INFORMATION TO USERS

UMI Number: 9938941

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

SEPARATE ESTIMATION AND INFERENCE IN ECONOMETRIC MODELS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Helle Bunzel

August 1999

BIOGRAPHICAL SKETCH

Helle Bunzel was born in Ålborg, Denmark. In 1993, she recieved her B.S in mathematical economics from the University of Århus, Denmark. She joined the doctoral program in economics at Cornell University in 1994. There, she received her M.A in economics in 1997. She is expecting to receive her Ph.D. in August 1999. Thereafter, she will be joining the economics faculty at Iowa State University.

ACKNOWLEDGEMENTS

I would not have come this far had it not been for Nick Kiefer and Tim Vogelsang. Nick's optimism and faith in the project and my ability to complete it has been never ending despite my best efforts to prove him wrong. For that I thank him. Tim has always provided down to earth guidance in all aspects of academic life and he has shown immense patience when helping me with specific projects. Having access to Tim has without a doubt made my life at Cornell better. I also thank Mike Conlin, George Jakubson and Yongmiao Hong for their guidance through the years.

I thank Ilaria Ossella and Vajesh Durbal for opening their home to me and providing invaluable friendship at a time when life was not easy for them either, and Gideon Saar for his hours of listening and thoughtful advice. Joydeep Bhattacharya was always there, he suffered with my pains even while he was the best cure and he shared with me the good times, making them unforgettable. I only hope that in the years to come I can manage to repay even a fraction of what I owe him.

Peter Morgan and Tom DiCiccio have both taken the time for long chats and guidance for which I am grateful. I also wish to thank Eduardo, Raji, Kirida, Mrs. Sudan, Yael and all the other people of Cornell and Ithaca who have made my life so much better.

I thank all my friends in Denmark, Anne, Birgitte, Karsten, Mariann, Pia and

Lars for making the effort to stay connected, even when I was so far away.

I gratefully acknowledge financial support from a Sage Fellowship and the Center for Analytic Economics. I also thank Nick Kiefer for two semesters of financial support. I thank Robert Masson and Patricia Paucke for their guidance through the paper jungle of graduate life, and Patricia Paucke and Amy Gowan for their patience and assistance while I was looking for job this past year.

Finally, I thank my parents for always going so far out of their way to support me in every imaginable way, and my grandparents for always being there for me, and I thank my whole family for their patience and understanding when I could not be near them. Without the support of my family, even starting down this road would not have been conceivable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Oftentimes econometricians are faced with econometric models in which both parameters that are of interest to the economist, and those that are not, called nuisance parameters, are present. While statistical inference is drawn only on the parameters of interest, the treatment accorded by the investigator to the nuisance parameters can significantly affect the results. As an example, consider MaCurdy's (1981) model of life-cycle labor supply as discussed in Hsiao (1986). Here the econometric model of labor supply is as follows:

$$y_{it} = \alpha_i + \beta z_{it} + u_{it}, \quad i = 1, ..., I, \quad t = 1, ..., T,$$

where y_{it} is the logarithm of the labor supply of individual i at time t, z_{it} is the logarithm of the real wage rate, and α_i is a combination of unknown regressors and parameters relating to the individual's utility function. Here β is the elasticity of labor supply with respect to the wage, and is clearly the parameter of interest. α_i, on the other hand, is a nuisance parameter in the sense that we have no special interest in the specific value it takes. However, as is well known, attention must be paid to α_i before statistical inference on β can be conducted.

The most common way of dealing with nuisance parameters is to replace the unknown nuisance parameters with consistent estimates of these same parameters. While such a technique may be justified asymptotically, it is possible that this method may create additional variation in the statistics used to draw inference on the parameter of interest. This could potentially impair the accuracy of the asymptotic approximation of the distribution of the test statistics (see Andrews (1991) for one example of this). This problem is typically more severe in the presence of infinite-dimensional nuisance parameters.

Infinite-dimensional nuisance parameters present yet another problem: It may simply not be possible to estimate them consistently from the available data. An example of such a situation is the panel data model presented above. If J is fixed and $T \to \infty$, clearly the α's cannot be estimated consistently. In such situations, solution techniques other than estimation are called for.

This dissertation, which consists of seven chapters, studies ways of dealing with nuisance parameters in econometric models without having to estimate them directly. More specifically, means of conducting inference on the parameters of interest, that are robust to the structure implied by the nuisance parameters, are studied. One line of research explores the possibilities of conducting separate inference on the parameters of interest in likelihood models, which contain nuisance parameters. This work builds upon the theory of local cuts. The other line of research concerns hypothesis testing in models with serial correlation or heteroskedasticity of unknown form. In this setting, a test statistic that is robust to different error structures, (and does not require an actual estimate of the error structure) is developed.

The second chapter explores the notions of local cuts and adaptivity in general models that contain both parameters of interest and nuisance parameters. A parameter estimate is said to be adaptive if it is efficient on the model where the nuisance parameters are known and fixed, for all possible values of the nuisance parameters. Adaptivity therefore ensures efficient, separate estimation in the presence of nuisance parameters, and thus imposes restrictions on both the model and the specific estimator used. In contrast, the concept of local cuts is defined at the level of the model. Suppose it is the case that the model function (e.g., the likelihood function) can be written as the product of a marginal component (the density of a sufficient statistic for a subset of the parameters) depending on one set of parameters, and a conditional component (the distribution of the data conditional on the sufficient statistic), depending on the rest of the parameters. Then the model is said to contain a proper cut. Because of the nature of the split of the model function, inference can be conducted using only one part of the model function (separate inference). When such a split is only asymptotically possible, the model is said to have a local cut, and separate, asymptotic inference is justified. Thus local cuts provide us with a framework where separate inference, as opposed to merely separate estimation, is possible. This chapter explores the linkages between the concepts of local cuts and adaptivity, and provides intuitive geometric conditions under which these two concepts are equivalent. Furthermore guidelines are provided for situations where it is necessary to reparametrize the model in order to obtain a local cut or an adaptive estimator. This chapter as well as the third chapter is based on joint work with Nicholas M. Kiefer.

The third chapter extends the concept of local cuts to an estimating equation environment. The primary advantage of working in the estimating equation environment compared to the likelihood environment is that less direct knowledge about the distribution of the data is required. This implies that the theory is applicable to a larger set of models, such as, for example, the standard regression models. One obstacle to defining local cuts in the estimating equation framework is that the estimates obtained from estimating equations are invariant to all full-rank multiplicative transformations of the estimating equation, while the asymptotic distribution of the estimating equation is not. This creates a problem because we define local cuts from the properties of the asymptotic distribution of the estimating equation, but these properties change from transformation to transformation. To overcome this obstacle, we define a transformation of the estimating equation which eliminates this indeterminacy. Because of the specific properties of this transformed estimating equation, a very clear one to one relationship between local cuts in the estimating equation framework and the ability to conduct separate inference appears. Finally, the dynamic regression model, which will be investigated in great detail in chapters four through six, is introduced. We show that there is a local cut in this model in the estimating equation sense, justifying separate inference on a transformation of the regression parameter. In later chapters we show that this allows us to conduct inference on the regression parameter itself, if the test statistic is constructed just right.

The fourth chapter lays the foundation for a new test statistic that is robust to serial correlation/heteroskedasticity of unknown form. The statistic is developed to test hypotheses in linear regression models of the form introduced in the third chapter. The novel aspect of these tests is that they are simple and do not require heteroskedasticity and autocorrelation consistent (HAC) estimators; hence the size distortion caused by the estimation of the correlation structure is eliminated. The development of the new test relies upon a data-dependent transformation of the ordinary least squares estimates of the parameters. The approach expands the class of available HAC-robust asymptotically pivotal statistics. It is established that the limiting null distributions of these new tests have distributions that depend only upon the number of restrictions. This method of testing is applied to an empirical example and it is illustrated that the size of the new test is usually less distorted than tests that utilize HAC estimators. Examples where the new tests have greater finite sample power than tests using HAC estimators are provided. This chapter as well as the next is based on joint work with Nicholas M. Kiefer and Timothy J. Vogelsang.

In the fifth chapter, the test statistics introduced in the fourth chapter are extended to a non-linear weighted regression environment. Again the error structure is allowed to contain heteroskedasticity and serial correlation of unknown form. It is established that the class of tests introduced in the fourth chapter is applicable in this framework as well. Furthermore, an empirical example illustrating this new test statistic is provided. Specifically, the long-run effect of GDP growth on the growth of total restaurant revenues is examined using the new method, as well as methods currently employed in the literature. Finally simulations are performed, establishing that the size of the new test is less distorted than that of tests currently in use. Finite sample power for the different methods is studied as well, and it is demonstrated that power of the new test can dominate tests currently in use.

In the sixth chapter, the techniques introduced in the fourth and fifth chapters are employed to develop a test statistic that is robust to serial correlation/heteroskedasticity of unknown form in a cointegration environment that incorporates linear polynomial trend functions. The test can be employed to conduct inference on the trend function or the cointegration vector in a cointegration relationship, and to test hypotheses about the parameters of the deterministic trend function of a univariate time series. Extensive simulation experiments investigate the properties of the new test statistic in finite samples. These reveal that size distortions are generally less than those of tests currently employed in the literature; moreover, there is no substantial reduction in power.

The final chapter provides a concise summary of the main results and suggests some directions for future research.

Chapter 2

Local Cuts and Adaptivity

2.1 Introduction

When empirical analysis of an economic model is conducted, the relevant econometric model often contains both parameters that are of interest and some that are not, called nuisance parameters. The treatment of the nuisance parameters can affect the results of inference on the parameters of interest in a non-trivial manner. In fact, the ability to conduct inference may be impaired even when consistent and efficient estimates of the nuisance parameters can be obtained.

In this chapter, we will examine closely two concepts pertaining to models with nuisance parameters. The two concepts on which we will concentrate are adaptivity and local cuts. Adaptivity ensures efficient, separate estimation in the presence of a nuisance parameter. Local cuts provides us with a framework where separate inference may be justified.

An estimator is said to be adaptive if it is asymptotically normal and the best estimator in terms of efficiency whether or not the nuisance parameter is known. It is intuitively clear, that when an estimator has this property, separate estimation is justified. Separate inference is not however, because the asymptotic distribution of the estimator could depend on the nuisance parameter.

A local cut is an asymptotic version of a proper cut. A model is said to contain a proper cut if the model function (e.g. the likelihood function) can be written as the product of a marginal component, the density of a sufficient statistic for a subset of the parameters, and a conditional component (the distribution of the data conditional on the sufficient statistic), depending on the rest of the parameters. Because of the nature of this split, separate inference is now justified on one set of parameters in the marginal component and on the other set of parameters in the conditional component (Barndorff-Nielsen (1978)). When this split occurs only asymptotically, as opposed to in finite samples, the model is said to contain a local cut, and separate *asymptotic* inference may be justified. The idea behind local cuts as introduced by Christensen and Kiefer (1994) is exactly this; that most of the inference we conduct is asymptotic, and it therefore seems unnecessarily strict to require the model function split in finite samples.

In this chapter we first investigate the properties of models containing local cuts, with special attention being paid to regular models. The main results of this investigation is that both adaptivity and local cuts are closely linked to block-diagonality of the asymptotic Fisher information matrix. The results of this investigation are then used to provide specific conditions under which local cuts and adaptivity are equivalent. We show that it is relatively simple to obtain an adaptive estimator in a model where there is a local cut. When we have an adaptive estimator, it is a little more complicated to obtain a local cut. The additional restrictions that must be placed on the model, are restrictions on the asymptotic Fisher information and the way it depends on the nuisance parameters.

The rest of the chapter is organized as follows. In Section 2.2, we set up the framework we will be using, and provide the mathematical definitions of local cuts and adaptivity. A specific example of a model allowing a local cut is also presented. In Section 2.3, we first explore the properties of models that allow local cuts. Then these properties are exploited to formally link local cuts and adaptivity. Section 2.4 concludes.

2.2 Framework and Definitions

We examine i.i.d. observations, each of which is generated by the density P_θ. The model, which is parametric and regular, is defined by $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^m$. In order to define local cuts and adaptivity, we will assume that θ consists of two sets of parameters, such that $\Theta = V \times \Psi$, $\theta^T = (\upsilon, \eta)^T$, $\upsilon \in V$ and $\eta \in \Psi$. Let $\mathbf{P}_1(\eta_0)$ be the model where η is kept fixed at η_0, implying $\mathbf{P}_1(\eta_0) = \{P_\theta : \eta = \eta_0, \upsilon \in V\}$. Throughout the chapter, $\hat{\upsilon}$ is an estimate of the parameter υ, and it is implicitly assumed that $\hat{\upsilon}$ can depend on the entire data set (typically n observations). Whenever we have a specific estimator in mind, we will make it clear in the text. We are now in a position to define various properties of estimators.

Letting L denote a limit law, local Gaussian regularity is defined as follows:

Definition 1 *Let $\{\theta_n\}$, $\theta_n \in \Theta$ be a sequence such that $\sqrt{n}\,|\theta_n - \theta_0|$ stays bounded. Then $\hat{\upsilon}$ is a locally Gaussian regular (LGR) estimate of υ at P_{θ_0} if,*

$$L_{\theta_n}\left(\sqrt{n}\,(\hat{\upsilon} - \upsilon\,(P_{\theta_n}))\right) \rightarrow L_0$$

where L_0 is Gaussian and does not depend on $\{\theta_n\}$.

Adaptivity may now be defined following Bickel et. al. (1993).

Definition 2 *$\hat{\upsilon}$ is an adaptive estimate of υ in the presence of η if $\hat{\upsilon}$ is LGR on $\mathbf{P}$ and efficient in the model $\mathbf{P}_1(\eta)$, $\forall \eta$.*

Note that it is the efficiency of $\hat{\upsilon}$ in the model $\mathbf{P}_1(\eta)$ as opposed to the model $\mathbf{P}$ that is important. An estimator is efficient on a given model when it obtains the Cramer-Rao bound for that specific model. The Cramer-Rao bound in the model $\mathbf{P}_1(\eta)$ is 'lower' than in the model $\mathbf{P}$ in the sense that the matrix difference is negative semi-definite. These two Cramer-Rao bounds are only equal when the asymptotic Fisher information is block-diagonal. Therefore, a necessary condition for an estimate to be adaptive, is that the information matrix be block diagonal (this result can be found in Bickel et. al. (1993) and is formally stated in Theorem 4 below). It is, however, perfectly possible that some or all of the non-zero entries of the information matrix depends on the nuisance parameter (η), and therefore the asymptotic variance of $\hat{\upsilon}$ depends on the specific value of η. It is in this sense that the concept of adaptivity justifies separate estimation, but, in general, not separate inference.

Local cuts impose conditions directly on the likelihood function, and do not require regularity or that we are in a Gaussian environment. In fact, we can define local cuts without having to specify which is the nuisance parameter and which is the parameter of interest. The reason for this is that local cuts treat the parameters symmetrically. Also, it is not required that we specify exactly which estimator will eventually be employed, although the relevant estimators will all be similar in spirit to maximum likelihood estimators.

Letting $p(\mathbf{x}; \upsilon, \eta)$ be the likelihood function for n data points, we define local cuts following Christensen and Kiefer (1994).

Definition 3 *Letting T be a statistic (function of the data), first define* $f_c(\eta)$,

$s_c(\upsilon)$, $s_m(\eta)$ and $f_m(\upsilon)$ by the following four equations:

$$\ln p\left(\mathbf{x};\upsilon,\eta+\frac{\delta}{\sqrt{n}}|T\right)-\ln p\left(\mathbf{x};\upsilon,\eta|T\right) = O\left(n^{-\frac{1}{2}f_c(\eta)}\right) \quad (2.1)$$

$$\ln p\left(\mathbf{x};\upsilon+\frac{\varepsilon}{\sqrt{n}},\eta|T\right)-\ln p\left(\mathbf{x};\upsilon,\eta|T\right) = O\left(n^{-\frac{1}{2}s_c(\upsilon)}\right) \quad (2.2)$$

$$\ln p\left(T;\upsilon,\eta+\frac{\delta}{\sqrt{n}}\right)-\ln p\left(T;\upsilon,\eta\right) = O\left(n^{-\frac{1}{2}s_m(\eta)}\right) \quad (2.3)$$

$$\ln p\left(T;\upsilon+\frac{\varepsilon}{\sqrt{n}},\eta\right)-\ln p\left(T;\upsilon,\eta\right) = O\left(n^{-\frac{1}{2}f_m(\upsilon)}\right), \quad (2.4)$$

where $\varepsilon=(\varepsilon_1,...,\varepsilon_{m_1})$ and $\delta=(\delta_1,...,\delta_{m_2})$ satisfy $\delta\delta'=\varepsilon\varepsilon'=1$. Then T constitutes a local cut when:

$$f_c(\eta)>0\geq s_c(\upsilon),\ f_m(\upsilon)>0\geq s_m(\eta) \quad (2.5)$$

Note that the notation is such that f_c is the fast rate (f) in the conditional distribution ($_c$), s_m is the slow (s) rate in the marginal ($_m$) distribution and so on. The first inequality in (2.5) requires the effect on the conditional log-likelihood of permutations of η to disappear faster than those of υ. The second inequality in (2.5) makes a similar statement about the marginal log-likelihood function. It is in the sense that the parts of the likelihood function depend more on one parameter than the other. The separation of the fast and the slow rates by 0 ensures that the model will have a cut asymptotically.

The requirement that the rates be separated by 0 was not explicitly included in Christensen and Kiefer (1994). From Definition 3, it is clear that if the fast rates exceed 0, the difference in the likelihoods caused by the permutation of the nuisance parameters is $o(1)$:

$$\ln p\left(\mathbf{x};\upsilon,\eta+\frac{\delta_i}{\sqrt{n}}|T\right)-\ln p\left(\mathbf{x};\upsilon,\eta|T\right) = O\left(n^{-\frac{1}{2}f_c(\eta)}\right)=o(1)$$

$$\ln p\left(T;\upsilon+\frac{\varepsilon_i}{\sqrt{n}},\eta\right)-\ln p\left(T;\upsilon,\eta\right) = O\left(n^{-\frac{1}{2}f_m(\upsilon)}\right)=o(1).$$

This implies that there will be no asymptotic dependence on the nuisance parameters in the conditional and marginal likelihoods, and hence asymptotically, T constitutes a cut.

In practice it turns out that we frequently have a parameter of interest and a nuisance parameter, where we are not interested in conducting inference on the nuisance parameter. In such situations, the assumptions of a local cut, which treats both sets of parameters symmetrically, are unnecessarily strict. We are thus interested in relaxing the assumptions in a way that reflects the fact that we are interested in conducting inference on only one parameter. This leads us to the concepts of marginal and conditional local cuts. Specifically, a marginal local cut is defined as follows:

Definition 4 *T constitutes a marginal local cut when*

$$f_c(\eta) \geq s_c(\upsilon), \quad f_m(\upsilon) > 0 \geq s_m(\eta), \quad f_c(\eta) \geq s_m(\eta), \quad f_m(\upsilon) \geq s_c(\upsilon), \tag{2.6}$$

where $f_c(\eta)$, $s_c(\upsilon)$, $f_m(\upsilon)$ and $s_m(\eta)$ are defined by (2.1)-(2.4).

In the same manner, a conditional local cut is defined in the following manner:

Definition 5 *T constitutes a conditional local cut when*

$$f_c(\eta) > 0 \geq s_c(\upsilon), \quad f_m(\upsilon) \geq s_m(\eta), \quad f_c(\eta) \geq s_m(\eta), \quad f_m(\upsilon) \geq s_c(\upsilon), \tag{2.7}$$

where $f_c(\eta)$, $s_c(\upsilon)$, $f_m(\upsilon)$ and $s_m(\eta)$ are defined by (2.1)-(2.4).

The concepts of marginal and conditional local cuts will be especially useful when examining the links to adaptivity below. This is apparent because adaptivity concentrates on the ability to estimate one parameter and treats the other purely as a nuisance parameter.

Clearly, the "quality" of the inference in a model containing a local cut is affected by the actual difference between the fast and the slow rates. The larger the difference, the closer we are to a situation where there is a proper cut in the model. At the other end of the scale, if the fast and slow rates are all identical, the partial likelihoods depends equally on the two sets of parameters, and separate inference is clearly not justified. To formalize this intuition accurately, we define the order of a local cut in the following manner:

Definition 6 *A local cut is of order q if*

$$f_c(\eta) - s_c(\upsilon) \geq q, \qquad f_m(\upsilon) - s_m(\eta) \geq q.$$

The definitions of marginal and conditional local cuts of order q will then require that only the marginal or the conditional inequality respectively holds with a difference of q.

The following example provides an example of the concepts presented in a well-known model.

