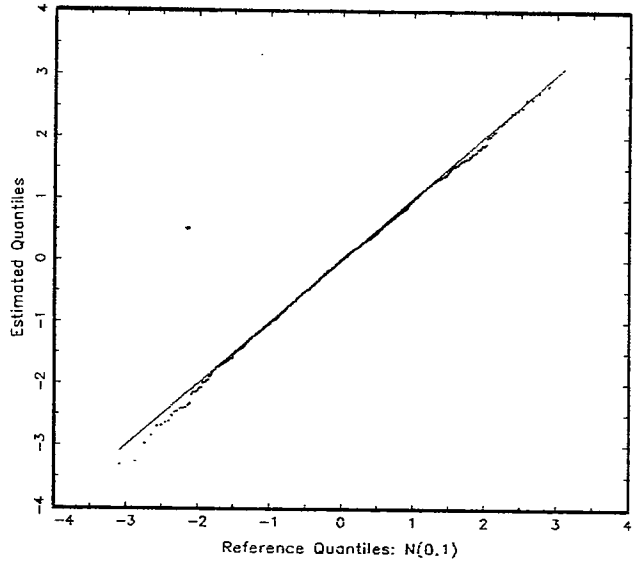


Figure 2C: QQ Plot – $\hat{\beta}_2$
Weight Least Squares – Design 2

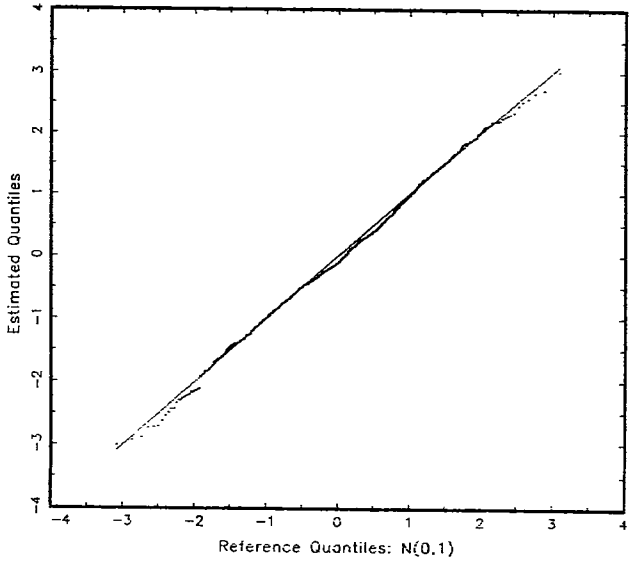


Figure 2D: Error Variance Estimation – Carrol (1982)
$E(\varepsilon^2|x)=Exp(x\zeta)$

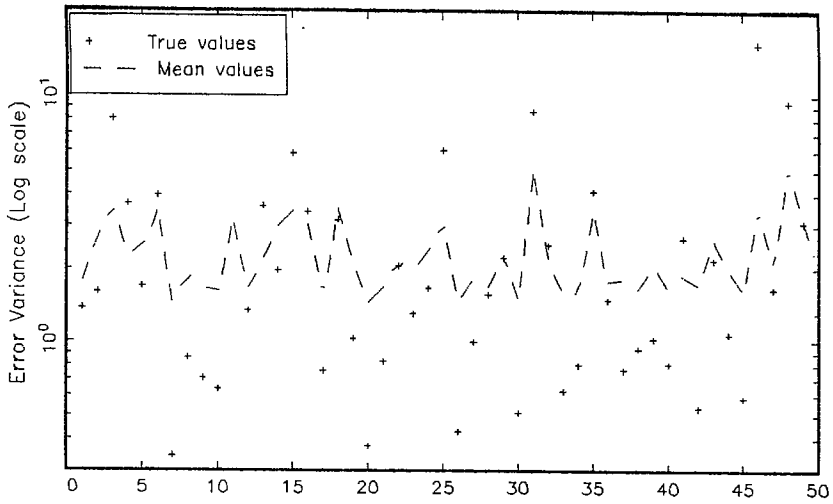


Figure 2E: QQ Plot – $\hat{\beta}_2$
Weight Least Squares – Design 2

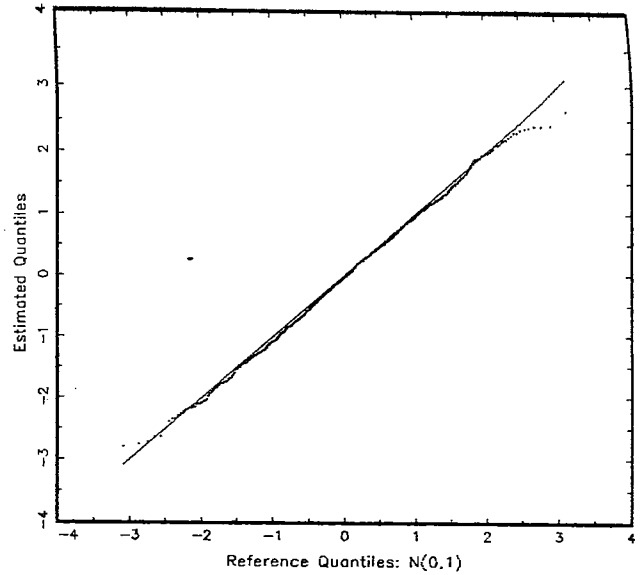
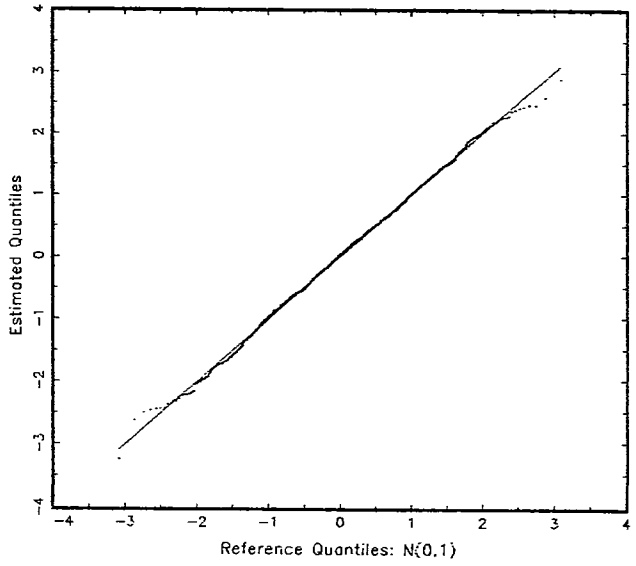


Figure 2F: QQ Plot – $\hat{\beta}_3$
Weight Least Squares – Design 2



5. CONCLUDING REMARKS

The WNLS estimator, $\widehat{\theta}$, presented in this chapter can be viewed as a generalization of Carrol's (1982) estimator where the variance is modelled as a function of the mean. The estimator is $\sqrt{n}$-consistent and asymptotic normal and if the conditional variance assumption is correct it attains (2.4) asymptotically. Furthermore, because $\widehat{\zeta}$ converge at the same rate as $\widehat{\theta}_{LS}$ it might be expected that the finite-sample performance of $\widehat{\theta}$ and Carrol's estimator are similar.

As can be seen from Section 2, the use of ADE provides another advantage if compared with Carrol's estimator. It provides a natural test for heteroskedasticity, simply by assessing the statistical significance of the ADE and it allow us a straightforward identification of the covariates responsible for the presence of heteroskedasticity.

The Wald and score type based tests for the detection of heteroskedasticity are shwon to be irrespective of the shape of $m(\cdot)$. To avoid incorrect standard errors a specification test for the variance assumption is also presented.

6. REFERENCES

Andrews, Donald W.K. (1994). Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity, *Econometrica*, 62, 43-72.

Andrews, Donald W.K. (1991). An Empirical Process Central Limit Theorem for Dependent Non-identically Distributed Random Variables, *Journal of Multivariate Analysis*, 38, 187-203.

Breusch, T.S. and Pagan, A.R. (1979). A simple test for Heteroskedasticity and Random Coefficient Variation. *Econometrica*, 47, 1287-94.

Carrol, R.J. (1982). Adapting for Heteroskedasticity in Linear Models. *Annals of Statistics*, 10, 1224-1233.

Chamberlain, (1987). Asymptotic efficiency in estimation with conditional moment restrictions, *Journal of Econometrics*, 34, 305-334.

Chesher, A. and Peters, S. (1994). Symmetry, regression design and sampling distributions, *Econometric Theory*, 10, 116-129.

Gourieroux, C.; Monfort, A. and Trognon, A. (1984). Pseudo-maximum likelihood methods: Theory. *Econometrica*, 52, 681-700.

Greene, W. H. (1993). Econometric Analysis. Macmillan, 2nd edition.

Härdle, Wolfgand (1990). Applied Nonparametric Regression. Cambridge, University Press.

Härdle, Wolfgand and Stoker, Thomas (1989). Investigating Smooth Multiple Rgression by the Method of Average Derivatives, *Journal of the American Statistical Association*, 84, 986-955.

Hidalgo, Javier (1992). Adaptative Estimation in the Time Series Regression Models with Heteroskedasticity of unknown Form. *Econometric Theory*, 8, 161-187.

Jennrich, R. (1969). Asymptotic properties of nonlinear least squares estimators. *Annals of Mathematical Statistics*, 40, 633-643.

Pötscher, B.M. and Prucha, I.R. (1989). A Uniform Law of Large Numbers for Dependent and Heterogeneous Data Processes, *Econometrica*, 57, 675-683.

Powell, L. James (1992). Estimation of Semiparametric Models. Princeton University. Draft.

To estimate θ, some method is required to handle the unknown density, $h(v)$. Two major approaches have been considered in the literature: one uses a parametric specification of the heterogeneity by assuming a known density function, $h(v)$. In this case, the marginal $g(y) = \int g(y, v)dv$ has a known expression and the estimated values of θ can be obtained from maximum likelihood methods [see for example Lancaster (1979)]; the other one uses nonparametric methods of estimation as the nonparametric maximum likelihood estimation (NPML) [see for example Heckman and Singer (1984a, 1984b)] and the maximum penalized likelihood method suggested by Huh and Sickles (1994).

The first one has the disadvantage of being too restrictive. In general, heterogeneity is modelled using a gamma specification of the density function. However there is not any reason to assume such a density other than because of its simplicity of computation and therefore this can easily lead to a misspecified model. Additionally, if the hazard function is fully parameterized, this way of handle the problem is in the origin of over parameterization and consequently to the observational equivalence of two different sets of distributions. This leads to the problem of identifiability and it will not be addressed here [for a good reference on identifiability in the context of duration models see for example Lancaster (1990)].

The second approach, popularized by Heckman and Singer (1984a) in the context of duration models, is based on the NPML estimator of $h(v)$. This estimator comes from the statistical literature on mixture models in particular from the works of Laird (1978) and Lindsay (1983a, 1983b) [see also Lindsay (1995)]. These authors showed that the NPML estimator of $h(v)$ is a finite mixture of m points of support with log-likelihood function

$$\ln L = \sum_{i=1}^{n} \ln \sum_{j=1}^{m} f(y_i \mid v_j) P_j \tag{1.1}$$

where m is estimated along with P and θ (if $y = y(\theta, x)$) and the P's are mass points associated with the v's. The algorithm usually used to estimate this mixture model is the EM algorithm. In the context of a duration model this algorithm is described, for example, in Lancaster (1990).

Heckman and Singer (1984b) showed, from the examples considered in their paper, that the application of the NPML estimator does not produce an adequate estimate of the underlying heterogeneity distribution. The problem is that if the true distribution is continuous the NPML estimates provides in general few points of positive mass[1]. Additionally, the shape generated by these points are in general not smoothed.

Recently, Huh and Sickles (1994) gave some contribution towards the lack of smoothness through Maximum Penalized Likelihood Estimation (MPLE)[2]. Basically, they estimate the parameters of a duration model by using a smoothed version of the joint density, $g(y, v)$, where the log-likelihood function of the conditional density is

$$\ln L = \sum_{i=1}^{n} \ln g(y_i, v_i) - \sum_{i=1}^{n} \ln h(v_i) \tag{1.2}$$

To estimate the parameters of this model, prior knowledge on $h(v)$ is necessary. To handle this
unknown quantity, they suggested the log penalized likelihood function,

$$\ln L = \sum_{i=1}^{n} \ln g(y_i, v_i) - \lambda \Omega(g) \tag{1.3}$$

where λ is the smoothness parameter and $\Omega(g)$ is a measure of roughness. In this case the effects of heterogeneity are being ignored where the penalty term smooth out the misspecified density of duration time. They suggested and followed another variant

[1]The method determines the points of support and they can not be controlled externally.

[2]This method was introduced by Good and Gaskins (1971) and developed by Montricher, Tapia and Thompson (1975).

of this by using a mass point method to integrate out the density, the first term in the right hand side of (1.3). This approach is a smoothed version of the method suggested by Heckman and Singer (1984b). For the penalized term they consider the squared norm of the second derivative of the hazard function with respect to the covariates.

Both of these approaches share a common problem. The NPML estimator is based on the estimation of the heterogeneity mass points but they have the problem of a lack of smoothness. Moreover, as shown from these authors, the practice dictates that the mass point method does not provides enough points of support that allows an easily identification of the density of the unobservables, particularly when the true underlying density is continuous. The MPLE applies smoothed methods, solving the first problem but not the second one.

In this chapter it is suggested an alternative method based on roughness penalties, as in Huh and Sickles (1994), but using a different methodology in order to achieve better estimates of the density of the unobservables. The purpose of this chapter is the estimation of density of the unobservables more than the estimation of the structural parameters. Some of the reasons for doing this are as follows: the use of the estimated density as a reference to access misspecifications of the model in a parametric specification of heterogeneity; to identify multimodality in the density of the heterogeneity. In this case, if the heterogeneity is due to differences among individuals, this method can be useful in the identification of groups of individuals with similar patterns.

This paper is organized as follows. Section 2 starts with the presentation of the model and the estimator in a general context of a mixture model. These results are then generalized to a regression model with an application to a duration model. In this case it is assumed that the implied duration distribution is in the Weibull family. Section 3 presents some examples to evaluate the performance of the method in prac-

tice. Section 4 concludes this chapter with some additional remarks and conclusions.

2. MODEL AND ESTIMATOR

Consider that $y_1, y_2, \ldots, y_n$ are realizations of a random variable Y with density $g(y)$. It is assumed that the density $g(y)$ is unknown but satisfies the equation

$$g(y) = \int f(y \mid v)\ h(v)\ dv \tag{2.1}$$

where $f(y \mid v)$ is a known density, $h(v)$ an unknown density function and $v \in V$. The problem that I address in this chapter is the estimation of the mixing density function, $h(v)$. Results for the existence of solutions of (2.1) can be viewed, for example, in Wahba (1990).

Methods of estimating $h(v)$ can be related to the structure of $f(y \mid v)$. A special and very well known case is derived if $f(y \mid v) = f(y - v)$. This is a convolution problem and finds application in econometrics when the covariates are measured with error. In this particular case, estimation methods are available in the literature if the distribution of $U = Y - V$ is known [for an overview on deconvolution see for example Carrol et al (1995), Chapter 12].

In the general case, the solution to this problem by inversion of (2.1) is not new in the statistical and mathematical literature. This problem also appears designated as Fredholm Integral Equation of the first kind in the mathematical literature or as mixture model in the statistical literature. It is also known to be an *ill-posed* problem [see for example Tikhonov, et al. (1977)] in the sense that small perturbations in $g(y)$ can result in large fluctuations in the solution, $h(v)$. As an example let $h(v)$

be the solution of (2.1) and add to it the function $h_m(v) = \sin(m\ v)$. For any integrable function $f(y \mid v)$ it is well known that $g_m(v) = \int f(y \mid v)\ h_m(v)\ dv \to 0$ as $m \to \infty$. Therefore only an infinitesimal change g_m in g causes a finite change h_m in h.

To stabilize the solution of (2.1) the regularization method have been proposed [for an overview see for example Tikhonov and Arsenin (1977)][3].