Example 1 An example of a local cut can be found in the standard one-dimensional normal distribution. The joint density of n observations is:

$$p\left(x;\mu,\sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum\left(x_i-\mu\right)^2\right)$$

Define the statistic, $T(x) = \frac{1}{n}\sum x_i$. Then the marginal density of $T(x)$ is:

$$p\left(T(x);\mu,\sigma^2\right) = \left(2\pi\frac{\sigma^2}{n}\right)^{-\frac{1}{2}} \exp\left(-\frac{n}{2\sigma^2}\left(T(x)-\mu\right)^2\right).$$

The density of the data conditional on $T(x)$ is:

$$\begin{aligned} p\left(x;\mu,\sigma^2|T(x)\right) &= \frac{p\left(x;\mu,\sigma^2\right)}{p\left(T(x);\mu,\sigma^2\right)} \\ &= \left(2\pi\sigma^2\right)^{-\frac{(n+1)}{2}} n^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{n}{2\sigma^2}T(x)^2\right) \end{aligned}$$

First we will compute the rate as defined by (2.1):

$$\ln p\left(x;\mu+\frac{\delta}{\sqrt{n}},\sigma^2|T(x)\right)-\ln p\left(x;\mu,\sigma^2|T(x)\right)=O_P\left(n^{-\frac{f_c(\mu)}{2}}\right).$$

Since $p\left(x;\mu,\sigma^2|T(x)\right)$ does not depend on μ at all, $f_c(\mu)=\infty$.

Correspondingly, we can find the slow rate in the conditional distribution as defined by (2.2):

$$\ln p\left(x;\mu,\sigma^2+\frac{\varepsilon}{\sqrt{n}}|T(x)\right)-\ln p\left(x;\mu,\sigma^2|T(x)\right)=O_P\left(n^{-\frac{s_c(\sigma^2)}{2}}\right)$$

Algebra (see Appendix A for details) proves that:

$$\ln p\left(x;\mu,\sigma^2+\frac{\varepsilon}{\sqrt{n}}|T(x)\right)-\ln p\left(x;\mu,\sigma^2|T(x)\right)$$

$$=-\tfrac{(n-1)}{2}\ln\left(1+\tfrac{\varepsilon}{\sigma^2\sqrt{n}}\right)-\tfrac{1}{2}\left(\frac{\frac{\varepsilon}{\sqrt{n}}}{\sigma^2\left(\sigma^2+\frac{\varepsilon}{\sqrt{n}}\right)}\right)\sum x_i^2+\tfrac{n}{2}\left(\frac{\frac{\varepsilon}{\sqrt{n}}}{\sigma^2\left(\sigma^2+\frac{\varepsilon}{\sqrt{n}}\right)}\right)T(x)^2$$

$$=\tfrac{\varepsilon\sqrt{n}}{2}\left[\frac{\varepsilon/\sqrt{n}}{\sigma^2(\sigma^2+\varepsilon/\sqrt{n})}\right]+\tfrac{1}{2}\left[\frac{\varepsilon}{\sigma^2(\sigma^2+\varepsilon/\sqrt{n})}\right]\left\{\tfrac{1}{\sqrt{n}}\sum\left[x_i^2-(\mu^2+\sigma^2)\right]\right\}+o(1)$$

$$-\tfrac{1}{2}\left[\frac{\varepsilon/\sqrt{n}}{\sigma^2(\sigma^2+\varepsilon/\sqrt{n})}\right]\left[\left\{\tfrac{1}{\sqrt{n}}\sum(x_i-\mu)\right\}\right]^2$$

$$-\left[\frac{\varepsilon\mu}{\sigma^2(\sigma^2+\varepsilon/\sqrt{n})}\right]\left\{\tfrac{1}{\sqrt{n}}\sum(x_i-\mu)\right\}$$

$$=O_P(1)$$

This implies that $-\frac{s_c(\sigma^2)}{2}=0$ and therefore that $s_c(\sigma^2)=0$. We have now obtained both rates for the conditional distribution, and can turn our attention to the marginal distribution.

The fast rate, defined by (2.4), is computed as follows:

$$\ln p\left(T(x);\mu,\sigma^2+\frac{\varepsilon}{\sqrt{n}}\right)-\ln p\left(T(x);\mu,\sigma^2\right)$$

$$=-\tfrac{1}{2}\ln\left(1+\tfrac{\varepsilon}{\sigma^2\sqrt{n}}\right)+\tfrac{1}{2}\left(\tfrac{\varepsilon/\sqrt{n}}{\sigma^2\left(\sigma^2+\frac{\varepsilon}{\sqrt{n}}\right)}\right)\left(\tfrac{1}{\sqrt{n}}\sum(x_i-\mu)\right)^2$$

$$=O(n^{-\frac{1}{2}})+O(n^{-\frac{1}{2}})O_P(1)=O_P(n^{-\frac{1}{2}})$$

This implies that $-\frac{1}{2}=-\frac{f_m(\sigma^2)}{2}$ or $f_m(\sigma^2)=1$. Now for the final rate, defined by 2.3:

$$\begin{aligned}&\ln p\left(T(x);\mu+\frac{\delta}{\sqrt{n}},\sigma^2\right)-\ln p\left(T(x);\mu,\sigma^2\right)\\ =\;&-\tfrac{n}{2\sigma^2}\left(T(x)-\mu-\tfrac{\delta}{\sqrt{n}}\right)^2+\tfrac{n}{2\sigma^2}\left(T(x)-\mu\right)^2\\ =\;&-\tfrac{\delta^2}{2\sigma^2}+\tfrac{\delta}{\sigma^2}\left(\frac{1}{\sqrt{n}}\sum(x_i-\mu)\right)=O_P(1)\end{aligned}$$

This gives us the final rate: $s_m(\mu)=0$. Holding all these calculations together, the rates are

$$f_c(\mu)=\infty,\; s_c(\sigma^2)=0,\; f_m(\sigma^2)=1,\; s_m(\mu)=0.$$

Comparing with 2.5, we note that $T(x)$ constitutes a local cut.

Note that the local cut in Example 1 is of order 1, but if we were interested in only σ^2, we could also view it as a conditional local cut of order infinity. Finally, since the asymptotic distribution of the estimate of μ obtained from the marginal distribution depends on σ^2, separate inference is not justified for μ, but it is in fact justified for σ^2 in the conditional distribution. The following theorem clarifies that this is no coincidence. In fact, we need a little more than a local cut in a model to be able to conduct separate inference. The following theorem states the conditions.

Theorem 1 *Separate inference is justified for both parameters in models containing local cuts of order strictly greater than 1, if* $\frac{\partial^2}{\partial \upsilon^2} \ln p(x|T)$ *does not depend on* η *asymptotically, and* $\frac{\partial^2}{\partial \eta^2} \ln p(T)$ *does not depend on* υ *asymptotically.*

This theorem gives the exact conditions under which separate inference is justified. Clearly for marginal inference alone, only a marginal local cut of order strictly greater than one, along with the requirement that $\frac{\partial^2}{\partial \eta^2} \ln p(T)$ does not depend on υ asymptotically, is required. The result for conditional inference is exactly analog to this.

2.3 Linking local cuts and adaptivity

In this section, we examine closely the links between local cuts and adaptivity. Adaptivity is a property of locally Gaussian regular estimators. For this reason it is useful to explore the rates of the local cut when the model is such that the maximum likelihood estimator is LGR. The following theorem states the result:

Theorem 2 *When the model allows a local cut and furthermore is such that the maximum likelihood estimates of both parameters are LGR, then the slow rates are both equal to 0.*

This theorem provides a benchmark for the numerical values of the slow rates in our subsequent discussion of local cuts. Additionally, if we have a proper cut, the fast rates will equal infinity. In this sense, a proper cut is nothing but a local cut of order infinity. Such a result is useful since it pins down the rates of local cuts in the standard frameworks.

It seems intuitively clear that adaptivity and local cuts are closely linked to information orthogonality (after all, some level of orthogonality between the parameters should be expected in situations where separate estimation and/or inference is justified). The restrictions a local cut places on the model function imply that the asymptotic information matrix is block diagonal. The result is stated in the following theorem.

Theorem 3 *When the model allows a local cut with slow rates equal to 0, the asymptotic information matrix is block-diagonal.*

We restrict ourselves to the case where the slow rates are 0, simply because this is the case where the information matrix is interesting. Recall that it is when the slow rates are equal to 0 that we are in a LGR environment. The following theorem provides the connection between adaptivity and block diagonality of the asymptotic information matrix.

Theorem 4 *Whenever an adaptive estimator exists, the asymptotic information matrix is block-diagonal.*

An important point which can be inferred from the proof of this theorem is that given a LGR estimator, the block-diagonality property is instrumental in delivering the efficiency of the estimator on $\mathbf{P}_1(\eta)$. Notice that when the information matrix is block-diagonal, the Cramer-Rao lower bounds on the variance are the same in $\mathbf{P}_1(\eta)$ and $\mathbf{P}$. This is a useful observation, which we state in the following theorem.

Theorem 5 *If $\hat{v}$ is LGR, efficient on* $\mathbf{P}$ *and the information matrix is block diagonal, $\hat{v}$ is adaptive.*

We have now established that both adaptivity and local cuts are closely linked to information orthogonality. In fact, our results indicate that in order to obtain either a local cut or an adaptive estimator, we need a model where the asymptotic information matrix is block-diagonal. In practice, it is often necessary to reparametrize the model for it to allow a local cut or an adaptive estimator. As a practical matter, therefore, when attempting to find the specific reparametrization, a good starting point would be a model with a block-diagonal information matrix. As such, out results have significant practical potency. The following theorem provides a general reparametrization which always produces a block-diagonal information matrix.

Theorem 6 *Let* $(\gamma_1, \gamma_2) = \left(\upsilon, \eta + I_{\upsilon\eta} (I_{\eta\eta})^{-1} \upsilon\right)$ *where*

$$I(\theta) = \begin{bmatrix} I_{\upsilon\upsilon}(\theta) & I_{\upsilon\eta}(\theta) \\ I_{\eta\upsilon}(\theta) & I_{\eta\eta}(\theta) \end{bmatrix}$$

is the asymptotic information matrix of the original model. Then the asymptotic information matrix of the model parametrized by (γ_1, γ_2) *is block-diagonal.*

Note that an important feature of this specific transformation is that it does not affect the parameter of interest. Another noticeable feature is that the scores of the transformed parameters are equal to the efficient scores in the original model:

$$\ln p(\upsilon, \eta) = \ln p\left(\gamma_1, \gamma_2 - I_{\upsilon\eta} (I_{\eta\eta})^{-1} \gamma_1\right)$$

$$\frac{\partial}{\partial \gamma_1} \ln p\left(\gamma_1, \gamma_2 - I_{\upsilon\eta} (I_{\eta\eta})^{-1} \gamma_1\right) = \frac{\partial}{\partial \upsilon} \ln p(\upsilon, \eta) - I_{\upsilon\eta} (I_{\eta\eta})^{-1} \frac{\partial}{\partial \eta} \ln p(\upsilon, \eta)$$

$$\frac{\partial}{\partial \gamma_2} \ln p\left(\gamma_1, \gamma_2 - I_{\upsilon\eta} (I_{\eta\eta})^{-1} \gamma_1\right) = \frac{\partial}{\partial \eta} \ln p(\upsilon, \eta)$$

It is well known (see Bickel et. al. (1993)) that this model is the starting point in the search for adaptive estimates. We have now demonstrated that it is the appropriate place to start when trying to obtain a local cut as well.

We are now ready to start examining the connections between adaptivity and local cuts. The following theorem provides the tie from local cuts to adaptivity.

Theorem 7 *If the model allows a local cut, any LGR estimate which is efficient on* **P** *will be adaptive.*

The fact that we can find an LGR estimate in a model which allows a local cut implies that the slow rates will be equal to 0. This by Theorem 3 implies that the Fisher information matrix is block-diagonal. Theorem 7 then follows directly from Theorem 5.

Theorem 7 fully describes the additional conditions required to obtain an adaptive estimate in a model with a local cut. The insight provided by Theorem 7 is that when the model allows for a local cut, only the additional requirement that an efficient LGR estimate exists, is required to permit separate estimation as well. It is worth noting that this extra requirement is needed only because we have insisted on defining adaptivity in a regular Gaussian environment. One could imagine defining a broader form of adaptivity where the basic intuition of adaptivity, namely that efficient estimation be possible as if the nuisance parameter was known, but where the restriction of a regular Gaussian framework does not hold. This generalization would require that efficiency be defined on a broader class of models. For an example where this has been done, see Saikkonen (1991).

In terms of linking adaptivity and local cuts, what remains to be done is to examine what additional conditions are required to ascertain whether we have a local cut in a model with an adaptive estimator.

Theorem 8 *If $\hat{\upsilon}$ is adaptive, and $I_{\upsilon\upsilon}(\theta)$ does not depend on η, then:*

a) $f_m(\eta)$ *defined by*

$$\ln p\left(\hat{\upsilon};\upsilon,\eta+\tfrac{\delta_i}{\sqrt{n}}\right)-\ln p\left(\hat{\upsilon};\upsilon,\eta\right)=O\left(n^{-\frac{1}{2}f_m(\eta)}\right)$$

is strictly greater than 0.

b) $s_m(\upsilon)$ *defined by*

$$\ln p\left(\hat{\upsilon};\upsilon+\tfrac{\varepsilon_i}{\sqrt{n}},\eta\right)-\ln p\left(\hat{\upsilon};\upsilon,\eta\right)=O\left(n^{-\frac{1}{2}s_m(\upsilon)}\right)$$

is 0.

Note that the restriction that $I_{\upsilon\upsilon}(\theta)$ not depend on η corresponds to a flatness in the direction of the nuisance parameter on the space where the metric is the information metric. Part a) follows directly from the fact that asymptotic distribution of $\hat{\upsilon}$ does not depend on η. Part b) follows from the asymptotic normality of $s_c(\upsilon)$.

The following theorem adds the restrictions that are necessary to obtain a marginal local cut in a model where the MLE is adaptive:

Theorem 9 *If* $\hat{\upsilon}_{MLE}$ *is adaptive,* $\hat{\eta}_{MLE}$ *is LGR and* $I_{\upsilon\upsilon}(\theta)$ *does not depend on* η*, then* $\hat{\upsilon}_{MLE}$ *constitutes a marginal local cut.*

Upon examination of Theorems 8 and 9, it is immediate that quite a few additional conditions are required to obtain a local cut, even when the model admits an adaptive estimator. This should not come as a surprise. After all, local cuts are a property of the underlying model, while adaptivity only provides information about the regularity of the model and the properties of one specific estimator.

2.4 Conclusion

In this chapter, we have carefully defined local cuts and adaptivity as well as marginal and conditional local cuts. We have examined the properties of models that allow local cuts or adaptive estimators. We stated the exact conditions under which a local cut implies that separate inference is justified. We have shown that block diagonality of the Fisher information matrix is a central requirement for both local cuts and adaptivity when dealing with regular models. We also outlined a method which may be used to reparametrize the model in order to obtain block diagonality of the Fisher information matrix. Finally we have studied the links between local cuts and adaptivity. Since adaptivity is a concept providing justification for separate estimation, while local cuts provide a basis for separate inference, one expects that several additional conditions are required to obtain a local cut in a model which admits an adaptive estimator. This expectation is fulfilled. Specifically, we show that the information matrix needs to be insensitive to the nuisance parameters, and the maximum likelihood estimates of all parameters need to be "well behaved". On the other hand, obtaining an adaptive estimate in a model which allows for a local cut is relatively easy. The only additional requirement is that a well behaved estimate of the parameters can be found. If this is the case, this estimate will then be adaptive.

It is often the case that when local cuts are used to justify separate inference, some estimate of the nuisance parameter is required. Alternatively it might be possible to condition on some statistic which would make the dependence on the nuisance parameter disappear. This need for a statistic or an estimate arises because a local cut only ensures that the score function will be free of nuisance parameters asymptotically. However, it is quite possible that the finite sample score used for estimation may depend on the nuisance parameters. It would be useful to explore what properties this estimate of the nuisance parameter must possess. In other words, what the 'worst' estimate of the nuisance parameter we could tolerate and still conduct good quality inference on the parameter of interest. This issue is left for future research.

Chapter 3

Local Cuts and Estimating Equations

3.1 Introduction

In the previous chapter, local cuts were analyzed in a likelihood setting. The likelihood environment requires the researcher to have substantial information about the data generating process; as such it provides an excellent environment in which to conduct very detailed analysis. Sometimes, however, we may not be able to specify the properties of the data generating process in as much detail as the likelihood environment requires. When this is the case, the estimating equation environment provides a good alternative. Furthermore, well-known methods of estimation such as maximum likelihood estimation and ordinary least squares estimation are all special cases of estimation using estimating functions. Therefore it seems like the natural next step in the analysis of local cuts to define a similar concept in an estimating equation environment.

The rest of the chapter is organized as follows. Section 3.2 provides the basic framework and notation, Section 3.3 defines local cuts in the estimating equation environment, and Section 3.4 provides an example of a dynamic model where a local cut in the estimating equation sense can be obtained.

3.2 Framework and definitions

Most of the literature on estimating equations is concerned with finite sample properties of the estimating equations and the estimates obtained therefrom (see for example Godambe (1991)).[1] This typically implies that assumptions are needed which ensure that various finite sample expectations are known. Instead, we focus only on asymptotic properties. The main reason for this choice is that the concept of local cuts is asymptotic; as such, it fits more naturally into a setting where the focus is on asymptotic properties.

Most of the notation in this section follows Heyde (1997). Unless otherwise specified, we operate in the class of estimating functions $\mathcal{G}$ of zero mean, square integrable estimating functions $G^n = G^n(x, \theta)$. At times, we will restrict ourselves to a subclass $\mathcal{H}$ of $\mathcal{G}$. As in the previous section, all expectations are taken with respect to P_θ. G^n is an $m-$dimensional vector with mean 0 for all $P_\theta \in \mathbf{P}$. Furthermore we assume throughout that $E(\partial G^n / \partial \theta)$ and $E\left(G^n (G^n)'\right)$ are non-singular. We also assume that integration and differentiation can be interchanged when necessary.

As before, we assume that θ consists of two sets of parameters, such that $\theta^T = (\upsilon, \eta)^T$. Next, we partition all the relevant terms into the parts involving υ and the parts involving η, so we write

$$G^n(\theta) = \begin{pmatrix} G^n_\upsilon(\theta) \\ G^n_\eta(\theta) \end{pmatrix}.$$

Furthermore, we assume that

$$\tilde{n}_\upsilon = diag(n^{q_1}, ..., n^{q_{m_1}}), \qquad \tilde{n}_\eta = diag(n^{p_1}, ..., n^{p_{m_2}})$$

[1] An execption to this is Heyde (1997), who has a chapter on asymptotic quasi-likelihood methods.

are two vectors, such that all stochastic variables of the form

$$\begin{aligned} n^{q_i} \left\{G_v^n(\theta)\right\}_i, \; i &= 1,..,m_1 \\ n^{p_i} \left\{G_\eta^n(\theta)\right\}_i, \; i &= 1,..,m_2 \end{aligned}$$

converge to a non-degenerate distribution. Finally let $\tilde{n} = diag(\tilde{n}_v, \tilde{n}_\eta)$.

3.3 Local Cuts and Estimating Equations

As a parallel to the definition of local cuts in the likelihood framework, we want to define local cuts in the estimating equation framework as a property of the asymptotic distribution of the estimating equation. A property of estimating equations that complicates this process is the following. Take a given estimating equation, $G^n(\theta) = 0$, and multiply it by an arbitrary non-singular function (deterministic or stochastic), which may or may not depend on the parameters. It is well known that we will still retain an estimating equation, which will provide us with the exact same estimate of θ. While such transformations do not affect the estimation of the parameter, they may affect the asymptotic distribution of the estimating equation. As such, defining local cuts here requires much more care than in the likelihood framework. The ultimate goal is to impose constraints on the asymptotic distribution of the estimating equation, such that we obtain a distribution of the estimate of the parameter of interest which does not depend on the nuisance parameter. The following definition satisfies this goal:

Definition 7 *The model allows a* EE-Local Cut *for the estimating function G, if*

$$-\tilde{n}\left[\frac{\partial}{\partial\theta}G^n(\theta)\right]^{-1} G^n(\theta) \Rightarrow \begin{bmatrix} X_1(v) \\ X_2(\eta) \end{bmatrix},$$

where $X_1(\upsilon)$ is a stochastic variable which only depends on υ and $X_2(\eta)$ is a stochastic variable which only depends on η.