Using some method of discretization, consider (2.1) rewritten as a linear operator, operating on the vector $h \equiv h(v)$ of order $m \times 1$,

$$g_n = A\ h \tag{2.2}$$

where g_n of order $n \times 1$ is some nonparametric estimate of $g \equiv g(y)$ (or g itself if known) and A is a $n \times m$ matrix with elements $f(y_i \mid v_j)$, $i = 1, \ldots, n$ and $j = 1, \ldots, m$. It is assumed that $g_n = g + \varepsilon_n$ where ε_n is the error incurred in approximating g by g_n and $\varepsilon_n \to 0$ as $n \to \infty$. As a first approach, if the same number of y's and v's is recorded, $m = n$, the solution to this problem can be viewed as an inversion problem or if $m < n$ as a regression problem. However, this approach seems to be less straightforward than it appears because the densities in (2.1) are assumed continuous. One of the problems is that as the grid becomes finer the rows and/or columns of A become closer and the calculation will become unstable due to the singularity of A. The degree of this closeness depends on the variability of $f(y \mid v)$ with less variable functions meaning more unstable situations. Additionally, even if this is not reason for concern, the solution of the inversion problem is not in general smoothed.

One possible solution to the first problem can be attained through some appro-

[3]Tikhonov define regularization method as the method of constructing approximate solutions, in a mathematical point of view.

priated discretization method. One possible rule can be the consideration of a grid indexed to the variability of $f(y \mid v)$, with a finer grid in areas were the variability is high. To the second problem it is usual to add some penalty term that accounts for roughness.

For general functions, not necessary density functions, with $g(.)$ known, Phillips (1962) suggested an approximate solution, solving (2.1) at discrete points $y_i = i/n$, $i = 1, 2, \ldots, n$ by finding h at points $v = (1/n, \ldots, n/n)$ in order to minimize the quantity

$$R^*(h_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[g(\frac{i}{n}) - \frac{1}{n}\sum_{j=1}^{n} f(\frac{i}{n} \mid \frac{j}{n})\ h(\frac{j}{n})\right]^2 + \lambda \sum_{j=2}^{n\ 1}\left[h(\frac{j+1}{n}) - 2h(\frac{j}{n}) + h(\frac{j-1}{n})\right]^2$$

where the second term in the right hand side is a penalty term for roughness and λ controls the smoothness of the solution h_λ.

However, for the problem considered in this paper, some additional restrictions are required. Basically, it is necessary to impose that $h(v) \geq 0$ for all $v \in V$ and $\int_V h(v) = 1$. Moreover it is of interest not to restrict the domain of y and v to the $[0, 1]$ interval.

With these remarks in mind and using the notation (2.2) the idea is to find a solution h_λ that minimizes

$$R^c(h_\lambda) = \|g_n - A\ h_\lambda\| + \lambda\Omega(h_\lambda) \tag{2.3}$$

subject to

$$h(v) \geq 0\ , \qquad \int_V h(v) = 1$$

where $\Omega(h)$ is a measure of roughness. As a measure of roughness it is usual to consider the second derivative [see for example Silverman (1986)]

$$\Omega(h) = \int [h''(v)]^2 dv. \tag{2.4}$$

Other approaches to the problem include the work of Mendelsohn and Rice (1982). These authors suggested the use of B-splines[4] in the representation of h [see also Wahba (1990)]. In this case, the idea is to choose $\widehat{h}$ to minimize $\|g_n - A\ \widehat{h}\|$ where $\widehat{h}$ is represented as a linear combination of B-splines with fixed knot locations, $\widehat{h}(v) = \sum_{j=1}^{p} \beta_j B_j(v)$ and thus $(A\widehat{h})(v) = \sum_{j=1}^{p} \beta_j (AB_j)(v)$ where B_j is the j^{th} B-spline of degree $k-1$ corresponding to the knot sequence $\tau_1, \ldots, \tau_{p+k}$. The coefficients β_j are computed from a least squares fit of $A\ \widehat{h}$ to g_n and the integral $(AB_j)(v)$ is numerically evaluated using Simpson's rule. The instability of solving (2.1) is controlled by p that is the analog of λ in the roughness penalty approach. For $\widehat{h}$ to be a probability density additional constraints are imposed. Because $\int B_j(v)dv = (\tau_{j+k} - \tau_k)/k$ and $B_j(v) > 0$ the problem becomes $\min_\beta \|g_n - A\ \widehat{h}\|$ subject to $\beta \geq 0$ and $c'\beta = 1$, where $c_j = (\tau_{j+k} - \tau_j)/k$, $j = 1, 2, \ldots, p$.

In the roughness penalty approach, $(A\widehat{h})(v)$ is computed using summation rather than numerical integration (through its B-spline representation). To deal with it some discretization method is needed. One possible solution to overcome the problem of the unknown v's can be based on the knowledge of $f(y \mid v)$, by looking at v as if they were parameters. Given the data $y_1, \ldots, y_n$ and solving the first order condition

[4]B-splines (B stands for basis) can be defined in terms of truncated power functions and divided difference operators. As an example, the B-spline of degree 1 for the knots $t_i, \ldots, t_{i+2}$ is a tent function on $[t_i, t_{i+2}]$, defined as

$$B_i(x) = \frac{(t_{i+2} - x)_+ - (t_{i+1} - x)_+}{t_{i+2} - t_{i+1}} - \frac{(t_{i+1} - x)_+ - (t_i - x)_+}{t_{i+1} - t_i}$$

where $(u)_+ = u$ if $u \geq 0$ and 0 otherwise.

$\partial \ln f(y \mid v)/\partial v = 0$ in v it is possible to write the maximum and minimum values of v as a function of the maximum and minimum of y. For example, if $f(y \mid v) = v\exp\{-y\,v\}$ one has, from the first order condition, $v = 1/y$ and $\min(v) = 1/\max(y)$ and $\max(v) = 1/\min(y)$. In another way, the range of v is setting over which $f(y \mid v)$ should have its support. Knowing these values we can define a mesh of m equally spaced points[5] between $\min(v)$ and $\max(v)$. Therefore, defining $A_{ij} = f(y_i \mid v_j)$, $g_{ni} = g_n(y_i)$, $h_j = h(v_j)$ where $v_j \in (\min(v), \max(v))$, $\Delta^2 h_j = h_{j+1} - 2h_j + h_{j-1}$ and $\Delta v_j = v_j - v_{j-1}$ the minimization problem (2.3) becomes,

$$\min_{h_1,\ldots,h_m} R^c(h_\lambda) = \frac{1}{n}\sum_{i-1}^{n}\left(g_{ni} - \frac{1}{m}\sum_{j-1}^{m} A_{ij}\; h_j\right)^2 + \lambda \sum_{j-2}^{m\;\;1}\left(\Delta^2 h_j\right)^2 \qquad (2.5)$$

subject to

$$\begin{cases} \sum_{j=1}^{m} \Delta v_j h_j = 1 \\ \quad h_j \geq 0. \end{cases}$$

This is a quadratic optimization problem and methods of solution are available in the literature.

The consistency of the method being presented can be based in the results of Mendelsohn and Rice (1982). They prove the consistency of the method in a related problem where $h(v)$ is a B-spline. However, their proof is not dependent on a specific form assumed by $h(v)$ and therefore the same results apply here with $m = m(n) \to \infty$ and $\lambda = \lambda(n) \to 0$ as $n \to \infty$. Under the assumptions presented by Mendelsohn and Rice, $\|g - Ah_n\| \to 0$ a.s. and this implies[6] $\|h_o - h_n\| \to 0$ a.s., where for each n, h_n is the unique function minimizing $\|g_n - Ah_n\|$ and h_o is the true density, i.e., $Ah_o = g$.

[5]The method provides a range for m. The number of m remains an open question.

[6]As pointed out by Mendelsohn and Rice (1982), what makes this implication true is the fact that h_o, $\widehat{h}$ and g are probability densities.

No results on rates of convergence are available.

This method has however the inconvenience of the presence of the constraints conditions. It is possible to introduce some simplification through some appropriate transformation of the parameter space to eliminate these constraints. Instead of estimating h one can estimate a function of it, $P = P(h)$, with generic element

$$P_j = \frac{\exp(h_j)}{\sum_{j=1}^{m} \exp(h_j)} \quad , \quad j = 1, 2, \ldots, m. \tag{2.6}$$

Considering this transformation and using matrix notation, the unconstrained version of (2.5) can be written[7] as

$$R(h_\lambda) = (g_n - AP)'(g_n - AP) + \lambda P'W'WP \tag{2.7}$$

where W is a band matrix of order $(m-2) \times m$ that summarises the second difference operator, defined as

$$W = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix}$$

The parameter λ makes the trade-off between goodness of fit and roughness and its value can be given by the subjective choice method [see Bartoszynski, et al. (1981)] or estimated by minimizing the cross validation function [see for example Green and Silverman (1994)]

[7] Another possibility is to penalize h instead of P. However it is not clear, without further research, which one produces better results.

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(g_{ni} - A_i P(\widehat{h}_{(i)})\right)^2$$

where $P(\widehat{h}_{(i)})$ is the $m \times 1$ vector of estimated parameters computed without observation i. Since this is a non-linear problem and to avoid the computation of $\widehat{h}_{(i)}$ for each $i = 1, \ldots, n$ it is suggested the use of approximative methods where

$$\widehat{h}_{(i)} \approx \widehat{h} + \left(\frac{\partial^2 R(\widehat{h})}{\partial h \partial h'}\right)^{-1}_{(i)} \left(\frac{\partial R(\widehat{h})}{\partial h}\right)_{(i)}.$$

The indice (i) in the last two terms of the right hand side of the above expression means the inverse hessian and the score computed without observation i [a similar procedure can be viewed in Wahba (1990)].

Following an approach as in Chesher (1996), the estimated variance of the estimator presented, $\widehat{P} = P(\widehat{h})$, is based on a first order approximation of the score $\partial R(\widehat{h}_\lambda)/\partial h_\lambda$. This suggests a *sandwich estimator* of the covariance matrix [see Efron and Tibshirani (1993), page 310][8],

$$\widehat{Cov(P(\widehat{h}))} = \frac{\partial P(\widehat{h})}{\partial h}\left\{\left(\frac{\partial^2 R(\widehat{h})}{\partial h \partial h'}\right)^{-1}\left(\sum_{i=1}^{n}\frac{\partial R_i(\widehat{h})}{\partial h}\frac{\partial R_i(\widehat{h})}{\partial h}'\right)_{\lambda=0}\left(\frac{\partial^2 R(\widehat{h})}{\partial h \partial h'}\right)^{-1}\right\}\frac{\partial P(\widehat{h})}{\partial h}. \tag{2.8}$$

The extension to the regression case is straightforward and can be made by considering the observations given as $y_i = y(x_i, \theta)$ where x_i is a vector of covariates and θ a vector of parameters. This case introduces the complication of the presence of the unknown θ that should be estimated altogether with P. The following steps in the estimation process are suggested:

[8]Note that the parameter of interest is not the vector $\widehat{h}$ but the vectorial function $P(\widehat{h})$.

Step 1: given θ determine y;

Step 2: estimate $g(y)$ using some nonparametric method;

Step 3: estimate P as the minimizer of (2.5) or (2.7);

Step 4: given $\widehat{P}$ estimate θ by parametric maximum likelihood

$$l_i(\theta) = \sum_{j=1}^{m} f(y_i \mid v_j)\widehat{P}_j \tag{2.9}$$

Step 5: from $\widehat{\theta}$ form a new y and return to step 2.

As an example, let the spell duration, T, be a non-negative random variable with conditional hazard function defined as

$$\theta(t \mid x, v) = \eta t^{\eta-1} \exp\{x'\beta\} v\,, \qquad v \geq 0 \tag{2.10}$$

where t is the observed duration, x is a $k \times 1$ vector of observable covariates, v is a random variable that summarizes the unobserved characteristics, η is a scalar and β a $k \times 1$ vector of parameters. From these results, the density of spell duration is given as

$$f(t \mid x, v) = \eta t^{\eta-1} \exp\{x'\beta\} v \exp\{-t^{\eta} \exp\{x'\beta\} v\}. \tag{2.11}$$

Considering $y = t^{\eta} \exp\{x'\beta\}$ the above expression simplifies to

$$f(y \mid v) = v \exp\{-y\, v\} \tag{2.12}$$

and steps 1-5 can be applied to estimate η, β and P.

Censoring is also easily introduced in this model. If y_i represents the time of a completed spell, censored observations are given by $Y_i = \min(y_i, y_c)$ and $d_i = I(y_i < y_c)$, where y_c is the censored time of an incomplete spell and I is an indicator function with $d_i = 1$ if $y_i < y_c$. Assuming these results the conditional density function of y is

$$f(y_i \mid v) = \begin{cases} v_i \exp\{-y_i v\} & , \quad d_i = 1 \\ \\ \exp\{-y_i v\} & , \quad d_i = 0 \end{cases}$$

The range of v values can be computed using proposition 11 of Heckman and Singer (1984a). For uncensored observations the first order condition $\partial f(y_i \mid v)/\partial v = 0$ leads to $v = 1/y_i$ and $v \in [1/y_{\max}, 1/y_{\min}]$. For censored observations $v \in [0, 1/y_{\min}]$.

3. APPLICATION

This section presents some illustrative examples with two cases: model without covariates and model with covariates. In all cases it is assumed that the true density of heterogeneity is of gamma type,

$$h(v) = \frac{\alpha^\gamma}{\Gamma(\gamma)} \exp\{-\alpha v\}\, v^{\gamma-1} \tag{3.1}$$

and the density of spell duration is defined as in (2.12). From these assumptions, the true marginal distribution of spell duration[9], $F(y) = \int F(y \mid v)\, h(v)\, dv$, has the following expression,

$$F(y) = 1 - \left(1 + \frac{y}{\alpha}\right)^{-\gamma}.$$

[9]Using distribution functions, the marginal $F(y) = \int F(y \mid v) dH(v)$ can be easily estimated using histogram methods. This avoid the boundary problems resulting from the application of non-parametric kernel density estimation.

In practice the data y is known but not its distribution. For its estimation one can apply a large variety of non-parametric methods. In the examples that follows I considered

$$\widehat{F(y_i)} = \frac{1}{n}\#I(y < y_i) \ , \quad i = 1, 2, \ldots, n$$

where $\#I(y < y_i)$ is a function representing the number of y's less then y_i. In all cases it is considered $n = 100$ observations.

Case 1: Model without covariates and $(\alpha, \gamma) = (2, 2)$ - Figures 1.1 to 1.4 represent the true and estimated density of heterogeneity for different values of λ. The estimated density was computed using the unconstrained problem (2.7). The range of the v's values were computed as shown in section 2. Knowing the maximum and minimum values of v, $m = 40$ equally spaced points are used with a step of 0.15. Note that m and the step are chosen in order to avoid the presence of extreme points in the range of v. The reason is that for y near zero $v = 1/y$ is very high and these high values of v provide no additional information to the estimated density.

The method of cross validation, traditionally used to computed the optimal value of λ does not work very well in the this case. The problem is probably associated to the use of the first order approximation in the computation of $\widehat{h}_{(i)}$. These results are summarized in Figure 1.5. A very fine grid of λ values were chosen to show the problems associated with the application of this method.

Standard errors of $\widehat{P}$ shown in Figures 1.2 to 1.4 are plotted as upper bounds of pointwise 95% intervals around $\widehat{P}$, $\widehat{P} + 1.96\ se(\widehat{P})$, to give an idea of the variability of $\widehat{P}$. The standard errors were computed from expression (2.8).

Case 2: Model with one covariate and $(\alpha, \gamma) = (1, 3)$ - This case is summarized through Figures 2.1 to 2.4 and the estimated densities were computed using the constraint problem (2.5). Apart from the criterion function used, this case is rather similar to case 1 with the difference that now the data, y, depends on some unknown parameters that have to be estimated as seen in section 2. Considering the hazard function (2.10), this corresponds to case 1 with $y = t^{\eta}\exp\{x'\beta\}$. Moreover, it is assumed that $x_i = [1 \quad x_{1i}]$ and $x_{1i} \sim N(0,1)$, $i = 1, \ldots, 100$. The true parameters of the model are $\beta_0 = \beta_1 = \eta = 1$. For the v's values it is considered $m = 60$ equally spaced points with a step of 0.15.

Table1: Case2 - Estimated Parameters as a function of λ

λ	$\widehat{\eta}$	$\widehat{\beta}_1$	se$(\widehat{\eta})$	se$(\widehat{\beta}_1)$
0	1.126	0.945	0.143	0.147
0.00006	1.135	0.965	0.143	0.141
0.6	1.152	0.994	0.145	0.147
100	1.239	1.065	0.152	0.158

The first conclusion that can be extracted from these Figures is that the estimated densities seem to perform relatively well[10] (if compared with the true density) for some appropriate value of λ. Moreover, the shape of the estimated density seems not to change significantly for a wide range of λ values. This is a good property if λ is chosen subjectively.