While it may not be immediately obvious that this definition provides us with a justification for separate inference, the following Taylor expansion of the original estimating equation provides some additional insight into why a model with an EE-local cut allows separate inference:

$$\begin{aligned} G^n\left(\hat{\theta}\right) &= 0 = G^n(\theta) + \frac{\partial G^n(\bar{\theta})}{\partial \theta}\left(\hat{\theta}-\theta\right), \qquad \bar{\theta} \in \left[\hat{\theta};\theta\right] \Leftrightarrow \\ \left(\hat{\theta}-\theta\right) &= -\left[\frac{\partial G^n(\bar{\theta})}{\partial \theta}\right]^{-1} G^n(\theta) \end{aligned}$$

Now we can get at the distribution of the parameter estimates:

$$\tilde{n}\left(\hat{\theta}-\theta\right) = \begin{bmatrix} \tilde{n}_\upsilon(\hat{\upsilon}-\upsilon) \\ \tilde{n}_\eta(\hat{\eta}-\eta) \end{bmatrix} = -\tilde{n}\left[\frac{\partial G^n(\bar{\theta})}{\partial \theta}\right]^{-1} G(\theta) \Rightarrow \begin{bmatrix} X_1(\upsilon) \\ X_2(\eta) \end{bmatrix}.$$

Clearly, when we have a EE-local cut, $\tilde{n}_\upsilon(\hat{\upsilon}-\upsilon) \Rightarrow X_1(\upsilon)$ and $\tilde{n}_\eta(\hat{\eta}-\eta) \Rightarrow X_2(\eta)$, and separate inference is justified. Because of the invariance to transformations property of estimating equations, we generally have a lot of flexibility in our choice of estimating equations. Another way of viewing this definition of EE-local cuts, is that the relevant estimating equations to use for inference on $\upsilon(\eta)$ always is the first m_1 (last m_2) rows of the transformed estimating function $S^n(\theta) = \left\{-\tilde{n}\left[\frac{\partial G^n(\bar{\theta})}{\partial \theta}\right]^{-1} G^n(\theta)\right\}$. Furthermore, according to the definition, separate inference is justified when

$$S^n(\theta) \Rightarrow \begin{bmatrix} X_1(\upsilon) \\ X_2(\eta) \end{bmatrix}.$$

This transformed estimating function is the same as the one Heyde (1997) uses as the standardized estimating function when discussing asymptotic Quasi-likelihood.

Heyde goes on to note that this standardization is convenient. Our discussion above provides additional justification for using this specific standardization.

A useful property of $S^n(\theta)$ is that the transformation of $S^n(\theta)$ is $S^n(\theta)$ itself. To see this note that for $S^n(\theta)$, $\tilde{n}^s = 1$, such that

$$\begin{aligned} -\tilde{n}^s \left[\frac{\partial S(\bar{\theta})}{\partial \theta}\right]^{-1} S^n(\theta) &= -\left[\frac{\partial}{\partial \theta}\left\{-\tilde{n}\left[\frac{\partial G^n(\bar{\theta})}{\partial \theta}\right]^{-1} G^n(\theta)\right\}\right]^{-1} \times \\ &\quad \left\{-\tilde{n}\left[\frac{\partial G^n(\bar{\theta})}{\partial \theta}\right]^{-1} G^n(\theta)\right\} \\ &= -\tilde{n}\left[\tilde{n}\left[\frac{\partial G^n(\bar{\theta})}{\partial \theta}\right]^{-1} \frac{\partial}{\partial \theta} G^n(\theta)\right]^{-1} \left[\frac{\partial G^n(\bar{\theta})}{\partial \theta}\right]^{-1} G^n(\theta) \\ &= -\tilde{n}\left[\frac{\partial G^n(\bar{\theta})}{\partial \theta}\right]^{-1} G^n(\theta) = S^n(\theta). \end{aligned}$$

We have thus eliminated the freedom of choice in the selection of estimating functions and picked the one that is relevant for our purposes. Notice that the property of invariance to the transformation implies that, in practice, the function is unique and easy to find.

The transformation leading to the choice of $S^n(\theta)$ is similar in spirit to the parameter transformation in the likelihood framework leading to a model with a block-diagonal Fisher information matrix. Intuitively, separate inference requires some orthogonality, specifically information orthogonality, between the two parameters. The mathematical parallel to block diagonal Fisher information in this framework, is

$$\text{plim}\left[\frac{\partial}{\partial \eta} S_v^n(\theta)\right] = \text{plim}\left[\frac{\partial}{\partial v} S_\eta^n(\theta)\right] = 0.$$

To see that $S^n(\theta)$ in fact has this property, note that

$$\begin{aligned} \frac{\partial}{\partial \eta} S_v^n &= -\frac{\partial}{\partial \eta}\left\{\left(\frac{\partial G_v^n}{\partial v} - \frac{\partial G_v^n}{\partial \eta}\left(\frac{\partial G_\eta^n}{\partial \eta}\right)^{-1}\frac{\partial G_\eta^n}{\partial v}\right)^{-1}\left(G_v^n - \frac{\partial G_v^n}{\partial \eta}\left(\frac{\partial G_\eta^n}{\partial \eta}\right)^{-1} G_\eta^n\right)\right\} \\ &= -\left(\frac{\partial G_v^n}{\partial v} - \frac{\partial G_v^n}{\partial \eta}\left(\frac{\partial G_\eta^n}{\partial \eta}\right)^{-1}\frac{\partial G_\eta^n}{\partial v}\right)^{-1}\left(\frac{\partial G_v^n}{\partial \eta} - \frac{\partial G_v^n}{\partial \eta}\left(\frac{\partial G_\eta^n}{\partial \eta}\right)^{-1}\frac{\partial G_\eta^n}{\partial \eta}\right) \\ &= 0. \end{aligned}$$

In a similar manner, it can be shown that $\frac{\partial}{\partial v} S_\eta^n = 0$.

As in the likelihood framework, there will be situations where we are interested in only one of the parameters. In the likelihood framework, we defined marginal and conditional local cuts to deal with precisely this type of situation. Next, we introduce a parallel concept.

Definition 8 *The model allows a* One-Sided EE-Local Cut *for the estimating function* G, *if*

$$-\tilde{n}_v \left[\frac{\partial G_v^n(\theta)}{\partial v}\right]^{-1} G_v^n(\theta) \Rightarrow X_1(v).$$

This One-Sided EE-Local Cut provides justification for separate inference on v. Note that we are allowing for a situation where we do not even have an estimating equation for the nuisance parameter. Since this is not an uncommon situation, especially if the nuisance parameter is infinite dimensional, this greatly increases the usefulness of the concept.

We are now ready to look at an example of a situation where estimating equations are a useful way of thinking about the model.

3.4 A Dynamic Model

Consider the following dynamic regression model

$$y = X\beta + u,$$

where β is a $(k \times 1)$ vector of regression parameters, X is a $(T \times k)$ vector of regressors which may include a constant, and $u = \{u_t\}$ is a mean zero (conditional on X) random process. It is assumed that u does not have a unit root, but u may be serially correlated and conditional heteroskedastic. Let $v_t = X_t u_t$ and define $\Omega = \Lambda\Lambda' = \Gamma_0 + \sum_{j=1}^{\infty}(\Gamma_j + \Gamma_j^{'})$ where $\Gamma_j = E(v_t v_{t-j}')$ and Λ is a lower triangular matrix based on the Cholesky decomposition of Ω. Define $S_t = \sum_{j=1}^{t} v_j$ which are the partial sums of $\{v_t\}$. Let $W_k(r)$ denote a k-vector of independent Wiener processes, and let $[rT]$ denote the integer part of rT where $r \in [0,1]$. In this model, the parameter of interest is β, and the parameters relating to the correlation structure, namely Ω, are nuisance parameters. Clearly there is not enough available information to analyze this model using maximum likelihood methods. The standard method used to estimate β in this type of model is Ordinary Least Squares (see Hamilton (1994)). This implies that the first order conditions, or the estimating functions, relating to β, are

$$G_\beta(\beta) = X'(y - X\beta) = 0.$$

We note that under simple regularity conditions, (see Chapter 4)

$$T^{-\frac{1}{2}} X'(y - X\beta) = T^{-\frac{1}{2}} X'u \Rightarrow N_k(0, \Omega) = \Lambda W_k(1),$$

and therefore, separate inference is clearly not justified. Notice that this is a situation where we are clearly interested only in one set of parameters, namely β.

It turns out that it is necessary to transform the regression parameter to be able to conduct separate inference. To get at the relevant transformation, we need a few additional definitions:

$$\begin{aligned}\hat{v}_t &= X_t\hat{u}_t = \left(y_t - X_t\hat{\beta}\right) \\ \hat{S}_t &= \sum_{j=1}^{t} \hat{v}_j \\ \hat{C} &= T^{-2}\sum_{t=1}^{T} \hat{S}_t\hat{S}_t^{'} \\ \hat{M} &= \left(\frac{1}{T}X'X\right)^{-1}\hat{C}^{\frac{1}{2}}\end{aligned}$$

In Chapter 4, it will be shown that $\hat{M} \Rightarrow Q^{-1}\Lambda P_k^{\frac{1}{2}}$ where $Q = plim\left(\frac{1}{T}X'X\right)$ and $P_k^{\frac{1}{2}}$ is a stochastic variable that does not depend on any parameters. We now let $\tilde{\beta} = \hat{M}^{-1}\beta$ such that the estimating equation for $\tilde{\beta}$ becomes

$$G_{\tilde{\beta}}\left(\tilde{\beta}\right) = \hat{M}'X'\left(y - X\hat{M}\tilde{\beta}\right) = 0.$$

We can now find the asymptotic distribution of the transformed estimating equation.

$$\begin{aligned}-\tilde{n}_{\tilde{\beta}}\left[\frac{\partial G_{\tilde{\beta}}^{n}\left(\tilde{\beta}\right)}{\partial\tilde{\beta}}\right]^{-1} G_{\tilde{\beta}}^{n}\left(\tilde{\beta}\right) &= -T^{-\frac{1}{2}}\left[-\hat{M}'X'X\hat{M}\right]^{-1}\hat{M}'X'\left(y - X\hat{M}\tilde{\beta}\right) \\ &= T^{-\frac{1}{2}}\hat{M}^{-1}\left(X'X\right)^{-1}X'\left(y - X\hat{M}\tilde{\beta}\right) \\ &\Rightarrow P_k^{\frac{1}{2}}\Lambda^{-1}QQ^{-1}\Lambda W_k(1) = P_k^{\frac{1}{2}}W_k(1).\end{aligned}$$

Thus, the asymptotic distribution of the transformed estimating equations does not depend on any nuisance parameters, and therefore this model allows for a One-Sided EE-Local Cut. Hence, separate inference is justified. To verify this, notice that the asymptotic distribution of $\widehat{\tilde{\beta}} = \hat{M}^{-1}\hat{\beta}$ is $P_k^{\frac{1}{2}}W_k(1)$, which does not depend on any nuisance parameters. To foreshadow, in Chapter 4 we will show that this result actually allows us to construct a test for hypothesis on the untransformed parameter, β, which does not depend on nuisance parameters.

Chapter 4

Simple Robust Testing of Regression Hypotheses

4.1 Introduction

In this chapter we consider the problem of hypothesis testing in models with errors that have serial correlation or heteroskedasticity of unknown form. This situation is often encountered in regression models applied to economic time series data. It is a classic textbook result that while ordinary least squares (OLS) estimates of regression parameters remain consistent and asymptotically normal when errors are heteroskedastic or autocorrelated (provided usual regularity conditions hold and no lagged dependent variables are in the model), standard tests are no longer valid. If the true form of serial correlation/heteroskedasticity were known, then generalized least squares (GLS) provides efficient estimates and standard inference can be conducted on the GLS transformed model. But, in practice the form of serial correlation/heteroskedasticity is often unknown, and this has led to the development of techniques that permit valid asymptotic inference without having to specify a model of the serial correlation or heteroskedasticity. The most common approach is to estimate the variance-covariance matrix of the OLS estimates nonparametrically using spectral methods (heteroskedasticity and autocorrelation consistent (HAC) estimators) and construct standard tests using the asymptotic normality of the OLS estimates. HAC estimators have been extensively analyzed in the literature and important contributions are given by Andrews (1991), Andrews and Monahan (1992), Gallant (1987), Hansen (1992), Newey and West (1987), Robinson (1991) and White (1984) among others. The benefit of HAC estimator tests is asymptotically valid inference that is robust to general forms of serial correlation/heteroskedasticity in the errors.

We propose an alternative method of constructing robust test statistics. We apply a non-singular data dependent stochastic transformation to the OLS estimates. The asymptotic distribution of the transformed estimates does not depend on nuisance parameters. Then, test statistics which are asymptotically invariant to nuisance parameters (asymptotic pivotal statistics) are constructed. The resulting test statistics have nonstandard asymptotic distributions that only depend on the number of restrictions being tested, and critical values are easy to simulate using standard techniques. The main advantage of our approach compared to standard approaches is that estimates of the variance-covariance matrix are not explicitly required to construct the tests. This is potentially important since simulation studies have shown that sampling variability of HAC estimators in finite samples can lead to tests that have substantial size distortions (e.g. Andrews (1991), Andrews and Monahan (1992) and Den Haan and Levin (1997)). We report results from finite sample simulations which show that our new tests have better finite sample size properties than HAC estimator tests (including prewhitening).

The transformation of the OLS estimates used in this chapter is the same transformation used in the dynamic example of Section 3.4. There, it was shown that when this transformation is carried out, the model allows a One-sided EE-local cut. It should not come as a surprise therefore, that separate inference can be conducted on the transformed parameter. In this chapter, we show that separate inference can be conducted on the original regression parameter as well.

The remainder of the chapter is organized as follows. In the next section we lay out the model and review some well known OLS results. We show how the OLS estimates can be transformed so that their joint distribution becomes asymptotically invariant to serial correlation/heteroskedasticity nuisance parameters. Natural by-products of this transformation are t type statistics for testing hypotheses about individual parameters. In Section 4.3 we show how to construct tests of general linear hypotheses. Limiting null distributions are obtained, and asymptotic critical values are tabulated. The tests developed in these sections are natural extensions to regression models of the univariate trend function tests proposed by Vogelsang (1998). The tests in Vogelsang (1998) share the property that serial correlation parameters need not be estimated to carry out valid asymptotic inference. In Section 4.4 we show how our approach easily extends to GLS and instrumental variables (IV) estimation. We report results on local asymptotic power of the new tests compared to HAC estimator tests in Section 4.5. We show that the new tests have nontrivial local asymptotic power that is comparable but slightly below that of HAC estimator tests. We note that local asymptotic power calculations for HAC estimator tests are the same as those for tests with known variance-covariance parameters, while our statistic implicitly corrects for unknown variance-covariance parameters. In Sections 4.6 and 4.7 we report results on the finite sample behavior of the tests. Because the local asymptotic power approximation does not capture the influence of sampling variability of HAC estimators on finite sample power, we provide cases based on an empirical example where the power of the new tests dominates power of HAC estimator tests. Since our tests can be more powerful and they dominate HAC estimator tests in the accuracy of the asymptotic null approximation, our tests are very competitive in practice. Section 4.8 concludes with proofs given in Appendix B.

4.2 The Model and Some Asymptotic Results

Consider the regression model given by

$$y_t = X_t\beta + u_t, \qquad t = 1, 2, ..., T, \tag{4.1}$$

where β is a $(k \times 1)$ vector of regression parameters, X_t is a $(k \times 1)$ vector of regressors which may include a constant, and $\{u_t\}$ is a mean zero (conditional on X_t) random process. It is assumed that u_t does not have a unit root, but u_t may be serially correlated and conditional heteroskedastic. The following notation is used throughout the chapter. Let $v_t = X_t u_t$ and define $\Omega = \Lambda\Lambda' = \Gamma_0 + \sum_{j=1}^{\infty}(\Gamma_j + \Gamma_j')$ where $\Gamma_j = E(v_t v_{t-j}')$ and Λ is a lower triangular matrix based on the Cholesky decomposition of Ω. Note that Ω is equal to 2π times the spectral density matrix of v_t evaluated at frequency zero. Define $S_t = \sum_{j=1}^{t} v_j$ which are the partial sums of $\{v_t\}$. Let $W_k(r)$ denote a k-vector of independent Wiener processes, and let $[rT]$ denote the integer part of rT where $r \in [0, 1]$. We use $\Rightarrow$ to denote weak convergence.

The following two assumptions regarding X_t and u_t are sufficient for us to obtain our main results.

Assumption 1 $T^{-\frac{1}{2}} \sum_{t=1}^{[rT]} v_t \Rightarrow \Lambda W_k(r)$ *for all* $r \in [0, 1]$.

Assumption 2 $plim(T^{-1} \sum_{t=1}^{[rT]} X_t X_t') = rQ$ *for all* $r \in [0, 1]$ *and* Q^{-1} *exists.*

Assumption 1 holds under a variety of regularity conditions. One set of conditions are given by Phillips and Durlauf (1986) which require that v_t be weakly stationary, that the elements of v_t have a finite moment greater than two, and that v_t satisfies well known α-mixing conditions. These conditions permit conditional heteroskedasticity in $\{v_t\}$ but rule out most forms of unconditional heteroskedasticity. Andrews (1991) showed that consistent HAC estimators can be obtained under the stronger assumption that $\{v_t\}$ is fourth order stationary and α-mixing (see his Lemma 1). Assumption 1 is also satisfied by stationary and invertible ARMA processes with innovations with finite fourth moments (see Hall and Heyde (1980)). Assumption 1 rules out unit roots in $\{X_t\}$ and $\{u_t\}$.

Assumption 2 holds, for example, when X_t is a weakly (second order) stationary vector process and rules out trends in the regressors. However, the asymptotic results remain valid for certain hypotheses if the regressors are trend stationary. To be more precise suppose the regression model is $y_t = \mu + \gamma t + X_t'\beta + u_t$ and $X_t = \mu_x + \gamma_x t + \zeta_t$ where μ_x and γ_x are $(k \times 1)$ vectors, and $\{\zeta_t\}$ and $\{\zeta_t u_t\}$ satisfy Assumptions 1 and 2. In Appendix Bwe show that the new statistic proposed in this chapter is invariant to projections of subsets of regressors. Therefore, hypotheses involving β can be tested using the regression $\tilde{y}_t = \tilde{X}_t'\beta + \tilde{u}_t$ where $\{\tilde{y}_t\}$ and $\{\tilde{X}_t\}$ are residuals from the projection of $\{y_t\}$ and $\{X_t\}$ onto the space spanned by $\{1, t\}$. This regression satisfies Assumptions 1 and 2 because $\tilde{X}_t = \tilde{\zeta}_t$ and it is easy to show that $T^{-\frac{1}{2}} \sum_{t=1}^{[rT]} \tilde{\zeta}_t u_t = T^{-\frac{1}{2}} \sum_{t=1}^{[rT]} \zeta_t u_t + o_P(1)$ and $T^{-1} \sum_{t=1}^{[rT]} \tilde{\zeta}_t \tilde{\zeta}_t' = T^{-1} \sum_{t=1}^{[rT]} \zeta_t \zeta_t' + o_P(1)$. Once $\{t\}$ is included in the regression, the asymptotic results we obtain for tests of the β parameters do not apply to tests that involve the parameters μ (the intercept) or γ in which case the asymptotic distributions depend on the specific deterministic trends included in the regression.

Suppose regression (4.1) is estimated by OLS to obtain $\hat{\beta}$, the OLS estimate. The limiting distribution of $T^{\frac{1}{2}}(\hat{\beta}-\beta)$ follows directly from Assumptions 1 and 2 as

$$\begin{aligned} T^{\frac{1}{2}}(\hat{\beta}-\beta) &= (T^{-1}\sum_{t=1}^{T} X_t X_t')^{-1}\, T^{-\frac{1}{2}}\sum_{t=1}^{T} X_t u_t = (T^{-1}\sum_{t=1}^{T} X_t X_t')^{-1}\, T^{-\frac{1}{2}} S_T \quad (4.2) \\ &\Rightarrow Q^{-1}\Lambda W_k(1) \sim N(0, Q^{-1}\Lambda\Lambda' Q^{-1}) = N(0, Q^{-1}\Omega Q^{-1}) = N(0, V). \end{aligned}$$

This asymptotic normality result can be used to test hypotheses about β. To construct standard tests that are asymptotically invariant to nuisance parameters, an estimate of $V = Q^{-1}\Omega Q^{-1}$ is required. A natural estimator of Q^{-1} is $(T^{-1}\sum_{t=1}^{T} X_t X_t')^{-1}$. A HAC estimator of Ω can be constructed from $\hat{v}_t = X_t\hat{u}_t$ where $\hat{u}_t$ are the OLS residuals.

Consider the estimator $\hat{V} = (T^{-1}\sum_{t=1}^{T} X_t X_t')^{-1}\hat{\Omega}(T^{-1}\sum_{t=1}^{T} X_t X_t')^{-1}$ where $\hat{\Omega}$ is a HAC estimator of Ω. If we transform $T^{\frac{1}{2}}(\hat{\beta}-\beta)$ using $\hat{V}^{-\frac{1}{2}} = (T^{-1}\sum_{t=1}^{T} X_t X_t')^{-1}\hat{\Lambda}$ where $\hat{\Lambda}$ is obtained from the Cholesky decomposition of $\hat{\Omega}$, we have

$$\hat{V}^{-\frac{1}{2}} T^{\frac{1}{2}}(\hat{\beta}-\beta) \Rightarrow N(0, I_k). \quad (4.3)$$

Using (4.3), hypotheses about individual $\beta' s$ can be tested using t statistics in the usual way with standard errors given by square roots of the diagonal elements of tne matrix $\hat{V}/T$. The asymptotic theory does not explicitly account for the effects of sampling variation in $\hat{V}$, and this variation it is potentially important in finite samples.

We take a different approach to testing that is similar in spirit to the transformation in (4.3) except that we transform $T^{\frac{1}{2}}(\hat{\beta}-\beta)$ using a moment matrix constructed from the data that does not require an estimate of Ω. Define $\hat{S}_t = \sum_{j=1}^t X_j \hat{u}_j = \sum_{j=1}^t \hat{v}_j$. Using Assumptions 1 and 2, the limiting behavior of $T^{-\frac{1}{2}}\hat{S}_{[rT]}$ as $T \to \infty$ is

$$\begin{aligned} T^{-\frac{1}{2}}\hat{S}_{[rT]} &= T^{-\frac{1}{2}}\sum_{t=1}^{[rT]} X_t\hat{u}_t = T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\{X_t u_t - X_t X_t'(\hat{\beta}-\beta)\} \\ &= T^{-\frac{1}{2}}\sum_{t=1}^{[rT]} v_t - (T^{-1}\sum_{t=1}^{[rT]} X_t X_t')\ T^{\frac{1}{2}}(\hat{\beta}-\beta) \\ &= T^{-\frac{1}{2}}S_{[rT]} - (T^{-1}\sum_{t=1}^{[rT]} X_t X_t')\ T^{\frac{1}{2}}(\hat{\beta}-\beta) \\ &\Rightarrow \Lambda W_k(r) - rQQ^{-1}\Lambda W_k(1) = \Lambda(W_k(r) - rW_k(1)). \end{aligned} \tag{4.4}$$

Consider $\hat{C} = T^{-2}\sum_{t=1}^T \hat{S}_t\hat{S}_t'$. From (4.4) and the continuous mapping theorem we have

$$\hat{C} \Rightarrow \Lambda[\int_0^1 (W_k(r) - rW_k(1))(W_k(r) - rW_k(1))'dr]\Lambda'. \tag{4.5}$$

To simplify later developments let $P_k = \int_0^1 (W_k(r) - rW_k(1))(W_k(r) - rW_k(1))'dr$ which is the integral of the outer product of a k-dimensional multivariate Brownian bridge. In the univariate case P_1' is the limiting distribution of the Cramér-von Mises statistic and is related to the Anderson-Darling statistic. Because P_k is positive definite by construction, we can use a Cholesky decomposition to write $P_k = Z_k Z_k'$ or equivalently $P_k^{-1} = (Z_k')^{-1}Z_k^{-1}$ where Z_k is lower triangular.

Now consider $\hat{B} = (T^{-1}\sum X_t X_t')^{-1}\hat{C}(T^{-1}\sum X_t X_t')^{-1}$. Define $\hat{M} = (T^{-1}\sum X_t X_t')^{-1}\hat{C}^{\frac{1}{2}}$ with $\hat{C}^{\frac{1}{2}}$ lower triangular and the Cholesky decomposition of $\hat{C}$. Consider the transformation $\hat{M}^{-1}T^{\frac{1}{2}}(\hat{\beta}-\beta)$. It follows directly from

(4.2) and (4.5):

$$\hat{M}^{-1}T^{\frac{1}{2}}(\hat{\beta}-\beta) \Rightarrow Z_k^{-1}W_k(1). \tag{4.6}$$

This transformation results in a limiting distribution that does not depend on the nuisance parameters Q and Ω. The distribution of $Z_k^{-1}W_k(1)$ is nonstandard. Because $W_k(1)$ and $W_k(r)-rW_k(1)$ are Gaussian and $E[W_k(1)(W_k(r)-rW_k(1))] = 0$ for all $r \in [0,1]$, they are independent, and it follows that Z_k and $W_k(1)$ are independent as well. Therefore, conditional on Z_k, $Z_k^{-1}W_k(1) \sim N(0, P_k^{-1})$. If we let $p(P_k)$ denote the distribution function of P_k, we can write the unconditional distribution of $Z_k^{-1}W_k(1)$ as $\int_0^1 N(0, P_k^{-1})p(P_k)dP_k$ which is a mixture of normals. Thus, the distribution of $Z_k^{-1}W_k(1)$ is symmetric with thicker tails than a normal distribution. This result is analogous to Fisher's classic development of the t statistic. After using a data dependent stochastic transformation (dividing by a moment of the data proportional to the error variance), Fisher obtained a finite sample distribution free of nuisance parameters with fatter tails than a normal distribution, a t distribution. This analogy is not exact as we obtain a distribution free of nuisance parameters only asymptotically, and the distribution of $Z_k^{-1}W_k(1)$ is not equivalent to a multivariate t distribution. But, the analogy is accurate as a nuisance parameter is eliminated and this results in increased dispersion of the null limiting distribution.

Hypotheses about individual β's can be tested using t type statistics which we label t^* that are constructed in the same way as usual t statistics with the usual standard errors replaced with square roots of the diagonal elements of the $\hat{B}/T$ matrix. Because the t^* statistics are invariant to the ordering of the regressors, the limiting distribution of any t^* is given by the first element in the vector $Z_k^{-1}W_k(1)$. Using the fact that Cholesky decompositions are lower triangular, it

Table 4.1: Asymptotic Critical values of t*

1.0%	2.5%	5.0%	10.0%	50.0%	90.0%	95%	97.5%	99.0%
-8.544	-6.811	-5.374	-3.890	0.000	3.890	5.374	6.811	8.544

is easy to show that the first element of $Z_k^{-1} W_k(1)$ has the same distribution as $W_1'(1)/[\int_0^1 (W_1'(r) - rW_1'(1))^2 dr]^{\frac{1}{2}}$. Therefore, as $T \to \infty$

$$t^* \Rightarrow W_1(1)/[\int_0^1 (W_1(r) - rW_1(1))^2 dr]^{\frac{1}{2}}. \tag{4.7}$$

Critical values of (4.7) were computed using simulations and are tabulated in Table 4.1. The Wiener process, $W_1'(r)$, was approximated by normalized sums of i.i.d. $N(0,1)$ pseudo random deviates using 1,000 steps and 50,000 replications. The simulations were written in the GAUSS programming language using an initial seed of 1,000 for the random number generator.

We also computed the density of (4.7) and the density of (4.7) with variance normalized to one by smoothing the 50,000 realizations of (4.7) using standard kernel techniques.[1] These densities are plotted in Figure 4.1 along with the density of a standard normal random variable. The asymptotic distribution of the normalized t^* has tails slightly fatter than a standard normal random variable.

[1]We computed the variance of (4.7) to be 10.893 using the sample variance of the 50,000 replications.

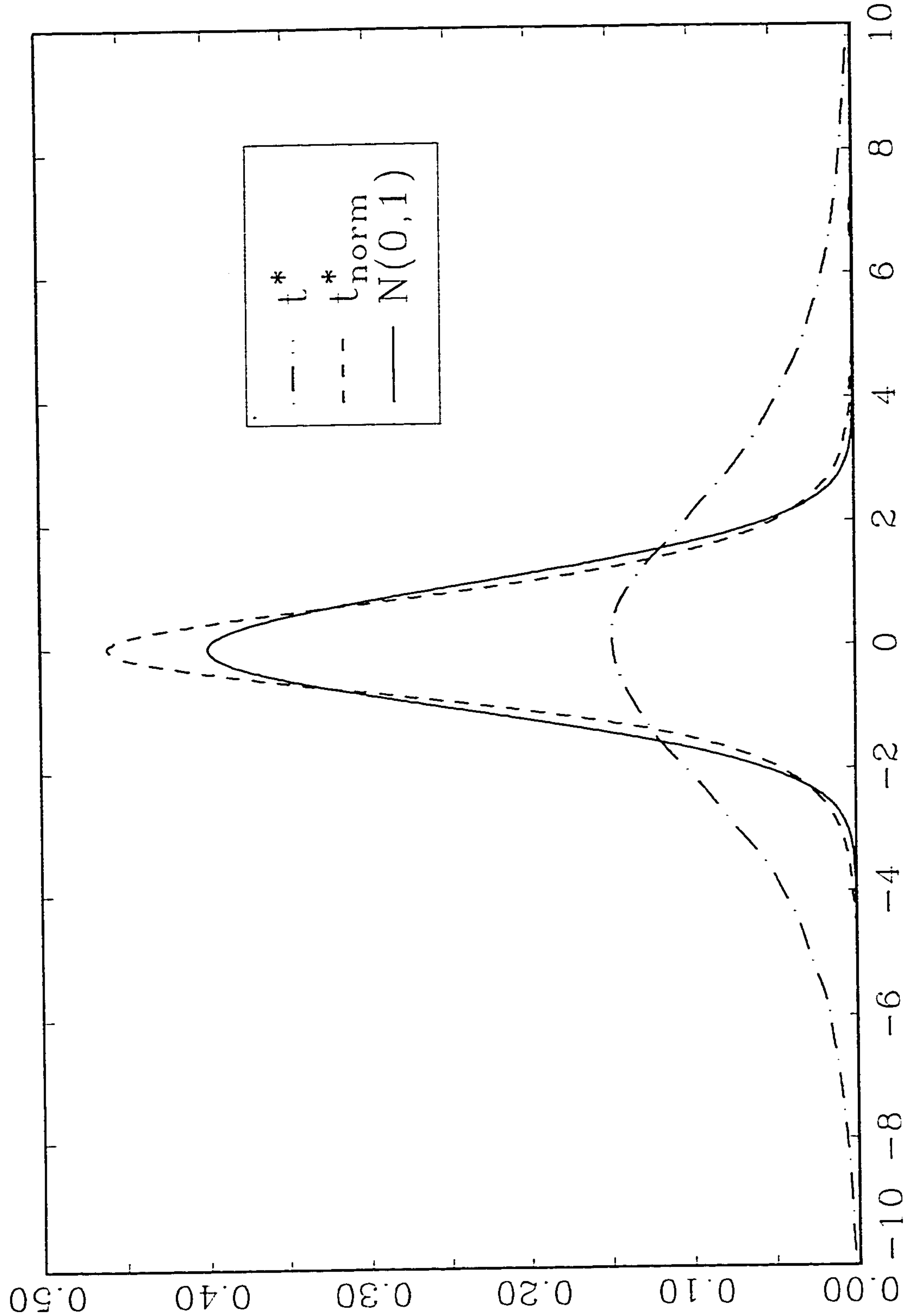

Figure 4.1: Densities of t^*, t^*_{norm}, and N(0,1).

4.3 Tests for General Linear Hypotheses

Suppose we are interested in testing more general linear hypotheses of the form

$$H_0 : R\beta = r, \qquad H_1 : R\beta \neq r,$$

where R is a $(q \times k)$ matrix with rank q and r is a $(q \times 1)$ vector. When the null hypothesis is true, we have $R\hat{\beta} - r = R(\hat{\beta} - \beta)$. To motivate the new statistic consider $T^{\frac{1}{2}}R(\hat{\beta} - \beta)$. From (4.2) it follows that $T^{\frac{1}{2}}R(\hat{\beta} - \beta) \Rightarrow RQ^{-1}\Lambda W_k(1)$. Because $W_k(1)$ is a vector of independent Wiener processes and is Gaussian, $RQ^{-1}\Lambda W_k(1)$ is equivalent in distribution to $\Lambda^* W_q^*(1)$ where $W_q^*(1)$ is a $(q \times 1)$ vector of independent Wiener processes and Λ^* is the $(q \times q)$ matrix square root of $RQ^{-1}\Lambda\Lambda' Q^{-1}R'$. Λ^* exists and is invertible because the matrix $RQ^{-1}\Lambda\Lambda' Q^{-1}R'$ has full rank of q. Now consider the matrix $R\hat{B}R'$. It is simple to show that $R\hat{B}R' \Rightarrow RQ^{-1}\Lambda P_k \Lambda' Q^{-1}R$ which is equivalent in distribution to $\Lambda^* P_q^* \Lambda^{*\prime}$ (see Appendix B). Let $\hat{M}^*$ denote the matrix square root of $R\hat{B}R'$ and note that $\hat{M}^* \Rightarrow [\Lambda^* P_q^* \Lambda^{*\prime}]^{\frac{1}{2}} = \Lambda^* Z_q^*$. Suppose we transform $T^{\frac{1}{2}}R(\hat{\beta} - \beta)$ using $\hat{M}^{*-1}$ giving $\hat{M}^{*-1}T^{\frac{1}{2}}R(\hat{\beta} - \beta)$. It evidently follows that $\hat{M}^{*-1}T^{\frac{1}{2}}R(\hat{\beta} - \beta) \Rightarrow Z_q^{*-1}W_q^*(1)$ which is free of nuisance parameters. Forming the usual quadratic form using $\hat{M}^{*-1}T^{\frac{1}{2}}R(\hat{\beta} - \beta)$ gives

$$[\hat{M}^{*-1}T^{\frac{1}{2}}R(\hat{\beta} - \beta)]'[\hat{M}^{*-1}T^{\frac{1}{2}}R(\hat{\beta} - \beta)] = T(R(\hat{\beta} - \beta))'[R\hat{B}R']^{-1}R(\hat{\beta} - \beta). \tag{4.8}$$

The quadratic form (4.8) suggests the following statistic for testing H_0 against H_1',

$$F^* = T(R\hat{\beta} - r)'[R\hat{B}R']^{-1}(R\hat{\beta} - r)/q.$$

Notice that F^* is the classic F test except that $\hat{B}$ replaces $\hat{V}$. (If $T^{\frac{1}{2}}R(\hat{\beta} - \beta)$ were transformed using the matrix square root of $R\hat{V}R'$, the quadratic form would lead to the construction of the classic F test based on a HAC estimate of V). We prove in Appendix B the following asymptotic result:

Theorem 10 *Suppose that Assumptions 1 and 2 hold. Then under the null hypothesis* $H_0 : R\beta = r$, $F^* \Rightarrow W_q(1)' P_q^{-1} W_q(1)/q$ *as* $T \to \infty$.

The limiting distribution of F^* is free of nuisance parameters and only depends on q. The distribution is nonstandard, but critical values can easily be simulated because the distribution is a function of independent standard Wiener processes. By approximating each Wiener process in the vector $W_q(r)$ using the same techniques that were used to simulate (4.7), critical values of $W_q(1)' P_q^{-1} W_q(1)/q$ were computed for $q = 1, 2, \ldots, 29, 30$ and are tabulated in Table 4.2. Since the distribution depends only on q, using Table 4.2 is no more difficult in practice than using a chi-square distribution table.

Construction of the F^* statistic amounts to replacing the HAC estimator, $\hat{\Omega}$, with $\hat{C}$ and using the scaling matrix $\hat{B}$ in place of the usual scaling matrix $\hat{V}$. The scaling matrix $\hat{B}$ converges to a random matrix rather than the fixed variance-covariance matrix. Viewed in this way, our approach creates a new class of statistics that are robust to serial correlation/heteroskedasticity in the errors and are asymptotically pivotal. In general, an asymptotically pivotal statistic can be obtained by replacing $\hat{\Omega}$ with any moment matrix of the data that has a limiting distribution of the form $\Lambda f(W_k(r))\Lambda'$ where $f(W_k(r))$ is a random matrix that is a functional of $W_k(r)$. Therefore, our particular choice of $\hat{C}$ is somewhat arbitrary and to some degree ad hoc, but it yields an elegant distribution theory with asymptotic distributions that do not depend on R, r or k. Other choices of $\hat{C}$ might not satisfy this property. In addition, we prove in Appendix B that our choice of $\hat{C}$ ensures that F^* is invariant to projecting out subsets of regressors,

Table 4.2: Asymptotic Critical values of F*

%/q	1	2	3	4	5	6	7	8	9	10
90.0	28.88	35.68	42.39	48.79	55.02	61.18	67.37	73.10	78.52	83.84
95.0	46.39	51.41	58.17	65.33	71.69	78.70	84.63	90.89	96.38	101.8
97.5	65.94	69.76	76.07	83.35	89.65	96.53	102.7	109.8	114.2	120.0
99.0	101.2	96.82	100.7	108.4	114.2	121.2	126.9	134.4	139.6	144.9
%/q	11	12	13	14	15	16	17	18	19	20
90.0	89.39	94.47	100.1	105.3	110.3	115.5	121.2	126.6	131.5	136.5
95.0	107.7	113.6	119.9	125.3	131.5	136.6	141.4	147.1	152.9	158.0
97.5	127.2	132.9	138.8	145.2	151.0	155.9	161.1	167.6	174.0	179.8
99.0	152.6	157.8	163.8	169.7	174.7	181.6	188.8	194.8	203.2	208.5
%/q	21	22	23	24	25	26	27	28	29	30
90.0	141.9	146.6	152.1	157.0	161.8	167.2	171.6	177.0	181.6	187.0
95.0	163.6	169.3	174.7	180.3	184.9	190.7	196.0	201.5	206.4	211.4
97.5	186.0	191.2	197.0	202.3	207.5	213.3	218.9	224.4	229.1	236.0
99.0	214.0	219.3	224.6	230.1	236.3	242.4	246.9	252.9	259.8	266.3

Note: q is the number of restrictions being tested.

i.e. F^* satisfies the Frisch-Waugh-Lovell (FWL) Theorem (see Davidson and MacKinnon (1993).[2] Also, F^* has the important practical property of invariance to rescaling of the regressors (i.e. invariance to units of measurement).[3] This discussion raises the natural question as to whether a theory of optimality can be created to help guide the choice of $\hat{C}$. We leave this interesting and challenging problem as an open research topic.

[2]HAC based tests satisfy the FWL Theorem only if a fixed truncation lag is used without prewhitening. Therefore, if an automatic truncation lag and/or prewhitening is used, different test statistics can result when one, say, first detrends regressors before estimating a regression as opposed to directly including a trend in the regression. See Section 7 for an example.

[3]As a referee pointed out, the F^* statistic is not invariant to the ordering of the observations (as is White's HC estimator). We anticipate F^* being used in time series settings where there is a natural ordering of the data. Should F^* be used in a pure cross section situation, ordering of the data becomes an issue.

4.4 Extensions to GLS and IV Estimation

In this section we briefly discuss how the F^* statistic can be applied to more general regression models which include GLS and IV estimation. Stack y_t, X_t, and u_t into matrices y, X, and u and consider a transformation of regression (4.1),

$$y^* = X^*\beta + u^*, \tag{4.9}$$

where $y^* = \Psi y$, $X^* = \Psi X$, $u^* = \Psi u$ and Ψ is a $(T \times T)$ transformation matrix. Estimating (4.9) by OLS is equivalent to minimizing $(y^* - X^*\beta)'\Psi'\Psi(y^* - X^*\beta)$. When $\Psi = \Sigma^{\frac{1}{2}}$ where $\Sigma = E(uu')$ we obtain the GLS estimate of β. When $\Psi = Z(Z'Z)^{-1}Z'$ where Z is a $(T \times m)$ vector of instruments with $m \geq k$ and $E(Z_t u_t) = 0$, we obtain the IV estimate of β. Provided that $v_t^* = X_t^* u_t^*$ and $T^{-1}\sum_{t=1}^{[rT]} X_t^* X_t^{*\prime}$ satisfy Assumptions 1 and 2, Theorem 10 still applies to F^* constructed from regression (4.9). In the case of IV estimation, sufficient conditions are stationary Z_t and $plim(T^{-1}\sum_{t=1}^{T} Z_t X_t') \neq 0$.

The natural extension beyond regression (4.9) is to consider a generalized methods of moments (GMM) framework which would include OLS, GLS and IV as special cases. This would be an important extension as GMM models are widely used in empirical macroeconomics. We conjecture that Theorem 10 generalizes to GMM models, but it is not clear whether standard GMM regularity conditions will be sufficient to obtain such a result. Furthermore, in overidentified GMM models, it is not obvious whether F^* should be constructed using sample analogs of the original moment conditions or sample analogs of the moment conditions implied by the weighted GMM minimization problem. It is unclear how the choice of weighting matrix will affect the asymptotic properties of F^*. An extension to GMM models is nontrivial and is beyond the scope of this chapter.