Another interesting property, already notice by Heckman and Singer (1984a) in the case of NPML estimator, is that in general, one has good estimates of the structural

[10]To corroborate the conclusions extracted from the examples presented, a Monte Carlo simulation is required.

parameters even if the $h(v)$ is poorly estimated. This is also true in this case as can be seen in Table 1 by varying the value of λ.

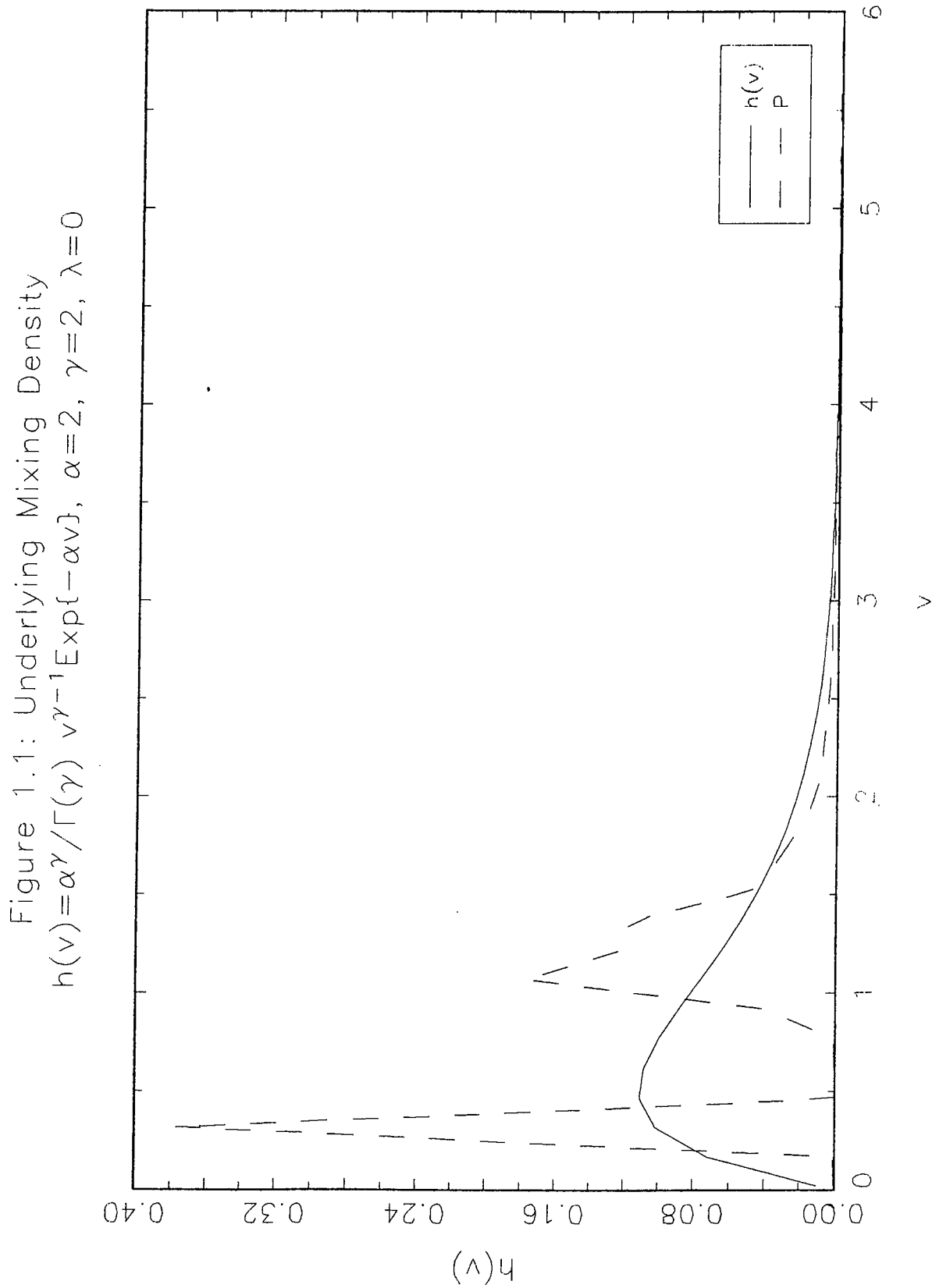


Figure 1.1: Underlying Mixing Density
$h(v)=\alpha^{\gamma}/\Gamma(\gamma)\ v^{\gamma-1}Exp[-\alpha v]$, $\alpha=2$, $\gamma=2$, $\lambda=0$