4.5 Local Asymptotic Power $k = 1$ Case

In this section we contrast the power properties of the t^* and HAC estimator t statistic using a local asymptotic framework. Of course, both tests have unit power against nonlocal alternatives. We restrict attention to the single regressor case ($k = 1$) as this special case sufficiently illustrates local asymptotic power comparisons. With $k = 1$ the regression becomes

$$y_t = \beta x_t + u_t, \tag{4.10}$$

where β and x_t are scalars and x_t is mean zero (this assumption has no effect on the local power results). We consider testing the null hypothesis $H_0 : \beta \leq \beta_0$ against the alternative $H_1' : \beta > \beta_0 + cT^{-\frac{1}{2}}$. Under the alternative, we model β as local to β_0 such that β converges to β_0 at rate $T^{-\frac{1}{2}}$ with local alternative parameter c. Let $\sigma_x^2 = E(x_t^2)$, and let $\sigma^2 = \gamma_0 + 2\sum_{j=1}^{\infty} \gamma_j$ where $\gamma_j = E(v_t v_{t-j})$ with $v_t = x_t u_t$. The parameter σ_x^2 is the variance of x_t, and the parameter σ^2 is equal to 2π times the spectral density of v_t evaluated at frequency zero. Define $\hat{\sigma}_x^2 = T^{-1}\sum_{t=1}^{T} x_t^2$ and let $\hat{\sigma}^2$ be a HAC estimator of σ^2 based on $\hat{v}_t = x_t \hat{u}_t$ where $\{\hat{u}_t\}$ are the OLS residuals from (10). Let $\hat{S}_t = \sum_{j=1}^{t} \hat{v}_j$ and define $\hat{C} = T^{-2}\sum_{t=1}^{T} \hat{S}_t^2$.

Using this notation, the HAC estimator t test, t_{HAC}, and t^* can be calculated as

$$t_{HAC} = T^{\frac{1}{2}}(\hat{\beta} - \beta_0)/(\hat{\sigma}_x^{-2}\hat{\sigma}^2\hat{\sigma}_x^{-2})^{\frac{1}{2}},$$

$$t^* = T^{\frac{1}{2}}(\hat{\beta} - \beta_0)/(\hat{\sigma}_x^{-2}\hat{C}\hat{\sigma}_x^{-2})^{\frac{1}{2}}.$$

In Appendix B we show under the local alternative and Assumptions 1 and 2, as $T \to \infty$,

$$t_{HAC} \Rightarrow c\sigma_x^2/\sigma + W_1(1) \sim N\left(c\sigma_x^2/\sigma, 1\right) \tag{4.11}$$

$$t^* \Rightarrow (c\sigma_x^2/\sigma + W_1(1))/[\int_0^1 (W_1(r) - rW_1(1))^2 dr]^{\frac{1}{2}}. \tag{4.12}$$

Results (4.11) and (4.12) show that local asymptotic power of both statistics depends on $c\sigma_x^2/\sigma$. Naturally, as c increases, power increases. As σ_x^2 increases, power also increases which follows from the standard regression result that more variability in the regressors leads to more efficient estimates. As σ^2 increases, power decreases which follows since variability in $\{u_t\}$ is higher.

The local asymptotic distributions were used to compute asymptotic power which is plotted in Figure 4.2. The power of t_{HAC} was computed analytically. The power of t^* was simulated using methods similar to those used to simulate the asymptotic critical values. Power was computed using the asymptotic 5% critical values.[4] As the figure shows, both statistics have monotonically increasing power functions. The power of t^* is nontrivial and is comparable to, but slightly below, that of t_{HAC}. In finite samples power of the tests is likely to be closer since the asymptotic power of t_{HAC} does not reflect the finite sample variability in $\hat{\sigma}^2$. In fact, we give examples in Section 4.7 where power of the $t^*(F^*)$ statistic exceeds the power of HAC estimator tests.

[4]We also computed asymptotic power for 1%, 2.5% and 10% significance levels. The relative power of the tests is similar to that depicted in Figure 1 and are available upon request.

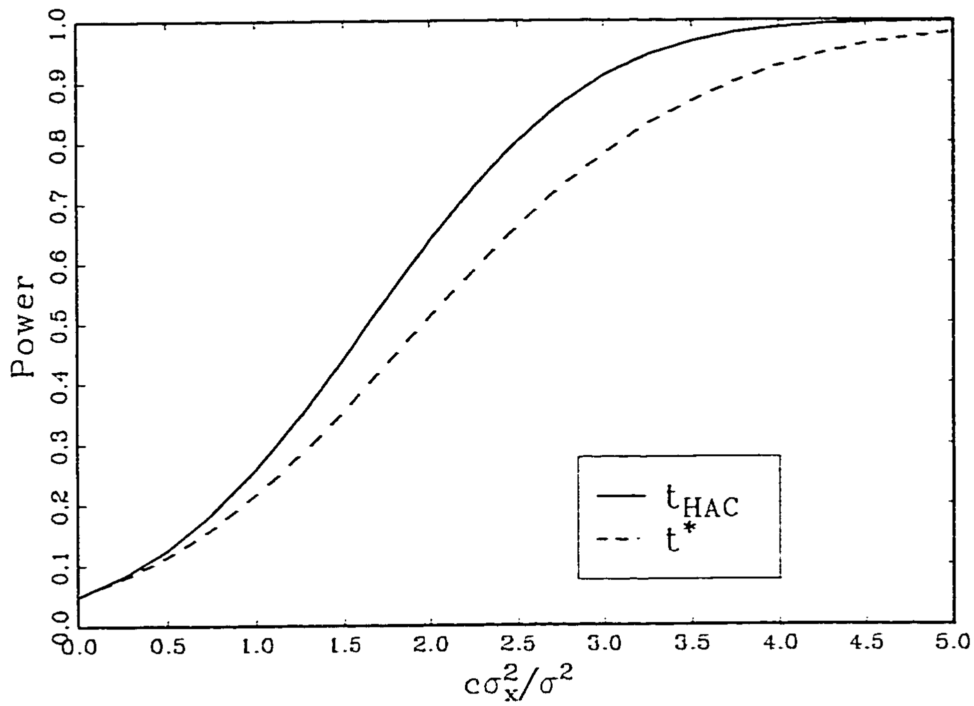

Figure 4.2: Local Asymptotic Power, 5% nominal size,
$k=1$, $y_t=\beta x_t+u_t$, H_0: $\beta \leqq \beta_0$, H_1: $\beta>\beta_0+cT^{-1/2}$

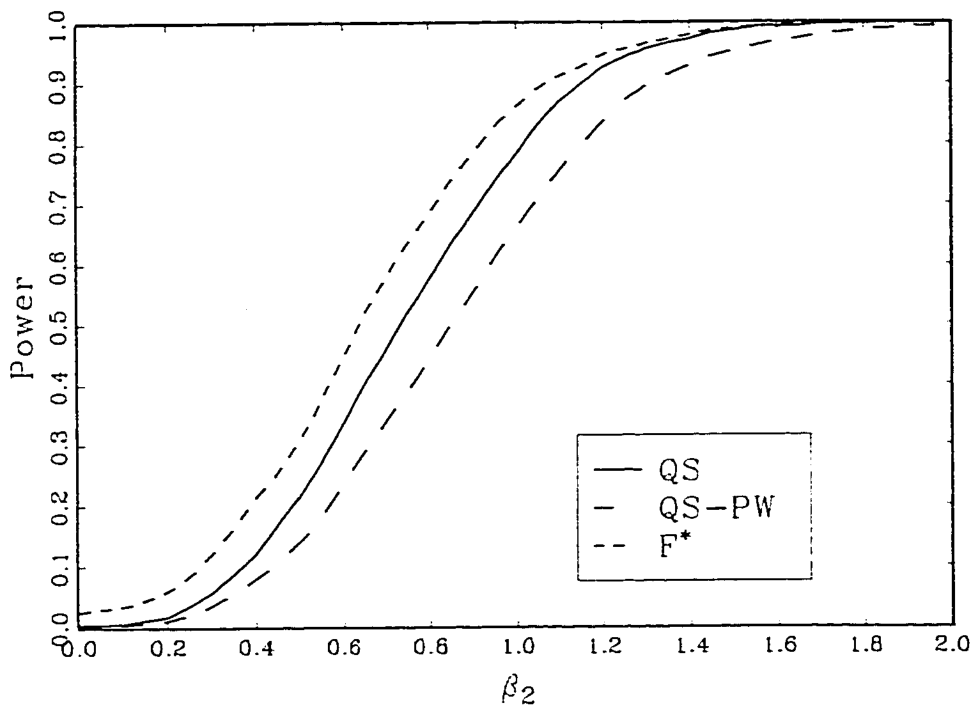

Figure 4.3: Finite Sample Power, T=103, 5% Nominal Size.
$y_t = \beta_1 + \beta_2 x_t + u_t$, H_0: $\beta_2=0$, $u_t \sim AR(4)$.
x_t is first difference of quarterly real GDP.

4.6 Finite Sample Size

In this section we report the results of an extensive simulation experiment with the purpose of comparing the finite sample size of HAC estimator tests and the F^* test. We designed the simulations so that they replicate the data generating processes (DGPs) and estimators used by Andrews (1991) and Andrews and Monahan (1992).

We consider a regression model with a constant and four stochastic regressors so that $k = 5$. We use the HAC estimator recommended by Andrews (1991) which uses the quadratic spectral kernel. The truncation lag (or bandwidth) was chosen using the automatic data-dependent procedure proposed by Andrews (1991) using the plug-in method based on univariate AR(1) models fit to the individual elements of $\hat{v}_t$. Tests based on this estimator are labeled *QS*. Consult Andrews (1991) for additional details. Following Andrews and Monahan (1992), we also computed HAC estimator tests using pre-whitening based on a VAR(1) parametric model of $\hat{v}_t$. We also employed the eigenvalue adjustment procedure used by Andrews and Monahan (1992) when fitting the VAR to $\hat{v}_t$. The pre-whitening tests are labeled *QS-PW*. Consult Andrews and Monahan (1992, p. 957) for additional details. Note that we are comparing our test with tests based on optimal HAC estimators.

We report results for six of the seven DGPs used by Andrews (1991) and Andrews and Monahan (1992). The models are: AR(1)-HOMO, where the errors and stochastic regressors are AR(1) homoskedastic processes; AR(1)-HET1 and AR(1)-HET2 where the DGPs are the same as the AR(1)-HOMO DGP except the error has multiplicative heteroskedasticity; MA(1)-HOMO, where the errors and stochastic regressors are MA(1) homoskedastic processes; and MA(1)-HET1 and

MA(1)-HET2 where the DGPs are the same as the MA(1)-HOMO DGP except the error has multiplicative heteroskedasticity. In all cases, the regressors and errors were drawn independently of each other. In the AR(1) models, the stochastic regressors and errors were generated according to the model $\eta_t = \rho\eta_{t-1} + e_t$ where e_t is drawn from i.i.d. $N(0,\ 1\text{-}\rho^2)$ random variables which results in η_t having unit variance. The initial condition was drawn from the stationary distribution of the AR(1) model. In each replication a new set of regressors were randomly drawn. We transformed the regressor matrix so that $T^{-1}\sum_{t=1}^{T} X_t X_t'$ is an identity matrix following Andrews and Monahan (1992, p. 959). For the HET1 and HET2 models, the errors were first drawn from the AR(1) process and then multiplied by $|X_{2t}|$ and $|\frac{1}{2}\sum_{i=2}^{5} X_{it}|$ respectively. We report results for ρ = -0.5, -0.3, 0.0, 0.3, 0.5, 0.7, 0.9, 0.95.

The MA(1) models were generated in a similar fashion with the stochastic regressors and errors generated according to the model $\eta_t = e_t + \theta e_{t-1}$ where e_t is drawn from i.i.d. $N(0,\ (1+\theta^2)^{-1})$ random variables which results in η_t having unit variance. We report results for θ = 0.3, 0.5, 0.7, 0.99. In all cases we used 2,000 replications.

Table 4.3: Finite Sample Null Rejection Probabilities AR(1) Models

Model	ρ	F^*	QS	QS-PW	Model	ρ	F^*	QS	QS-PW
	-0.50	0.067	0.094	0.079		-0.50	0.082	0.118	0.097
	-0.30	0.058	0.073	0.067		-0.30	0.062	0.090	0.080
AR(1)-	0.00	0.059	0.064	0.068	AR(1)-	0.00	0.054	0.070	0.069
HOMO	0.30	0.073	0.078	0.075	HOMO	0.30	0.065	0.090	0.085
$q = 1$	0.50	0.083	0.103	0.089	$q = 2$	0.50	0.090	0.134	0.109
	0.70	0.099	0.143	0.107		0.70	0.128	0.207	0.147
	0.90	0.197	0.302	0.211		0.90	0.273	0.440	0.322
	0.95	0.307	0.439	0.314		0.95	0.409	0.611	0.448
	-0.50	0.096	0.147	0.127		-0.50	0.104	0.184	0.153
	-0.30	0.074	0.109	0.107		-0.30	0.077	0.124	0.121
AR(1)-	0.00	0.059	0.083	0.093	AR(1)-	0.00	0.074	0.089	0.109
HOMO	0.30	0.071	0.115	0.108	HOMO	0.30	0.089	0.127	0.128
$q = 3$	0.50	0.097	0.169	0.131	$q = 4$	0.50	0.114	0.199	0.162
	0.70	0.141	0.262	0.195		0.70	0.169	0.313	0.237
	0.90	0.344	0.567	0.429		0.90	0.388	0.651	0.515
	0.95	0.491	0.748	0.570		0.95	0.543	0.832	0.658
	-0.50	0.075	0.108	0.093		-0.50	0.085	0.127	0.117
	-0.30	0.070	0.084	0.080		-0.30	0.071	0.091	0.093
AR(1)-	0.00	0.063	0.069	0.068	AR(1)-	0.00	0.062	0.087	0.086
HET1	0.30	0.075	0.093	0.089	HET1	0.30	0.073	0.099	0.097
$q = 1$	0.50	0.088	0.119	0.099	$q = 2$	0.50	0.095	0.136	0.118
	0.70	0.114	0.166	0.128		0.70	0.125	0.211	0.167
	0.90	0.217	0.338	0.267		0.90	0.283	0.450	0.352
	0.95	0.326	0.439	0.348		0.95	0.395	0.579	0.464
	-0.50	0.087	0.153	0.134		-0.50	0.098	0.174	0.157
	-0.30	0.065	0.100	0.103		-0.30	0.075	0.112	0.120
AR(1)-	0.00	0.068	0.088	0.101	AR(1)-	0.00	0.067	0.096	0.117
HET1	0.30	0.086	0.120	0.120	HET1	0.30	0.091	0.128	0.136
$q = 3$	0.50	0.10	0.165	0.142	$q = 4$	0.50	0.117	0.187	0.168
	0.70	0.14	0.262	0.208		0.70	0.163	0.317	0.242
	0.90	0.328	0.535	0.423		0.90	0.361	0.616	0.490
	0.95	0.449	0.687	0.546		0.95	0.506	0.769	0.616
	-0.50	0.073	0.090	0.078		-0.50	0.086	0.121	0.099
	-0.30	0.064	0.077	0.069		-0.30	0.070	0.085	0.087
AR(1)-	0.00	0.056	0.068	0.073	AR(1)-	0.00	0.069	0.077	0.079
HET2	0.30	0.068	0.086	0.082	HET2	0.30	0.077	0.097	0.094
$q = 1$	0.50	0.085	0.108	0.096	$q = 2$	0.50	0.089	0.130	0.116
	0.70	0.100	0.151	0.122		0.70	0.119	0.203	0.163
	0.90	0.203	0.305	0.234		0.90	0.257	0.421	0.318
	0.95	0.303	0.416	0.318		0.95	0.372	0.557	0.443

Table 4.3: (Continued)

Model	ρ	F*	QS	QS-PW	Model	ρ	F*	QS	QS-PW
	-0.50	0.083	0.142	0.124		-0.50	0.087	0.160	0.136
	-0.30	0.075	0.105	0.104		-0.30	0.079	0.115	0.117
AR(1)-	0.00	0.069	0.086	0.097	AR(1)-	0.00	0.077	0.090	0.110
HET2	0.30	0.079	0.114	0.114	HET2	0.30	0.083	0.123	0.131
$q = 3$	0.50	0.097	0.157	0.142	$q = 4$	0.50	0.101	0.196	0.170
	0.70	0.139	0.258	0.202		0.70	0.169	0.310	0.246
	0.90	0.311	0.529	0.406		0.90	0.351	0.610	0.489
	0.95	0.443	0.671	0.548		0.95	0.505	0.753	0.624

Following Andrews (1991) and Andrews and Monahan (1992) we computed type I error probabilities (they computed confidence interval coverage probabilities) for tests of the hypothesis $H_0 : \beta_2' = 0$. We extend the results of Andrews (1991) and Andrews and Monahan (1992) and also report results for tests of the hypotheses: $H_0 : \beta_2' = \beta_3 = 0$, $H_0 : \beta_2' = \beta_3 = \beta_4 = 0$, $H_0 : \beta_2' = \beta_3 = \beta_4 = \beta_5 = 0$. We label the hypotheses according to the number of restrictions being tested, i.e. $q = 1, 2, 3, 4$. The results for the AR(1) models with a sample size of $T = 128$ are reported in Table 4.3. Asymptotic critical values for the 0.05 nominal level were used.

Several patterns emerge from the table. First, in nearly every case, null rejection probabilities of F^* are less distorted and closer to 0.05 than the QS or QS-PW tests. The differences become larger as q increases. Although the F^* test has less distortions, there are many cases in which null rejection probabilities are much greater than 0.05. Nonetheless, the asymptotic approximation of the distribution of F^* is substantially better compared to QS and QS-PW. Second, as ρ approaches one, distortions of the null rejection probabilities increase for all the statistics. This is explained by the fact that the stationary asymptotic approximation becomes less accurate the closer the autoregressive root is to one. Third, for all three statistics, as q increases, null rejection probabilities also increase indicating the asymptotic approximation is less precise when testing joint hypotheses compared to testing simple hypotheses. This result suggests, in particular, that for joint hypotheses, size distortions of HAC estimator tests can be substantial even when there is only modest serial correlation in the errors.

Results for the MA(1) models with $T = 128$ are given in Table 4.4. Asymptotic critical values for the 0.05 nominal level were used. Similar patterns are seen as for the AR(1) models except that distortions overall are much less severe. Rejection probabilities of F^* are rarely above 0.10 while those of QS and $QS\text{-}PW$ often exceed 0.10 especially for large q.

In Table 4.5 we report results for the AR(1)-HOMO model for sample sizes $T = 256, 512$. Again asymptotic critical values for the 0.05 nominal level were used. The table indicates that the asymptotic approximation improves substantially for all the tests as T increases. For the most part, F^* has rejection probabilities close to 0.05 for $\rho \leq 0.5$. For $\rho > 0.5$ rejection probabilities are inflated but by much less compared to when $T = 128$. Rejection probabilities of QS and $QS\text{-}PW$ are, for the most part, more distorted than those of F^*, especially for $\rho \geq 0.9$ and $q \geq 3$.