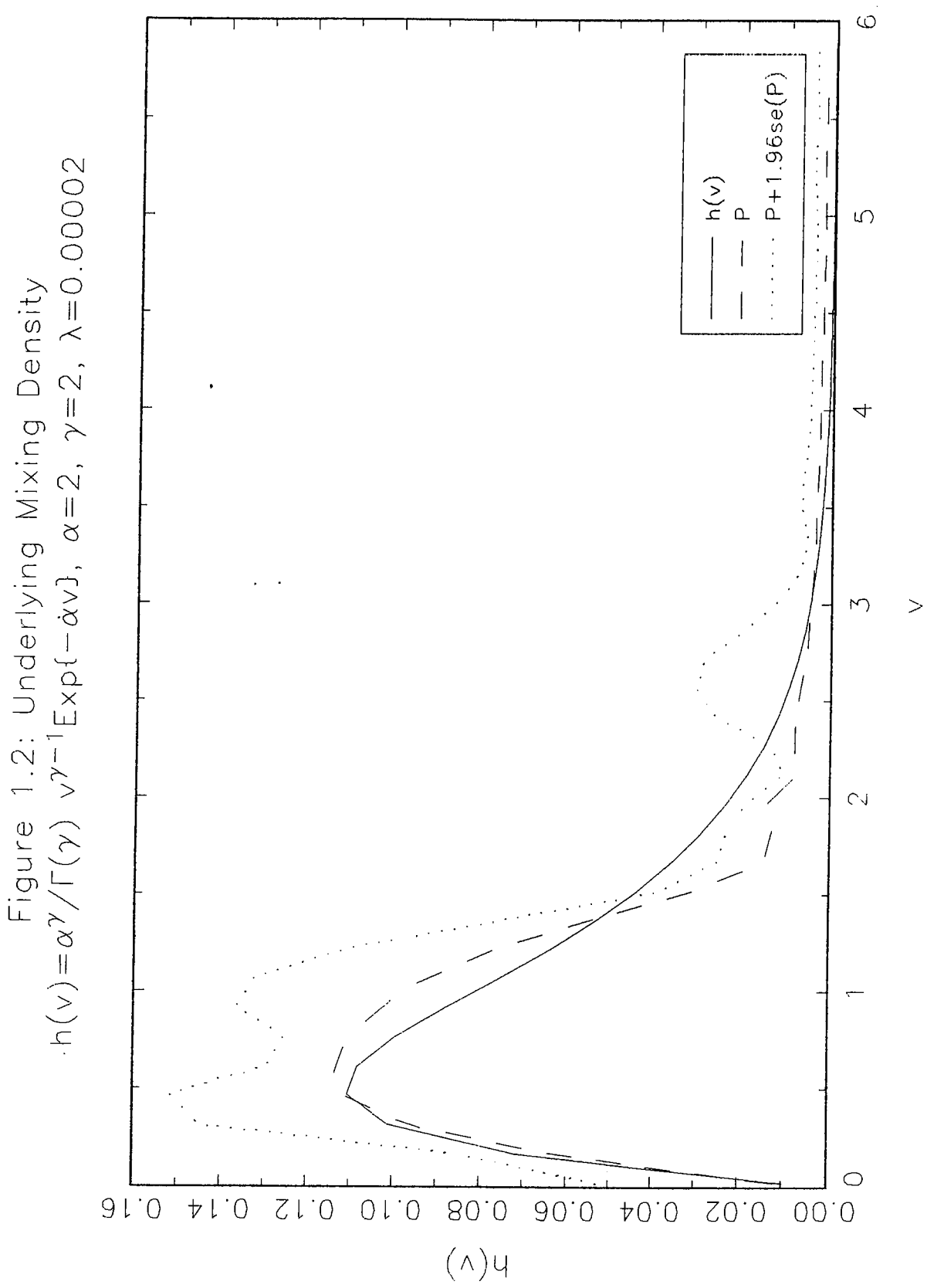
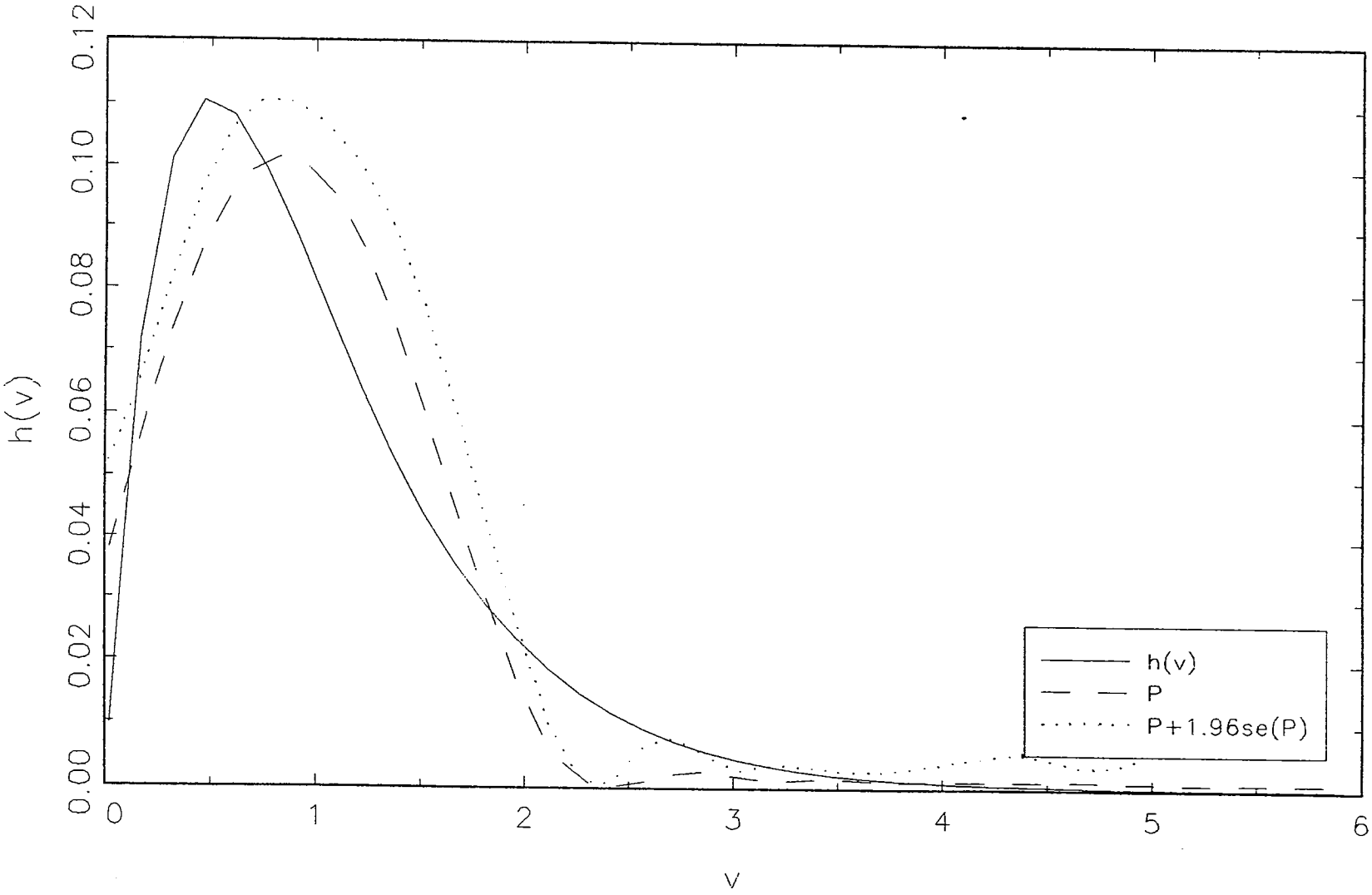


Figure 1.2: Underlying Mixing Density
$h(v)=\alpha^{\gamma}/\Gamma(\gamma)\ v^{\gamma-1}Exp\{-\alpha v\}$, $\alpha=2$, $\gamma=2$, $\lambda=0.00002$

Figure 1.3: Underlying Mixing Density
$h(v)=\alpha^{\gamma}/\Gamma(\gamma)\ v^{\gamma-1}Exp\{-\alpha v\},\ \alpha=2,\ \gamma=2,\ \lambda=0.002$



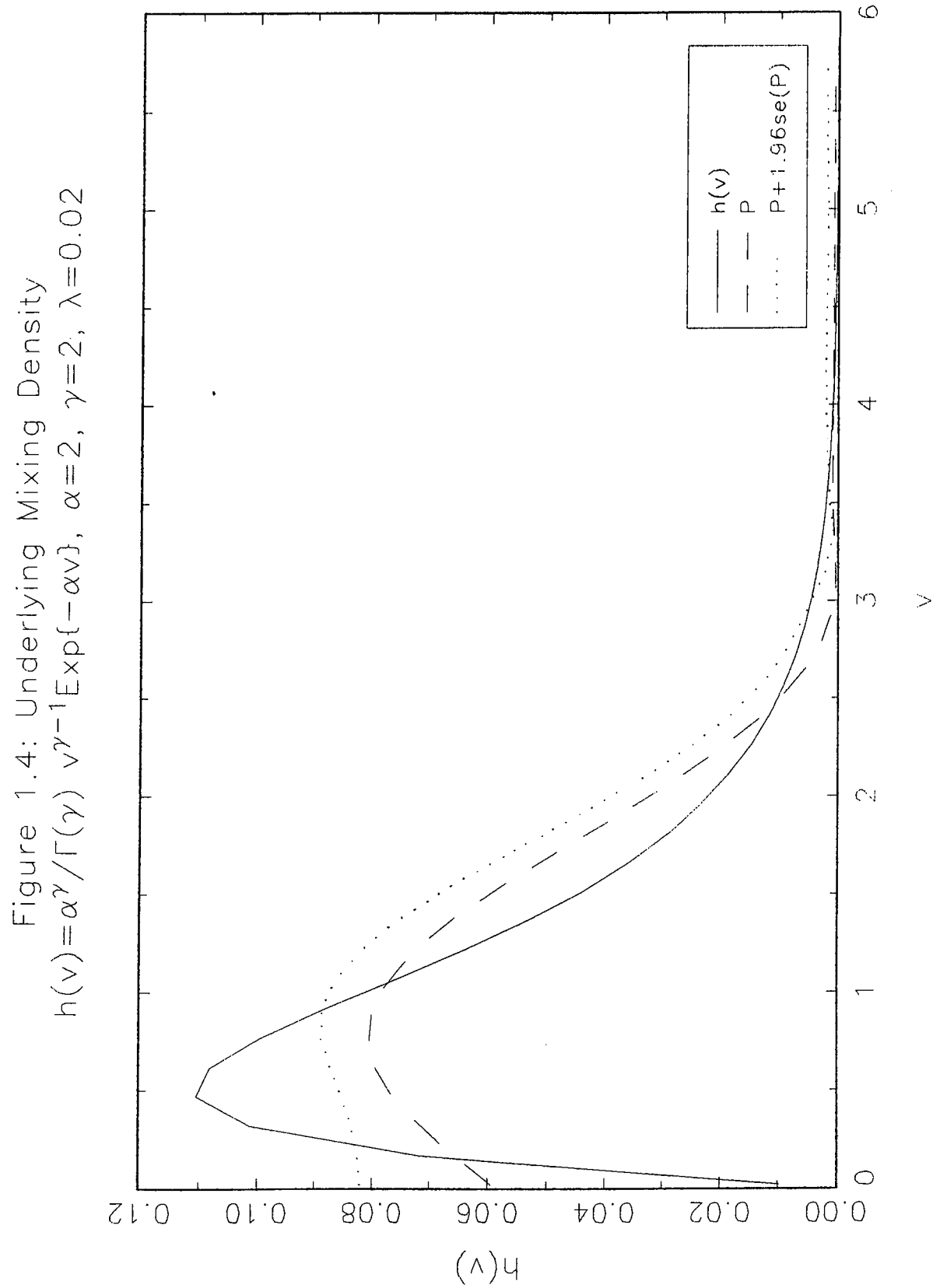


Figure 1.4: Underlying Mixing Density
$h(v)=\alpha^{\gamma}/\Gamma(\gamma)\ v^{\gamma-1}Exp\{-\alpha v\}$, $\alpha=2$, $\gamma=2$, $\lambda=0.02$