4.7 Finite Sample Power and Empirical Example

Using the DGPs from the previous section, we simulated size-adjusted power of the statistics and found that power rankings of the statistics followed patterns qualitatively similar to the local asymptotic power curve depicted in Figure 4.2. Therefore, we do not report those simulations here and instead report results on fi-

Table 4.4: Finite Sample Null Rejection Probabilities MA(1) Models

Model	θ	F^*	QS	QS-PW	Model	θ	F^*	QS	QS-PW
MA(1)-	0.30	0.072	0.074	0.071	MA(1)-	0.30	0.068	0.087	0.080
HOMO	0.50	0.073	0.084	0.074	HOMO	0.50	0.074	0.100	0.081
$q = 1$	0.70	0.073	0.089	0.073	$q = 2$	0.70	0.078	0.109	0.082
	0.99	0.073	0.090	0.073		0.99	0.078	0.112	0.083
MA(1)-	0.30	0.066	0.106	0.104	MA(1)-	0.30	0.083	0.117	0.117
HOMO	0.50	0.074	0.124	0.105	HOMO	0.50	0.090	0.135	0.124
$q = 3$	0.70	0.080	0.134	0.102	$q = 4$	0.70	0.093	0.160	0.124
	0.99	0.082	0.139	0.101		0.99	0.094	0.164	0.118
MA(1)-	0.30	0.068	0.087	0.082	MA(1)-	0.30	0.072	0.093	0.090
HET1	0.50	0.082	0.098	0.086	HET1	0.50	0.074	0.103	0.089
$q = 1$	0.70	0.083	0.102	0.084	$q = 2$	0.70	0.082	0.115	0.095
	0.99	0.080	0.104	0.084		0.99	0.080	0.122	0.097
MA(1)-	0.30	0.080	0.118	0.117	MA(1)-	0.30	0.086	0.112	0.122
HET1	0.50	0.085	0.132	0.119	HET1	0.50	0.093	0.139	0.130
$q = 3$	0.70	0.095	0.140	0.116	$q = 4$	0.70	0.095	0.151	0.130
	0.99	0.095	0.146	0.116		0.99	0.093	0.158	0.130
MA(1)-	0.30	0.068	0.084	0.081	MA(1)-	0.30	0.073	0.088	0.089
HET2	0.50	0.077	0.096	0.084	HET2	0.50	0.072	0.103	0.091
$q = 1$	0.70	0.082	0.098	0.083	$q = 2$	0.70	0.077	0.110	0.088
	0.99	0.081	0.098	0.076		0.99	0.085	0.104	0.089
MA(1)-	0.30	0.078	0.102	0.105	MA(1)-	0.30	0.077	0.119	0.126
HET2	0.50	0.077	0.122	0.115	HET2	0.50	0.082	0.142	0.138
$q = 3$	0.70	0.082	0.133	0.109	$q = 4$	0.70	0.087	0.155	0.136
	0.99	0.086	0.132	0.106		0.99	0.097	0.156	0.125

Table 4.5: Finite Sample Null Rejection Probabilities AR(1)-HOMO Model

Model	ρ	F^*	QS	QS-PW	Model	ρ	F^*	QS	QS-PW
	-0.50	0.050	0.069	0.059		-0.50	0.062	0.089	0.071
	-0.30	0.051	0.068	0.061		-0.30	0.049	0.064	0.062
AR(1)-	0.00	0.044	0.057	0.058	AR(1)-	0.00	0.053	0.053	0.056
HOMO	0.30	0.052	0.066	0.061	HOMO	0.30	0.057	0.075	0.067
q=1	0.50	0.054	0.081	0.064	q=2	0.50	0.070	0.095	0.077
T=256	0.70	0.067	0.101	0.078	T=256	0.70	0.095	0.137	0.106
	0.90	0.123	0.191	0.141		0.90	0.170	0.289	0.197
	0.95	0.184	0.297	0.207		0.95	0.263	0.442	0.311
	-0.50	0.063	0.104	0.080		-0.50	0.072	0.120	0.095
	-0.30	0.048	0.075	0.071		-0.30	0.056	0.086	0.078
AR(1)-	0.00	0.052	0.060	0.064	AR(1)-	0.00	0.057	0.064	0.068
HOMO	0.30	0.066	0.082	0.079	HOMO	0.30	0.067	0.090	0.086
q=3	0.50	0.073	0.101	0.096	q=4	0.50	0.084	0.132	0.106
T=256	0.70	0.100	0.173	0.125	T=256	0.70	0.122	0.202	0.146
	0.90	0.213	0.386	0.263		0.90	0.250	0.477	0.342
	0.95	0.330	0.565	0.406		0.95	0.394	0.682	0.502
	-0.50	0.062	0.070	0.057		-0.50	0.059	0.080	0.070
	-0.30	0.058	0.060	0.053		-0.30	0.047	0.061	0.056
AR(1)-	0.00	0.055	0.054	0.056	AR(1)-	0.00	0.047	0.053	0.056
HOMO	0.30	0.057	0.063	0.057	HOMO	0.30	0.058	0.062	0.058
q=1	0.50	0.047	0.067	0.062	q=2	0.50	0.058	0.073	0.060
T=512	0.70	0.064	0.081	0.064	T=512	0.70	0.065	0.097	0.072
	0.90	0.092	0.125	0.084		0.90	0.105	0.170	0.118
	0.95	0.124	0.193	0.132		0.95	0.165	0.278	0.195
	-0.50	0.057	0.087	0.073		-0.50	0.057	0.091	0.073
	-0.30	0.049	0.070	0.067		-0.30	0.045	0.076	0.068
AR(1)-	0.00	0.045	0.061	0.060	AR(1)-	0.00	0.060	0.063	0.066
HOMO	0.30	0.053	0.066	0.060	HOMO	0.30	0.057	0.082	0.071
q=3	0.50	0.050	0.077	0.065	q=4	0.50	0.054	0.089	0.071
T=512	0.70	0.068	0.104	0.077	T=512	0.70	0.073	0.117	0.091
	0.90	0.120	0.225	0.156		0.90	0.142	0.268	0.191
	0.95	0.194	0.365	0.240		0.95	0.232	0.452	0.309

nite sample power from simulations based on the following empirical example. Let $\Delta lrev_t$ denote the first difference of the natural logarithm of real aggregate restaurant revenues for the United States, and let $\Delta lgdp_t$ denote the first difference of the natural logarithm of (seasonally adjusted) real gross domestic product (GDP) for the United States. We obtained quarterly observations from 1971:1 to 1996:4 for the nominal versions of these series and constructed the real series by dividing by the implicit GDP deflator. We seasonally adjusted the nominal restaurant revenue series before constructing the real series. The restaurant revenue series was obtained from the *Current Business Reports* published by the Bureau of the Census, and the nominal GDP and deflator series were obtained from the *Survey of Current Business* published by the Bureau of Economic Analysis, U.S. Department of Commerce. The levels of the real revenue and real GDP series are clearly trending over time and may have unit root errors. Therefore, the first differences of the series are likely to be stationary and satisfy Assumptions 1 and 2, so we consider a regression model in first differences of the data. For simplicity, we are ignoring the possibility that the levels of the series are cointegrated.

Consider the regression

$$\Delta lrev_t = \beta_1 + \beta_2 \Delta lgdp_t + u_t. \tag{4.13}$$

In the notation of Section 4.2, $\beta = (\beta_1', \beta_2')'$ and $X_t = (1, \Delta lgdp_t)'$. The parameter β_2' measures the change in the growth of real restaurant revenues with respect to a unit increase in the real growth rate of GDP. Thus, β_2' measures the sensitivity of real restaurant revenue growth to changes in real GDP growth. Since shocks to the restaurant sector are likely to have little or no effect on GDP, it is reasonable to think of $\Delta lgdp_t$ as an exogenous regressor. Therefore, OLS provides a consistent estimate of β_2'.

We estimated (4.13) by OLS and obtained $\hat{\beta}_2' = 0.681$. Thus, an increase in the real growth rate of GDP by 1% results in a 0.681% increase in the real growth rate of restaurant revenues. To measure the sampling variability of $\hat{\beta}_2'$, we constructed the following 95% confidence intervals: QS: (0.059, 1.302), QS-PW: (-0.070, 1.431) and F^* : (0.305, 1.056). Confidence intervals based on QS and QS-PW were computed as $\hat{\beta}_2' \pm 1.96(\hat{V}_{22}/T)^{\frac{1}{2}}$ where $\hat{V}_{22}$ is the second diagonal element of $\hat{V} = (T^{-1}\sum_{t=1}^{T} X_t X_t')^{-1}\hat{\Omega}(T^{-1}\sum_{t=1}^{T} X_t X_t')^{-1}$, $\hat{\Omega}$ is the QS or QS-PW HAC estimator respectively of Ω, 1.96 is 97.5% critical value of a standard normal distribution, and $T = 103$. The confidence interval based on F^* was computed as $\hat{\beta}_2' \pm 6.811(\hat{B}_{22}/T)^{\frac{1}{2}}$ where $\hat{B}_{22}$ is the second diagonal element of $\hat{B} = (T^{-1}\sum_{t=1}^{T} X_t X_t')^{-1}(T^{-2}\sum_{t=1}^{T} \hat{S}_t \hat{S}_t')(T^{-1}\sum_{t=1}^{T} X_t X_t')^{-1}$ and 6.811 is the 97.5% asymptotic critical value taken from Table 4.1. Interestingly, the tightest confidence interval is obtained using F^*, and there are nontrivial differences in the HAC based confidence intervals whether or not prewhitening is used. This empirical example suggests a situation where power of the F^* statistic may be greater than power of HAC estimator tests and illustrates the sensitivity of inference to the way HAC estimators are constructed (see Den Hann and Levin (1997) for additional evidence on the latter).

To investigate the possibility that F^* is more powerful in the empirical example, we conducted the following simulation experiment. We fit a variety of ARMA models to the OLS residuals from (13) and found that an AR(4) model provided a good fit. We also fit a variety of ARMA models to $\Delta lgdp_t$ and found that an AR(1) and an ARMA(4,1) model provided good fits. We performed three power simulations using 2,000 replications and $T = 103$. We generated data according the model $y_t = \beta_2' x_t + u_t$ with $u_t = -0.3429u_{t-1} - 0.3301u_{t-2} -$

$0.2686u_{t-3} + 0.5947u_{t-4} + \epsilon_t$, $\epsilon_t \sim$ i.i.d $N(0, 0.0197)$. We generated x_t using three DGPs: DGP(1): x_t equal to the actual first differenced quarterly real GDP data ($x_t = \Delta lgdp_t$), DGP(2): $x_t = 0.3249x_{t-1} + \xi_t$, $\xi_t \sim$ i.i.d $N(0, 0.0089)$ and DGP(3): $x_t = 0.9952x_{t-1} - 0.1446x_{t-2} + 0.0411x_{t-3} - 0.1465x_{t-4} + \xi_t - 0.7410\xi_{t-1}$, $\xi_t \sim$ i.i.d $N(0, 0.0089)$. The null hypotheses was $H_0 : \beta_2' = 0$.

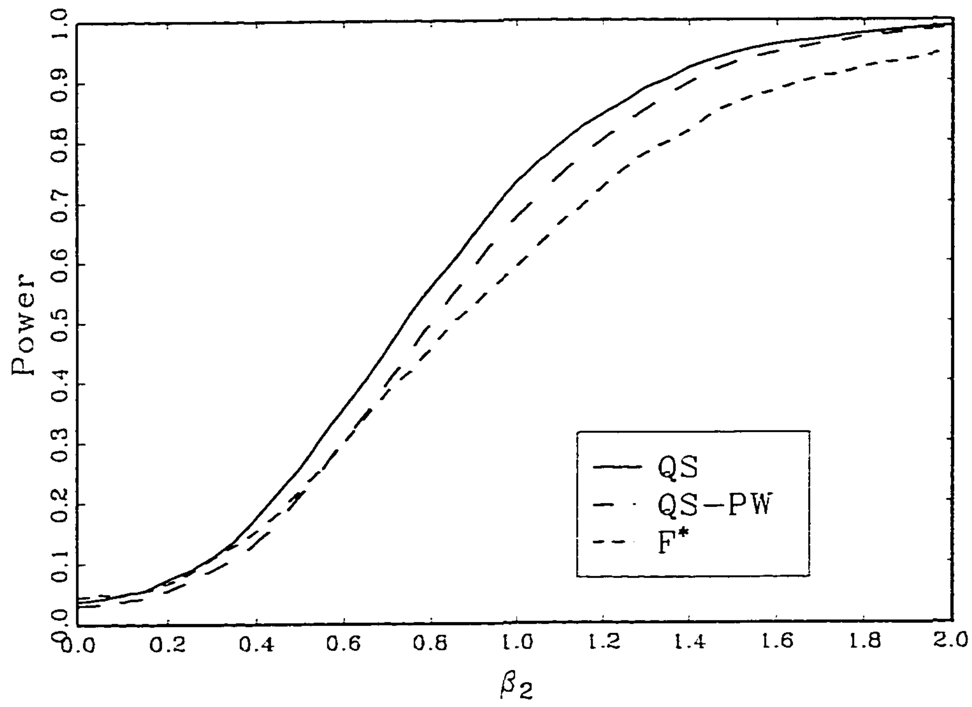

Figure 4.4: Finite Sample Power, T=103, 5% Nominal Size.
$y_t = \beta_1 + \beta_2 x_t + u_t$, H_0: $\beta_2=0$, $u_t \sim AR(4)$, $x_t \sim AR(1)$.

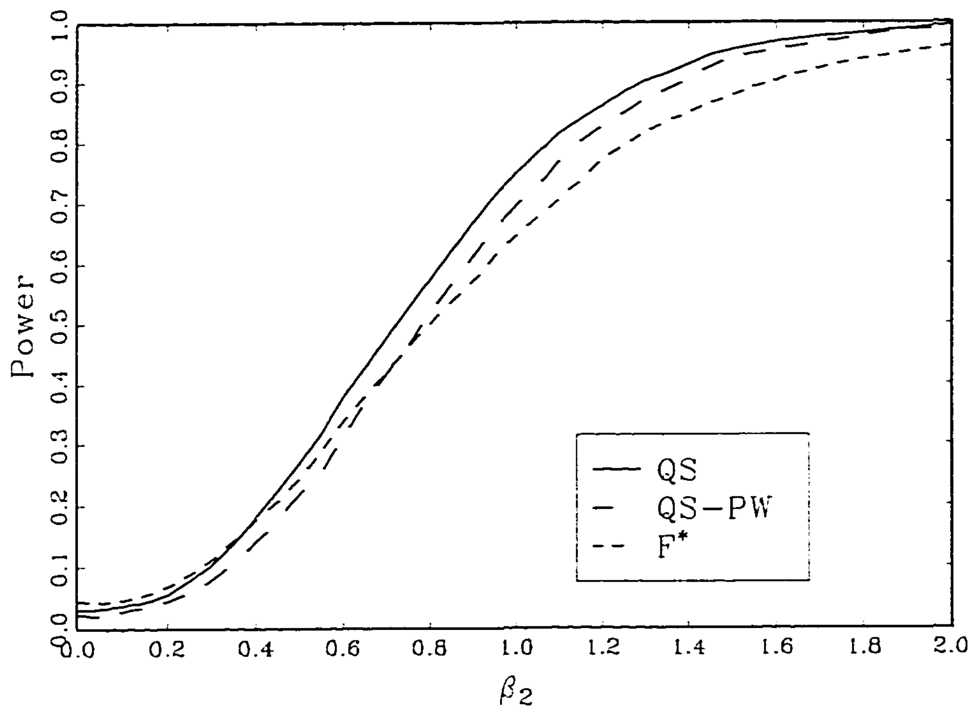

Figure 4.5: Finite Sample Power, T=103, 5% Nominal Size.
$y_t = \beta_1 + \beta_2 x_t + u_t$, H_0: $\beta_2=0$, $u_t \sim AR(4)$, $x_t \sim ARMA(4,1)$.

Table 4.6: Finite Sample Size

	β_2'	F^*	QS	QS-PW
DGP(1)	0.00	0.026	0.005	0.002
DGP(2)	0.00	0.044	0.038	0.030
DGP(3)	0.00	0.044	0.030	0.023

We computed finite sample null rejection probabilities of the statistics using 5% asymptotic critical values which we report in Table 4.6. Regardless of the DGP used for x_t, rejection probabilities of all the statistics are below 0.05 with those of QS and $QS\text{-}PW$ below that of F^*. Rejection probabilities of QS and $QS\text{-}PW$ are quite low when the actual GDP data is used (DGP(1)) which explains the wide confidence intervals using QS and $QS\text{-}PW$ in the empirical example. Because the tests are all conservative, it makes sense to compare power functions computed using the asymptotic critical values (this mimics the way the tests are used in practice). We simulated power for $\beta_2' = 0.1, 0.2, ..., 1.9, 2.0$. The resulting finite sample power curves are plotted in Figures 4.5, 4.6 and 4.7 corresponding to the three DGPs for x_t.. In Figure 4.5 we see that F^* dominates the HAC estimator tests in terms of power when actual GDP data is used for x_t. This is not an atypical example as real GDP is commonly used in empirical work. In the other cases where x_t is modeled as an ARMA process, the power ranking depends on how far β_2' is from zero.

We conclude this section with an example which illustrates the sensitivity of HAC estimator tests to projections of subsets of regressors in OLS regressions. Using the same data set as above, we regressed the level of nominal aggregate restaurant revenue on a constant, a time trend, and the level of nominal GDP and obtained the following 95% confidence intervals for the estimate of the coefficient on nominal GDP: QS: (1.077, 1.321), $QS\text{-}PW$: (1.040, 1.358), F^* : (1.031, 1.366).

(Because the nominal series almost certainly have unit root errors, this example does not satisfy Assumptions 1 and 2. It is illustrative nonetheless). We also detrended the data and regressed the detrended revenue series on the detrended constant and detrended GDP (we projected out the time trend). The OLS estimate of the GDP coefficient and F^* are invariant to the method of estimation by the FWL Theorem. Therefore, confidence intervals based on F^* are the same in the two cases. Confidence intervals based on QS and $QS\text{-}PW$ are not invariant which is illustrated by confidence intervals based on the detrended regression: QS: (1.069, 1.329), $QS\text{-}PW$: (0.795, 1.603). This lack of invariance arises from using an automatic bandwidth and/or pre-whitening and illustrates a pitfall when using HAC estimator tests.

4.8 Conclusions

In this chapter we propose new test statistics for testing hypotheses in regression models with serial correlation/heteroskedasticity of unknown form. The novel aspect of the new tests is that they are simple to compute and do not require spectral density (HAC) estimators. Our approach is to eliminate nuisance parameters asymptotically with a simple stochastic transformation of the parameter estimates. Since there are many conceivable transformations that will yield asymptotic pivotal statistics, our approach creates a new class of test statistics which are pivotal and robust to heteroskedasticity and serial correlation in the errors. An open research problem is to develop a theory of optimality for this new class of tests. We derived the limiting null distributions of the new tests and showed that while they have nonstandard distributions, the distributions only depend on the number of restrictions being tested and critical values were easily simulated. Our results easily extend to GLS and IV estimation, and we conjecture that our approach can be extended to the GMM framework. A simulation experiment showed that the asymptotic approximation of the new test is better (nearly uniformly) than that of more standard HAC estimator tests. But, like HAC estimator tests, the new tests suffer from serious size distortions (although less so) if the data has highly persistent serial correlation and is close to being nonstationary. This is a common problem in time series models when the true form of serial correlation is unknown. Finally, the new tests retain respectable power, and we provide a relevant empirical example where finite sample power of our test dominates finite sample power of HAC estimator tests. Given that new tests compare favorably to HAC methods in finite samples and are simpler to compute, they should become serious competitors to HAC estimator tests in practice.

Chapter 5

Simple Robust Testing of Hypotheses in Non-Linear Models

5.1 Introduction

It is a well known result that in models with autocorrelation/heteroskedasticity of unknown form, standard estimators remain consistent and are asymptotically normally distributed under weak regularity conditions. However, the usual results required for testing hypotheses in the usual manner no longer holds. In this chapter we develop new hypothesis tests in weighted, nonlinear regression models with serial correlation/heteroskedasticity of unknown form. Included in this class of models are non-linear GLS, IV-estimation in nonlinear models, and some Quasi-likelihood models.

When the entire covariance structure is known, the model can be transformed and standard testing results can be obtained using GLS methods. This is usually not possible in practice, as the serial correlation or heteroskedasticity encountered is frequently of unknown form. To obtain valid testing procedures, the most common approach in the literature to date has been to estimate the variance-covariance matrix of the parameter estimate. This is usually done nonparametrically, using spectral methods which lead to heteroskedasticity and autocorrelation consistent (HAC) estimators. Using these estimators, standard tests are constructed based on the asymptotical normal distribution of the of the weighted NLS estimator. HAC estimators and their properties have recently attracted a lot of attention in the literature. Among the important contributions are Andrews (1991), Andrews and Monahan (1992), Gallant (1987), Hansen (1992), Newey and West (1987), Robinson (1991) and White (1984). The direct contribution of this literature has been the construction of asymptotically valid tests that are robust to serial correlation/heteroskedasticity of unknown form.

The main limitation of the HAC approach is that, while the variance-covariance matrix is estimated, the resulting variation in finite samples is not taken into account. Asymptotically, this clearly is not a problem; in fact, once the variance-covariance matrix has been estimated, it can be assumed to be known. In finite samples, however, this can cause substantial size distortions. In this chapter, we develop an alternative method of creating hypothesis tests that are robust to serial correlation or heteroskedasticity of unknown form, and which do not require a direct estimate of the variance-covariance matrix.

The approach we take is similar to Fisher's classic construction of the t test. A data-dependent transformation is applied to the NLS estimates of the parameters of interest. This transformation is chosen such that it ensures that the asymptotic distribution of the transformed estimator does not depend on nuisance parameters. The transformed estimator can then be used to construct a test for general hypotheses on the parameters of interest. The asymptotic distribution of the resulting test statistic turns out to be symmetric, but with fatter tails than the normal distribution; it is not normal, but has the form of a scale mixture of normals. Furthermore, it depends only on the number of restrictions that are being tested. We are therefore able to tabulate the critical values in the usual manner; as a function of the number of restrictions and the level of the test.

We provide an empirical example illustrating this new test statistic. Specifically, we examine the effect of GDP growth on the growth of total restaurant revenues. We use quarterly data over 26 years for the analysis, and there is reason to suspect the presence of autocorrelation and heteroskedasticity. We do not, however, have any knowledge of the specific forms of autocorrelation and heteroskedasticity we may encounter in this data set, hence making it an excellent candidate for the application of both the HAC estimators and the newly introduced test statistic.

We estimate the long-run relationship between growth of restaurant revenues and growth of GDP using our method and the HAC estimator with and without prewhitening (as introduced by Andrews and Monahan, 1992). We then perform simulations, confirming that the size of our test is less distorted than that of the tests currently in use. We also examine finite sample power for the different methods, and find that power of our test can dominate HAC estimator tests.

The rest of the chapter is organized as follows. In section 5.2, we introduce the model and prove basic asymptotic results. In section 5.3, we develop the test statistic, first for simple, 1-dimensional hypotheses and then for general, non-linear hypotheses. In section 5.4, we describe the empirical example and the simulation based on it. Section 5.5 concludes. Some proofs are included in Appendix C.

5.2 The Model and Some Asymptotic Results

Consider the nonlinear regression model given by

$$y_t = f(X_t, \beta) + u_t = f_t(\beta) + u_t; \ t = 1, ..., T, \tag{5.1}$$

where f denotes the nonlinear function of regressors and parameters. β is a $(k \times 1)$ vector of parameters and X_t is a $(k_2 \times 1)$ vector of exogenous variables and conditional on X_t, u_t is a mean zero random process. We assume that u_t does not have a unit root, but u_t may be serially correlated or heteroskedastic. At times, it will be useful to stack the equations in (5.1) and rewrite it as

$$y = \mathbf{f}(\beta) + u. \tag{5.2}$$

We will use weighted non-linear least squares to obtain an estimate of β. The estimate, $\hat{\beta}$, is defined as

$$\hat{\beta} = \arg\min_{\beta} (y - \mathbf{f}(\beta))' W (y - \mathbf{f}(\beta)) \tag{5.3}$$

where W is a symmetric, positive definite $T-$dimensional weighting matrix. Depending on the choice of W, the following are examples of estimation techniques covered by this framework:

Example 1: *Nonlinear Least Squares.*

If we let W be the identity matrix, (5.3) takes the well-known form

$$\hat{\beta} = \arg\min_{\beta} (y - \mathbf{f}(\beta))' (y - \mathbf{f}(\beta)).$$

This is the case of standard, non-linear least squares.

Example 2: *Non-linear IV estimation, Lagged Dependent Variables.*

If we have a model corresponding to (5.2), and a $T \times l$ matrix of instruments Z, $l \geq k$, with the matrix projecting onto the space spanned by the instruments defined as $P_Z = Z(Z'Z)^{-1} Z'$, the IV estimator takes the form

$$\hat{\beta}_{IV} = \arg\min_{\beta} (y - \mathbf{f}(\beta))' P_Z (y - \mathbf{f}(\beta))$$

corresponding to $W = P_Z$.

A special case of this is models including lagged dependent variables Here we are interested in a model

$$Y = X\beta + Y_2\Gamma + U, \tag{5.4}$$

where Y_2 is a matrix containing lagged values of Y and β and Γ are parameters. If we let $A = [X,\ Y_2]$ and $\delta = (\beta,\ \Gamma)'$, we can rewrite (5.4) as

$$Y = A\delta + U.$$

Using the method of instrumental variables with instruments Z, the estimate of δ is defined as

$$\hat{\delta} = \arg\min_{\delta} (Y - A\delta)' P_Z (Y - A\delta) \tag{5.5}$$

where $P_J = Z(Z'Z)^{-1} Z'$. Comparing (5.5) and (5.3), we see that $\hat{\delta}$ is the weighted least squares estimator in a model with weighting matrix P_Z.

These examples illustrate that several well-known models and estimation techniques are special cases of the framework we use.

In developing the results, we also work with the following transformed model

$$W^{\frac{1}{2}}y = W^{\frac{1}{2}}\mathbf{f}(\beta) + W^{\frac{1}{2}}u$$

or, simplifying the notation

$$\tilde{y} = \tilde{\mathbf{f}}(\beta) + \tilde{u} \tag{5.6}$$

where $\tilde{y}_t$, $\tilde{f}_t$ and $\tilde{u}_t$ are defined in the natural way. The following additional notation is used throughout the chapter. Let $\tilde{F}_t(\beta) : k \times 1$ denote the derivative of $\tilde{f}_t(\beta)$ with respect to β and $\mathbf{F}(\beta)$ the derivative of $\mathbf{f}(\beta)$ with respect to β. In addition, let $v_t \equiv \tilde{F}_t(\beta)\tilde{u}_t$ and define $\Omega = \Lambda\Lambda' = \Gamma_0 + \sum_{j=0}^{\infty}(\Gamma_j + \Gamma_j')$ where $\Gamma_j = E(v_t v_{t-j})$. For later use, note that Ω is equal to 2π times the spectral density matrix of v_t evaluated at frequency zero. Define $S_t = \sum_{j=1}^{t} v_j$ and let $W_k(r)$ denote a k-vector of independent Wiener precesses, and let $[rT]$ denote the integer part of rT, where $r \in [0,1]$. We let β_0 denote the true value of the parameter β. We use $\Rightarrow$ to denote weak convergence.

The following two assumptions will be sufficient to obtain the main results of the chapter.

Assumption 3 $\text{plim}\left[T^{-1}\sum_{t=1}^{[rT]}\tilde{F}_t(\beta_0)\tilde{F}_t'(\beta_0)\right] = rQ$, where Q is invertible.

Assumption 4 $T^{-\frac{1}{2}}S_{[rT]} = T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\tilde{F}_t(\beta_0)\tilde{u}_t = T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}v_t \Rightarrow \Lambda W_k(r)$.

Assumption 3 rules out trends in the *linearized* regression function of the transformed model, but not necessarily in the X_t process. Note that Assumption 3 implies the standard assumption, namely $\text{plim}[\frac{1}{T}\mathbf{F}'(\beta_0)W\mathbf{F}(\beta_0)] = Q$. Assumption 4 states that a functional central limit theorem holds for the sequence $\{v_t\}$. This is the case, for example, if v_t is weakly stationary, the elements of v_t have a finite moment greater that 2 and $\{u_t\}$ and $\{X_t\}$ satisfy well-known mixing conditions. One set of conditions under which Assumption 4 holds can be found in Phillips and Durlauf (1986).

Using Assumptions 3 and 4, we can obtain the well known asymptotic distribution of $\hat{\beta}$:

$$\sqrt{T}\left(\hat{\beta} - \beta\right) \Rightarrow Q^{-1}\Lambda W_k(1) \sim N\left(0, Q^{-1}\Lambda\Lambda' Q^{-1}\right) = N\left(0, Q^{-1}\Omega Q^{-1}\right) = N(0, V). \tag{5.7}$$

For the derivation of (5.7) see Davidson MacKinnon (1993). We see that the asymptotic distribution of $\hat{\beta}$ is a k-variate normal distribution with mean β and variance covariance matrix $V = Q^{-1}\Omega Q^{-1}$. The asymptotic distribution of $\hat{\beta}$ can now be used to test hypothesis on β. To do this, an estimate of V (and therefore Q^{-1} and Ω) is required. A natural estimate of Q^{-1} is $\left[\frac{1}{T}\mathbf{F}'\left(\hat{\beta}\right)W\mathbf{F}\left(\hat{\beta}\right)\right]^{-1}$. Ω can be estimated by a HAC estimator, $\hat{\Omega}$. Letting $\hat{u}_t$ be the residuals of the transformed model, the HAC estimate would utilize $\hat{v}_t = \tilde{F}_t\left(\hat{\beta}\right)\hat{u}_t$ to estimate nonparametrically the spectral density of v_t at frequency zero, and hence Ω. To test hypotheses on β using $\hat{V} = \left[\frac{1}{T}\mathbf{F}'\left(\hat{\beta}\right)W\mathbf{F}\left(\hat{\beta}\right)\right]^{-1}\hat{\Omega}\left[\frac{1}{T}\sum_{j=1}^{T}\mathbf{F}'\left(\hat{\beta}\right)W\mathbf{F}\left(\hat{\beta}\right)\right]^{-1}$, transform $\sqrt{T}\left(\hat{\beta}-\beta_0\right)$ to obtain

$$\hat{V}^{-\frac{1}{2}}\sqrt{T}\left(\hat{\beta}-\beta_0\right) \Rightarrow \Omega^{-\frac{1}{2}}QQ^{-1}\Lambda W_k(1) = W_k(1) \sim N(0, I_k). \tag{5.8}$$

Using (5.8), hypotheses can be tested in the usual manner with a t-test.

To test hypotheses on β, we use a method that is similar; we also transform $\sqrt{T}\left(\hat{\beta}-\beta_0\right)$ in such a manner that the asymptotic distribution no longer depends on unknown parameters. The essential difference between the two approaches is that our approach does not require an explicit estimate of Ω and takes the additional sampling variation associated with not knowing the covariance matrix into account. HAC estimates, on the other hand, treat the variance-covariance matrix as known asymptotically.

We now proceed to obtain the relevant transformation. Consider $T^{-\frac{1}{2}}\hat{S}_{[rT]} = T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\hat{v}_t = T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\tilde{F}_t\left(\hat{\beta}\right)\hat{u}_t$. In Appendix C, we prove the following lemma:

Lemma 1

$$T^{-\frac{1}{2}}\hat{S}_{[rT]} = T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\hat{v}_t = T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\tilde{F}_t\left(\hat{\beta}\right)\hat{u}_t = T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\tilde{F}_t(\beta_0)\hat{u}_t + l_{[rT]}$$

where $l_{[rT]}$ is a residual term, with the property that

$$plim\left(l_{[rT]}\right) = 0$$

We now use the asymptotic distribution of $\hat{\beta}$, specified in (5.7) together with Assumptions 3 and 4, to determine the asymptotic distribution of $T^{-\frac{1}{2}}\hat{S}_{[rT]}$.

$$\begin{aligned} T^{-\frac{1}{2}}\hat{S}_{[rT]} &= T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\tilde{F}_t(\beta_0)\hat{u}_t + l_{[rT]} \\ &= T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\tilde{F}_t(\beta_0)\left[\tilde{f}_t(\beta_0)+\tilde{u}_t-\tilde{f}_t\left(\hat{\beta}\right)\right] + l_{[rT]} \\ &= T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\tilde{F}_t(\beta_0)\tilde{u}_t + T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\tilde{F}_t(\beta_0)\left[\tilde{f}_t(\beta_0)-\tilde{f}_t\left(\hat{\beta}\right)\right] + l_{[rT]} \end{aligned}$$

Using a Taylor expansion of $\tilde{f}_t(\hat{\beta})$ around β_0, we see $\tilde{f}_t(\hat{\beta}) = \tilde{f}_t(\beta_0) + \tilde{F}_t'(\beta_0)(\hat{\beta} - \beta_0) + \tilde{l}_t$, where $\tilde{l}_t$ represents the higher order terms of the expansion. From the assumptions at the beginning of this section, it is clear that $\tilde{l}_t$ is $O_P(T^{-1})$; hence $T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\tilde{F}_t(\beta_0)\tilde{l}_t$ will be $O_P\left(T^{-\frac{1}{2}}\right)$, implying that the term can be ignored. This allows us to write

$$\begin{aligned} T^{-\frac{1}{2}}\hat{S}_{[rT]} &= T^{-\frac{1}{2}}S_{[rT]} - T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\tilde{F}_t(\beta_0)\left[\tilde{F}_t'(\beta_0)\left(\hat{\beta}-\beta_0\right)+\tilde{l}_t\right] + l_{[rT]} \qquad (5.9) \\ &= T^{-\frac{1}{2}}S_{[rT]} - r\left[\frac{1}{[rT]}\sum_{t=1}^{[rT]}\tilde{F}_t(\beta_0)\tilde{F}_t'(\beta_0)\right]\left[T^{\frac{1}{2}}\left(\hat{\beta}-\beta_0\right)\right] + o_P(1) \\ &\Rightarrow \Lambda W_k(r) - rQ\left[Q^{-1}\Lambda W_k(1)\right] = \Lambda\left(W_k(r) - rW_k(1)\right) \end{aligned}$$

Note that $(W_k(r) - rW_k(1))$ is a k-dimensional Brownian Bridge.

Now consider $\hat{C} \equiv T^{-2}\sum_{t=1}^{T}\hat{S}_t\hat{S}_t'$. From (5.9) and the continuous mapping theorem, it follows that

$$\begin{aligned} \hat{C} &= T^{-1}\sum_{t=1}^{T}\left[T^{-\frac{1}{2}}\hat{S}_t\right]\left[T^{-\frac{1}{2}}\hat{S}_t\right]' \\ &\Rightarrow \Lambda\int_0^1 (W_k(r) - rW_k(1))(W_k(r) - rW_k(1))'\,dr\Lambda'. \end{aligned}$$

Define $P_k \equiv \int_0^1 (W_k(r) - rW_k(1))(W_k(r) - rW_k(1))' dr$. The asymptotic distribution of $\hat{C}$ can be written as $\Lambda P_k \Lambda'$. We have now obtained a matrix whose asymptotic distribution is a quadratic form in Λ. In what follows, this will enable us to make a transformation eliminating Λ from the asymptotic distribution of the test statistic.

To that end, note that because P_k is constructed as the integral of the outer product of a k-dimensional Brownian Bridge, it is positive definite. This permits us to use the Cholesky decomposition to write $P_k = Z_k Z_k'$.

To eliminate Q from the asymptotic distribution as well, we now turn our attention to the following matrix
$\hat{B} = \left[\frac{1}{T}\mathbf{F}'\left(\hat{\beta}\right) W\mathbf{F}\left(\hat{\beta}\right)\right]^{-1} \hat{C} \left[\frac{1}{T}\mathbf{F}'\left(\hat{\beta}\right) W\mathbf{F}\left(\hat{\beta}\right)\right]^{-1}$. Now define
$\hat{M} = \left[\frac{1}{T}\mathbf{F}'\left(\hat{\beta}\right) W\mathbf{F}\left(\hat{\beta}\right)\right]^{-1} \hat{C}^{\frac{1}{2}}$, where $\hat{C}^{\frac{1}{2}}$ is the Cholesky decomposition of $\hat{C}$. Therefore, $\hat{M}\hat{M}' = \hat{B}$. Note that since $\hat{B} \Rightarrow Q^{-1}\Lambda P_k \Lambda' Q^{-1}$, $\hat{M} \Rightarrow Q^{-1}\Lambda Z_k$. We are now ready to examine a transformation of $\sqrt{T}\left(\hat{\beta} - \beta\right)$, namely $\hat{M}^{-1}\sqrt{T}\left(\hat{\beta} - \beta\right)$.

$$\hat{M}^{-1}\sqrt{T}\left(\hat{\beta} - \beta\right) \Rightarrow \left[Q\Lambda Z_k\right]^{-1} Q^{-1}\Lambda W_k(1) = Z_k^{-1} W_k(1). \tag{5.10}$$

The limiting distribution given by (5.10) does not depend on the nuisance parameters Q and Ω. It is trivial to show that $W_k(1)$ and P_k^{-1} are independent, so conditional on Z_k, $\hat{M}^{-1}\sqrt{T}\left(\hat{\beta} - \beta\right)$ is distributed as $N\left(0, P_k^{-1}\right)$. If we denote the density function of P_k by $p(P_k)$, the unconditional distribution of $\hat{M}^{-1}\sqrt{T}\left(\hat{\beta} - \beta\right)$ is $\int_0^1 N\left(0, P_k^{-1}\right) p(P_k)\, dP_k$. This is a mixture of normals, which is symmetric, but has thicker tails than the normal distribution. It is important to note that $\hat{M}$ is easy to compute from data.

The above derivation is similar to of Fisher's classical development of the t-statistic. Fisher utilized a data dependent transformation to avoid unknown variance parameters and obtained a distribution with fatter tails than the normal distribution. Although our approach is similar, note that we do not obtain the finite sample distribution of our teststatistic, and that $Z_k^{-1}W_k(1)$ is not a multivariate t distribution. In what follows, we will use the distribution obtained in (5.10) to develop tests for hypotheses on β.

5.3 Tests for General Hypotheses

We will now construct the relevant test statistic for simple hypotheses on the β's. In order to construct a test statistic t^* for hypotheses about the individual β's, we let the square root of the diagonal elements of $T\hat{B}^{-1}$ assume the role of the usual standard errors, i.e. $t^* = \left(\hat{\beta}_i - \beta_i\right) / \left(T^{-1}\hat{B}_{ii}^{-1}\right)^{\frac{1}{2}}$. Because the t^* statistic is invariant to the ordering of the individual β's, its asymptotic distribution is given by the first element in the vector $Z_k^{-1}W_k(1)$. Making use of the fact that Z_k^{-1} is lower triangular, it is straightforward to show that

$$t^* \Rightarrow W_1(1)\left[\int_0^1 \left(W_1(r) - rW_1(1)\right)^2 dr\right]^{-\frac{1}{2}} = W_1(1)\, P_1^{-\frac{1}{2}}.$$

The critical values of $W_1(1)\, P_1^{-\frac{1}{2}}$ are easy to simulate and are tabulated in Table 4.1. We now consider more general hypotheses in this framework.

We are interested in testing general non-linear hypotheses. We examine hypotheses of the form

$$H_0 : r(\beta_0) = 0, \qquad H_1 : r(\beta_0) \neq 0,$$

where $r(\cdot) : \mathbb{R}^k \rightarrow \mathbb{R}^q$ imposes q restrictions on the parameter vector β. We restrict our attention to hypotheses where $r(\cdot)$ is twice continuously differentiable with bounded second derivatives near β_0, and $R(\cdot) \equiv \frac{\partial}{\partial\beta'}r(\cdot)$ has full rank q in a neighborhood around β_0, implying that there are genuinely q restrictions. For later use, we let $\hat{R} = R\left(\hat{\beta}\right)$ and $R_0 = R(\beta_0)$.

We wish to obtain a test using the result from the previous section that the asymptotic distribution of $\hat{M}^{-1}T^{\frac{1}{2}}\left(\hat{\beta}-\beta_0\right)$ does not depend on any of the nuisance parameters.

The relevant test statistic is analogous to the standard Wald statistic for non-linear models. We simply substitute $\hat{B}$ for the standard variance-covariance matrix to obtain

$$F^* = T\, r\left(\hat{\beta}\right)'\left[\hat{R}\hat{B}\hat{R}'\right]^{-1} r\left(\hat{\beta}\right)/q.$$

The asymptotic distribution of F^* follows from the fact that

$$\begin{aligned} F^* &= T\left[\hat{R}\left(\hat{\beta}-\beta_0\right)\right]'\left[\hat{R}\hat{B}\hat{R}'\right]^{-1}\left[\hat{R}\left(\hat{\beta}-\beta_0\right)\right]/q + O_P\left(T^{-1}\right) \\ &\Rightarrow \left[R_0Q^{-1}\Lambda W_k(1)\right]'\left[R_0Q^{-1}\Lambda P_k\Lambda'Q^{-1}R_0'\right]^{-1}\left[R_0Q^{-1}\Lambda W_k(1)\right]/q \end{aligned}$$

But since $R_0Q^{-1}\Lambda$ has rank q and $W_k(1)$ is a vector of independent Wiener processes that are Gaussian, we can rewrite $R_0Q^{-1}\Lambda W_k(1)$ as $\Lambda^* W_q^*(1)$, where $W_q^*(1)$ is a $q-$dimensional vector of independent Wiener processes, and Λ^* is the $q \times q$ matrix square root of $R_0Q^{-1}\Lambda\Lambda'Q^{-1}R_0'$. Using the same arguments, it is then possible to establish that

$$R_0Q^{-1}\Lambda P_k\Lambda'Q^{-1}R_0' = \Lambda^* P_q\left(\Lambda^*\right)'.$$

and we therefore obtain the asymptotic distribution of F^* as follows:

$$\begin{aligned} F^* &\Rightarrow \left[\Lambda^* W_q^*(1)\right]'\left[\Lambda^* P_q\left(\Lambda^*\right)'\right]^{-1}\left[\Lambda^* W_q^*(1)\right]/q \\ &= W_q^*(1)' P_q^{-1} W_q^*(1)/q. \end{aligned}$$

We provide a fully detailed proof in Appendix C, and the following theorem formally states the result.

Theorem 11 : *Suppose that Assumptions 3 and 4 hold. Then under the null hypothesis* $H_0 : r(\beta) = 0$, $F^* \Rightarrow W_q(1)' P_q^{-1} W_q(1)/q$ *as* $T \to \infty$.

We have in this manner constructed a test for general nonlinear hypotheses on the parameters in a broad class of non-linear models, whose asymptotic distribution depends only on the number of restrictions. Table 4.2 tabulates the asymptotic critical values.