Figure 1.5: GCV(λ)

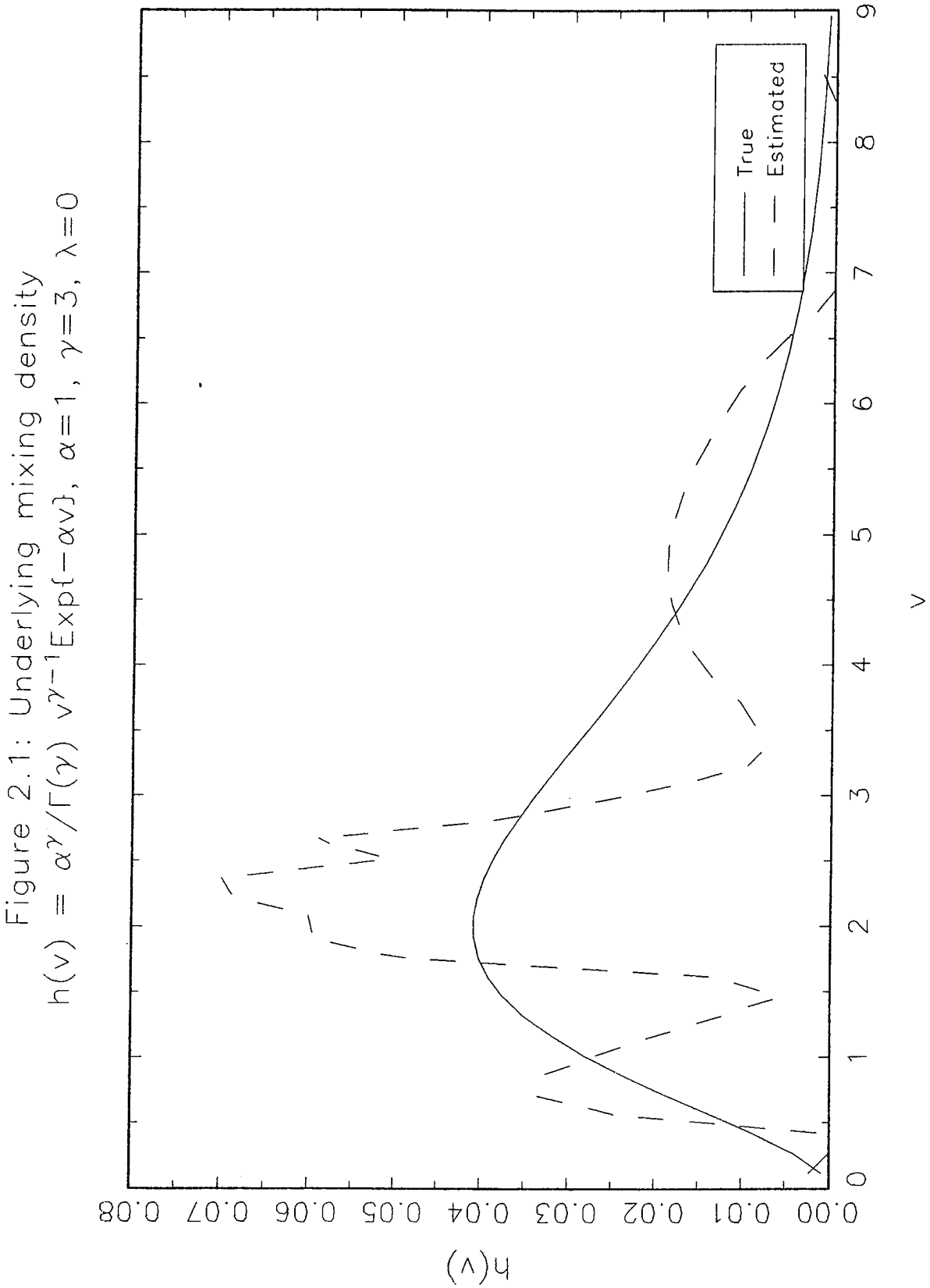


Figure 2.1: Underlying mixing density

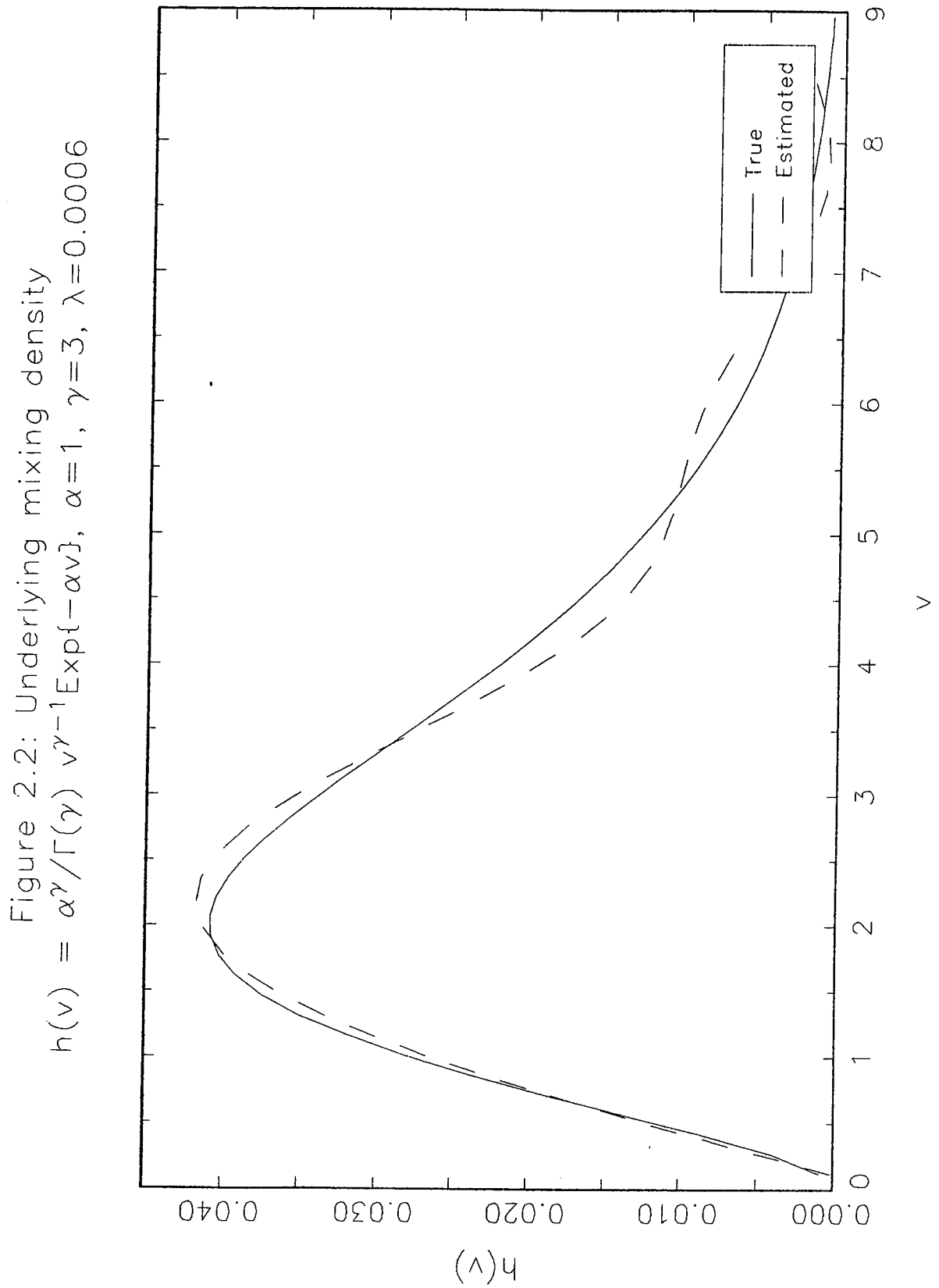


Figure 2.2: Underlying mixing density

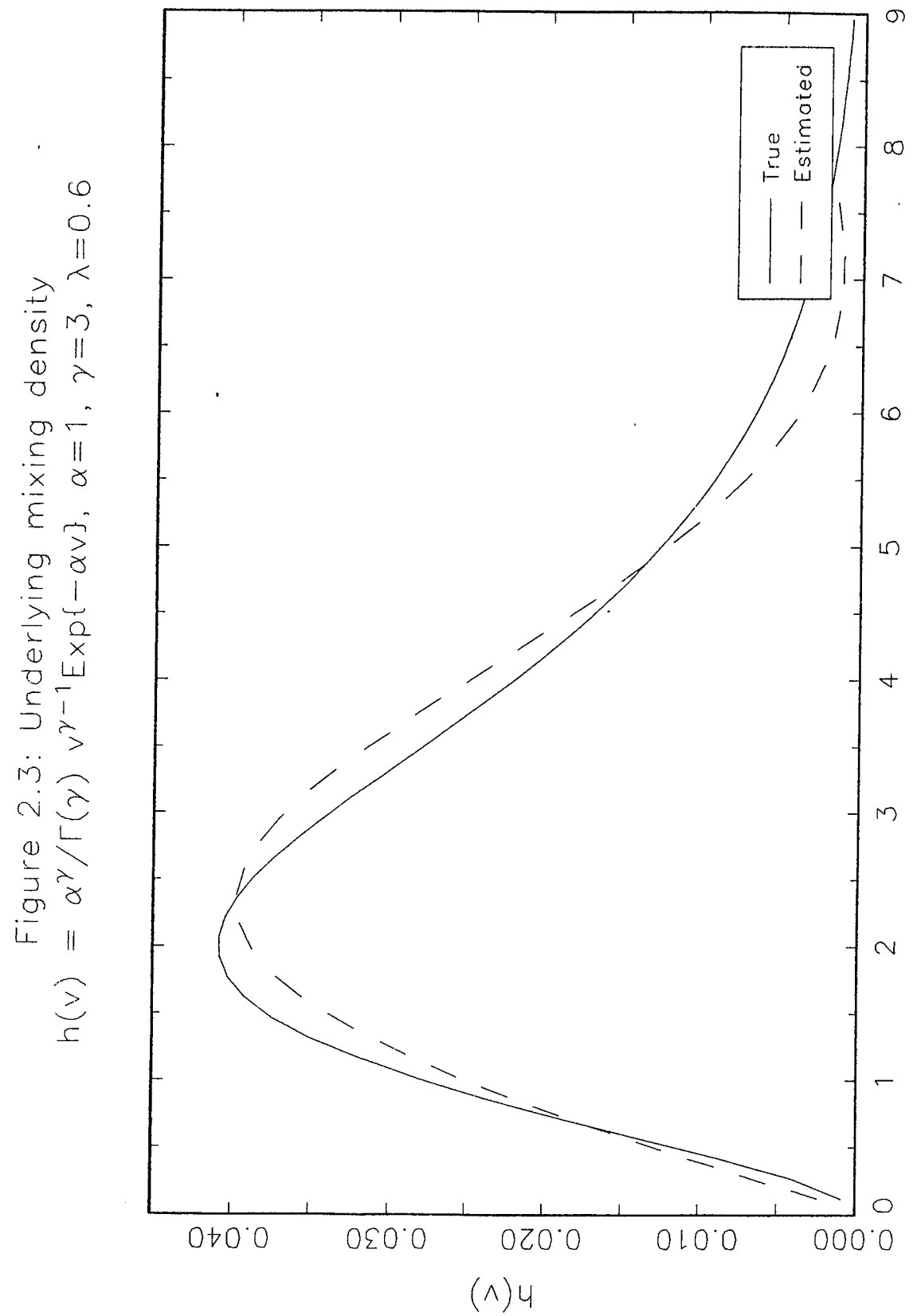


Figure 2.3: Underlying mixing density
$h(v) = \alpha^{\gamma}/\Gamma(\gamma)\ v^{\gamma-1}\mathrm{Exp}\{-\alpha v\}$, $\alpha=1$, $\gamma=3$, $\lambda=0.6$

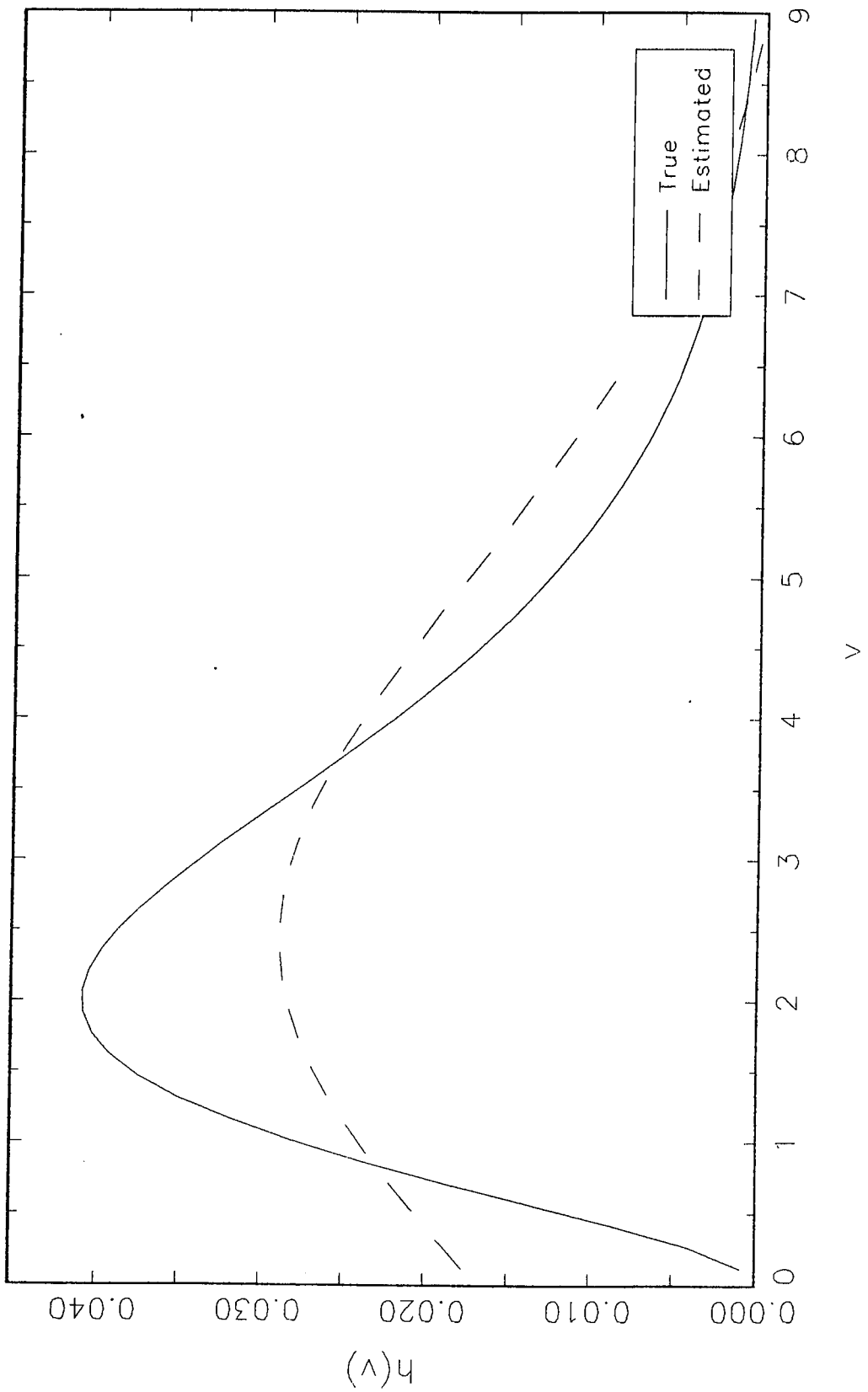


Figure 2.4: Underlying mixing density
$h(v) = \alpha^{\gamma}/\Gamma(\gamma)\ v^{\gamma-1}\mathrm{Exp}\{-\alpha v\}$, $\alpha=1$, $\gamma=3$, $\lambda=100$