5.4 Empirical Illustration

In this section we illustrate the theoretical results with an empirical example. We wish to examine the effects of GDP growth on the growth of aggregate restaurant revenues. We let ΔRR denote the first difference of the natural logarithm of real, seasonally adjusted aggregate restaurant revenues for the United States, and ΔGDP denote the first difference of the natural logarithm of real, seasonally adjusted gross domestic product (GDP). Initially, we consider the basic regression model

$$\Delta RR = \beta_1 + \beta_2 \cdot \Delta GDP + u, \tag{5.11}$$

where β_1 is an intercept term, while β_2 is the parameter measuring the long run effect of GDP growth on restaurant revenues. It is unreasonable to think that the error term of this regression is i.i.d, so both HAC estimator tests and the method of testing introduced in this chapter are relevant methods of testing. To eliminate some of the autocorrelation in the error structure, we consider an AR(1) guess GLS transformation of the model. The idea is to "soak" up some of the auto correlation using the AR(1) transformation and then use HAC robust test to deal with any remaining autocorrelation in the model. Therefore, consider the model:

$$\begin{aligned} \Delta RR_t - \rho \cdot \Delta RR_{t-1} &= \beta_1 (1-\rho) + \beta_2 \cdot (\Delta GDP_t - \rho \cdot \Delta GDP_{t-1}) + u_t, \quad (5.12) \\ (1-\rho^2)^{\frac{1}{2}} \Delta RR_1 &= (1-\rho^2)^{\frac{1}{2}} \beta_1 + (1-\rho^2)^{\frac{1}{2}} \beta_2 \cdot \Delta GDP_1 + u_1 \end{aligned}$$

Table 5.1: Confidence Intervals

	$\hat{\beta}_2$	$\hat{\rho}$	t^*	$t-HAC$	$t-HAC-PW$
Model (5.11)	0.681		[0.31; 1.06]	[0.06; 1.30]	[−0.07; 1.43]
Model (5.12)	0.694	−0.293	[0.39; 1.00]	[0.18; 1.20]	[0.20; 1.19]

where ΔRR_{t-1} and ΔGDP_{t-1} are lagged values of first differences of real log-restaurant revenue and real log-GDP respectively. We estimate model (5.12) by non linear least squares, utilizing a grid search over values of ρ.

For both series, we use quarterly data from 1971 through 1996. The restaurant revenues are total for all sectors and the source is *Current Business Reports*, published by the Bureau of the Census. The GDP series and the GDP inflator were obtained from the *Survey of Current Business* published by the Bureau of Economic analysis, US Department of Commerce. The GDP series is nominal and seasonally adjusted, and we use the GDP deflator to obtain real GDP. We adjust the restaurant revenue data for seasonal flunctuations and again use the GDP deflator to obtain the real series. We are interested in the long-run effect of GDP growth on the growth of restaurant revenues (β_2). We compute confidence intervals using our t^* and HAC estimator test. We implement the HAC estimators as recommended by Andrews (1991) which uses the quadratic spectral kernel, and also with VAR(1) prewhitening as suggested by Andrews and Monahan (1992). The following table summarizes the results: We see that the different methods of calculating confidence intervals result in different intervals, and that our method provides a much tighter confidence interval than the two methods using HAC estimators. Using this empirical example as relevant data generating processes, we compare finite sample size and power of these different methods, using simulations. To this end, we fit the residuals from (5.11) to several different ARMA models and find that an AR(4) model provides a good fit. We also fit ΔGDP to several

ARMA processes and find that an AR(1) model renders a good fit. To perform the simulations, we generate data according to the two models. Model 1 is as follows: $\Delta RR_t = \beta_2 \cdot \Delta GDP_t + u_t, u_t = -0.343u_{t-1} - 0.330u_{t-2} - 0.269u_{t-3} + 0.595u_{t-4} + \xi_t$, $\xi_t \sim N(0, 0.0367746)$ and ΔGDP_t are the original regressors. In model 2, we simulate the regressors, and the model becomes: $\Delta RR_t = \beta_2 \cdot \Delta GDP_t + u_t, u_t = -0.343u_{t-1} - 0.330u_{t-2} - 0.269u_{t-3} + 0.595u_{t-4} + \xi_t$, $\xi_t \sim N(0, 0.0367746)$ and $\Delta GDP_t = -0.21 \cdot \Delta GDP_{t-1} + \zeta_t$, $\zeta_t \sim N(0, 0.007888)$. We generate data using $\beta_2 = .0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2$. We calculate finite sample size and power and asymptotic power for testing $\beta_2 = 0$. In all cases we use 2,000 replications.

The results are summarized in Table 5.2. We see that our method generally dominates the HAC estimators with respect to size. There is significantly less distortion. Size adjusted power is always bigger for the our method at very small values of β_2, but dominated by the HAC method, followed by the pre-whitened HAC at larger values. The same pattern hold for asymptotic power, when using the simulated regressors. When using the original regressor, however, The t^* achieves both better size and uniformly better power than the tests using HAC estimators.

5.5 Conclusion

In this chapter, we have developed a test statistic to test possibly non-linear hypotheses in nonlinear, weighted regression models with serial correlation/heteroskedasticity of unknown form. These tests are simple and do not require use of heteroskedasticity autocorrelation consistent (HAC) estimators. We derive the limiting null distributions of these new tests in a general nonlin-

Table 5.2: Simulation Results

		Finite Sample Size and Power			Finite Sample Size and Asymptotic Power		
Error Model[1]	β_0	t^*	HAC	HAC PW	t^*	HAC	HAC PW
1	0.0	0.028	0.006	0.005	0.028	0.006	0.005
	0.2	0.158	0.177	0.174	0.101	0.039	0.028
	0.4	0.429	0.504	0.502	0.339	0.187	0.155
	0.6	0.744	0.813	0.806	0.652	0.473	0.407
	0.8	0.925	0.961	0.954	0.871	0.750	0.690
	1.0	0.978	0.993	0.992	0.957	0.911	0.874
	1.2	0.994	1.000	0.999	0.988	0.979	0.958
	1.4	1.000	1.000	1.000	0.996	0.995	0.985
	1.6	1.000	1.000	1.000	1.000	1.000	0.995
	1.8	1.000	1.000	1.000	1.000	1.000	0.998
	2.0	1.000	1.000	1.000	1.000	1.000	1.000
2	0.0	0.046	0.031	0.027	0.046	0.031	0.027
	0.2	0.108	0.106	0.107	0.098	0.070	0.059
	0.4	0.252	0.289	0.286	0.233	0.223	0.203
	0.6	0.449	0.542	0.536	0.423	0.459	0.432
	0.8	0.628	0.739	0.732	0.607	0.678	0.657
	1.0	0.768	0.872	0.867	0.753	0.833	0.815
	1.2	0.867	0.941	0.944	0.854	0.916	0.908
	1.4	0.924	0.976	0.977	0.913	0.963	0.959
	1.6	0.954	0.990	0.992	0.947	0.985	0.983
	1.8	0.972	0.997	0.997	0.967	0.995	0.994
	2.0	0.980	0.998	0.998	0.977	0.997	0.997

ear setting, and show that while the tests have nonstandard distributions, the distributions depend only upon the number of restrictions.

The test presented in this chapter introduces a new class of tests, utilizing stochastic transformations, some of which can be used in situation with autocorrelation/heteroskedasticity of unknown form in the errors. While the selection of the specific statistic within this class is somewhat arbitrary, its properties with respect to invariance to nuisance parameters, and very little size distortion are highly desirable. Future work in this area should examine this class of statistics more closely in order to compare different statistics within this class of statistics.

We apply this method of testing to an empirical example and illustrate that the size of the new test is less distorted than tests utilizing the HAC estimators. The lesser size distortion of the newly introduced test makes it an attractive alternative to the currently used HAC test.

Chapter 6

Robust Inference in Models of Cointegration

6.1 Introduction

In empirical analysis of economic phenomena, economic theory often dictates the presence of both parameters that are of interest to the economist, and those that are not, called nuisance parameters. While statistical inference is drawn only on the parameters of interest (separate inference), the treatment accorded by the investigator to the nuisance parameters can significantly affect the results. In fact, the ability to conduct inference may be impaired even when consistent estimates of the parameters of interest can be obtained. This chapter proposes a way of dealing with this problem within the context of a specific environment.

The particular setup explored here is one where time series data are generated by unit root processes and the variables of interest captured by these data series show co-movement over the entire length of the time horizon considered. In such environments[1] it is well-known that the ability of the economist to conduct proper statistical inference is impaired by the all-too-common presence of heteroskedasticity and serial correlation. The parameters describing the specific form of the heteroskedasticity and serial correlation are often deemed as nuisance parameters

[1] Examples of such environments are frequently encountered in the macroeconomics and the asset-pricing literatures.

in this context. This chapter suggests a way of conducting proper statistical inference on the parameters of interest without having to directly estimate these nuisance parameters.

More specifically, let y be a unit root process, $f(t)$ a trend function, and X a set of regressors that also contain unit roots. A subcase of this model, which is also treated in this chapter, is the case with no regressors, i.e., y is trend-stationary. For models with regressors included, assume that there is exactly one cointegrating relationship such that the model can be represented as a standard univariate regression model with stationary error terms.[2] Furthermore assume that X is "exogenous" with respect to the cointegrating vector of parameters (the parameters of interest). This chapter proposes a new test statistic for testing hypotheses in models which fit the above description.

The development of the new test relies upon a data-dependent transformation of the ordinary least squares estimates of the parameters. The asymptotic distribution of the transformed estimates depends only on the parameters of the cointegrating vector, and therefore a test statistic which is invariant to the specific form of the correlation structure (the nuisance parameters) can be obtained.

To evaluate the finite sample performance of the new test, simulation experiments are performed. These simulations are repeated for tests currently employed in the literature, thus providing a basis for comparison. It is shown that in general, size distortions are much less severe than those of tests currently employed in the literature and the size-adjusted power of the new test is only marginally lower than that of tests currently employed.

Even though the size distortions of the new test are generally less than those

[2]The model introduced in this paper is designed primarily for the analysis of the long term relationship between integrated variables.

of tests currently employed in the literature, size remains inflated when serial correlation is high. A way of correcting this problem is provided for the model with a constant, a linear trend, and no regressors. Simulations confirm that the correction alleviates the problem of inflated size.

In order to place the new test statistic in perspective, it seems appropriate at this stage to describe ways in which the literature currently tackles the presence of heteroskedasticity and autocorrelation of unknown form in the type of models that are the focus of this chapter. In this connection, two very different frameworks stand out: the single equation framework, which forms the basis of this chapter, and the systems framework. The standard approach used to deal with heteroskedasticity and serial correlation in single equation models with exogenous regressors is to estimate the correlation structure of the error terms using non-parametric heteroskedasticity and autocorrelation consistent (HAC) estimators.[3] These estimators furnish consistent estimates of the correlation structure. Using these estimates, inference on the cointegrating vector is carried out using conventional tests. Specifically, after the covariance matrix has been estimated, it is treated as if it is known, and testing proceeds from there. Inference conducted in this manner is robust to heteroskedasticity and serial correlation of unknown form. Even though tests that use HAC estimators are valid *asymptotically*, they may potentially display substantial size distortions. Simulation studies that document these size distortions have been carried out for stationary models, see for example, Andrews (1991), Andrews and Monahan (1992) and Den Haan and Levin (1997). The fact that the estimation of the correlation structure causes size distortions carries with it the implication that there may be significant benefits from circum-

[3]HAC estimators have been thoroughly examined in the literature. Among the major contributions are Andrews [1991], Andrews and Monahan [1992], Hansen [1992b], Newey and West [1987], Robinson [1991] and White [1984].

venting this procedure. The test proposed here achieves exactly that. In fact, the distinctive feature of the test developed in this chapter is that even though it does not require estimation of the correlation structure, it is still robust to serial correlation of unknown form.

In the general literature on estimation and testing in cointegrating systems, a commonly used method is the full information maximum likelihood (FIML) approach developed in Johansen (1988, 1991) and Johansen and Juselius (1990, 1992).[4] A discussion and comparison seems warranted here even though the procedure outlined in these chapters is tailored for models that are somewhat different from those within the scope of this chapter. The FIML approach models multivariate structural models, and is therefore very flexible in the sense that it allows the empiricist to test for the number of cointegrating relationships, without requiring exogeneity of any kind. It estimates the correlation structure as part of the maximum likelihood estimation, and is therefore robust to heteroskedasticity and serial correlation. Johansen's method does however require the researcher to express the model as a VAR model for which a lag length must be chosen. While modeling the cointegration relationship as a VAR model is useful for examining the short term dynamics and while inference conducted on such a model is robust to heteroskedasticity and serial correlation, it is clear that the choice of lag length will influence the results.[5] The test statistic developed in this chapter circumvents the issue of lag length choice; as such it may be viewed as an alternative to the Johansen approach when dealing with single equation models.

The rest of the chapter is organized as follows. In Section 6.2 the model is described in detail, the required assumptions are stated, and some of the basic

[4]It should be noted that this approach can also be used to estimate single equation models.
[5]For a recent application of Johansen's method in which the choice of lag length matters for inference, see Metin [1998].

asymptotic results are presented. In Section 6.3, the test statistic is derived and the asymptotic distribution thereof stated. In Section 6.4, the results of the finite sample simulation experiments are reported. Section 6.5 explores the cause of the inflated size when serial correlation is high and implements a size correction procedure for one specific model. Section 6.6 concludes. Proofs of important results are collected in Appendix D.

6.2 The Model Setup

Consider the following regression model containing a single cointegrating relationship:

$$\begin{aligned} y_t &= f(t)'\alpha + X_t'\beta + u_t, \qquad t = 1, ..., T \\ X_t &= X_{t-1} + v_t, \end{aligned} \tag{6.1}$$

where $f(t)$ denotes a $(k_1 \times 1)$ vector of trend functions, X_t a $(k \times 1)$ vector of regressors[6] and α and β are $(k_1 \times 1)$ and $(k \times 1)$ vectors of parameters respectively. The following assumptions will be maintained throughout the chapter. Conditional on X_t, u_t is a scalar, mean zero random process. The sequences $\{u_t\}$ and $\{v_t\}$ do not have unit roots, but may exhibit serial correlation or heteroskedasticity. Furthermore, it is assumed that the cross-spectral density of $\{u_t\}$ and $\{v_t\}$ is zero at frequency zero. This implies that the regressors are strictly exogenous in the sense of Phillips and Park (1988).

A single equation model like (6.1) can arise either directly from an economic model or as the result of a multivariate analysis.[7] The latter case may obtain if

[6]Although the model as it is characterized in (6.1) does not allow for trends in the regressors, the asymptotic results derived in this paper remain valid for hypotheses on β if the trends in the regressors are included in $f(t)$. This stems from the fact that the test statistic is invariant to projections of subsets of regressors in linear models. See Kiefer, Vogelsang and Bunzel [1998] for details.

[7]For a recent empirical application, see Brouwer and Ericsson [1998]. Here a single equa-

either the multivariate model directly delivers a system like the one in (6.1) above, or when some parameters estimated from the full system are taken as given and the model accordingly transformed.

At times, it will be useful to stack the equations in (6.1) and rewrite them as

$$y = \mathbf{f}(T)\,\alpha + X\beta + u. \tag{6.2}$$

Here $\mathbf{f}(T)$ is the $(T \times k_1)$ stacked vector of trend functions, and X is the $(T \times k_1)$ matrix of regressors. Furthermore, θ will denote $[\alpha' \ \beta']'$ and ordinary least squares will be used to obtain an estimate of θ (denoted by $\hat{\theta}$).

The following notation is required to state the central assumptions of the chapter. Denote $S_t' = \sum_{j=1}^{t}[f(j), X_j]u_j$, and let $w_k(r)$ be a k-vector of independent Wiener processes, and $[rT]$ be the integer part of rT, where $r \in [0,1]$. " $\Rightarrow$ " is used to denote weak convergence. The following assumptions will be sufficient to obtain the main results of the chapter.

Assumption 5 :

$\begin{bmatrix} T^{-\frac{1}{2}}\sum_{j=1}^{[rT]} v_j \\ T^{-\frac{1}{2}}\sum_{j=1}^{[rT]} u_j \end{bmatrix} \Rightarrow \Omega \begin{bmatrix} w_k(r) \\ w_1(r) \end{bmatrix}$, where $w_k(s)$ and $w_1(s)$ are independent, $\Omega = \begin{bmatrix} \Lambda & 0 \\ 0 & \sigma \end{bmatrix}$, and $\Lambda\Lambda'$ and σ^2 are 2π times the spectral density (evaluated at frequency 0) of v and u respectively.

Assumption 6 :

There exists a $(k_1 \times k_1)$ diagonal matrix τ_T and a vector of functions F,

tion is obtained from a multivariate model of cointegration, and inference is conducted on the cointegration vector.

such that

$$\tau_T f(t) = F\left(\tfrac{t}{T}\right) + o(1)$$

$$\int_0^1 F_i(s)\,ds < \infty,\ i = 1, ..., k_1$$

$$\det\left[\int_0^1 F(s)\,F(s)'\,ds\right] > 0$$

The functional central limit theorem stated in Assumption 5 can be obtained by placing mixing restrictions on the errors. The restrictions on the trend function stated in Assumption 6 are general enough to allow polynomial trends, but rule out ill-behaved trend functions like $f_i(t) = 1/t$ and trends that are linearly dependent asymptotically.[8] These assumptions can be relaxed, but as they stand, are sufficiently general to cover most commonly used models. Both assumptions will be maintained henceforth. For later use, let $\mathbf{F}(T)$ be the matrix of the stacked $F(t/T)$ functions.

Using Assumptions 5 and 6, the asymptotic distribution of the least squares estimate of θ, $\hat{\theta}$, is derived in Lemma 2. To that end, it is necessary to develop some additional notation. $w_k^F(s)$ is defined as the residual from the projection of $w_k(s)$ on the subspace generated by $F(s)$ in the Hilbert space of square integrable functions on [0,1] with the inner product $(f, g) = \int_0^1 fg$. Correspondingly, $F(s)^X$ is the residual from the projection of $F(s)$ onto the space generated by $w_k(s)$. The exact expressions can be found in Appendix D.1.

[8]These assumptions follow Vogelsang [1998].

Lemma 2 *Suppose Assumptions 5 and 6 hold. Then*

$$D_T T^{\frac{1}{2}}\left(\hat{\theta}-\theta\right) \Rightarrow \left(\sigma\left(\Sigma'\right)^{-1}\right)\left[\begin{array}{c}\left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 F^X(s)\,dw_1(s)\\ \left(\int_0^1 w_k^F(s)\,w_k^F(s)'\,ds\right)^{-1}\int_0^1 w_k^F(s)\,dw_1(s)\end{array}\right],$$

where

$$\Sigma=\begin{bmatrix} I_{k_1} & 0\\ 0 & \Lambda\end{bmatrix} \text{ and } D_T=\begin{bmatrix}\tau_T^{-1} & 0\\ 0 & T^{\frac{1}{2}}I_k\end{bmatrix}.$$

The asymptotic distribution of $(\hat{\theta}-\theta)$ is a function of the independent Wiener processes $w_1(s)$ and $w_k(s)$, the normalized trend functions, $F(s)$, and the parameters relating to the correlation structure, σ and Λ. Note that the asymptotic distribution of $(\hat{\theta}-\theta)$ is proportional to $\left(\sigma\left(\Sigma'\right)^{-1}\right)$, the only unknown parameters entering into the asymptotic distribution of $(\hat{\theta}-\theta)$. This proportionality plays an important role in the construction of the test statistic, as will become apparent later when a data-dependent matrix with the same asymptotic proportionality property is found. To foreshadow, such a matrix will then be used to transform $(\hat{\theta}-\theta)$ in such a way that the asymptotic distribution of $(\hat{\theta}-\theta)$ no longer depends on σ and Σ.

If the errors are not serially correlated, standard Wald tests can be constructed; such tests on β are asymptotically χ^2-distributed.[9] However, if the errors *are* serially correlated, the standard procedure to date has been to estimate the asymptotic covariance matrix.[10] Using the estimated covariance matrix, standard Wald and $t-$type tests can be carried out.

Instead of following this route and proceeding to estimate the correlation structure, the test developed in this chapter relies on a data-dependent transformation

[9] For this result and the derivation thereof see Hamilton [1994; Chapter 19].

[10] Hamilton [1994; pp. 607] suggests using non-parametric covariance matrix estimates as a possible way of dealing with serial correlation.

which transforms the parameter estimates in such a way that the asymptotic distribution of the transformed parameters no longer depends on unknown parameters. As will be established below, the transformation matrix has a non-degenerate asymptotic distribution (as opposed to converging to a fixed matrix). The transformation is derived in the next section.

6.3 The Test Statistic

The transformation of the parameters of interest (θ) that will eliminate the nuisance parameters relating to the correlation structure in the asymptotic distribution of the test statistic is now derived. Consider $T^{-\frac{1}{2}}\hat{S}_{[rT]}$, where $\hat{S}'_{[rT]} = \sum_{t=1}^{[rT]}[f(t)\ X_t]\hat{u}_t$, and $\hat{u}_t$ are the OLS residuals obtained from estimation of the model in (6.1). In Appendix D.2, the following lemma is proven:

Lemma 3 *Suppose Assumptions 5 and 6 hold. Then*

$$T^{-\frac{1}{2}}D_T^{-1}\hat{S}_{[rT]} = T^{-\frac{1}{2}}D_T^{-1}\sum_{t=1}^{[rT]}\begin{bmatrix} f(t) \\ X_t \end{bmatrix}\hat{u}_t \Rightarrow (\sigma\Sigma)\, Q_k^F(r),$$

where

$$Q_k^F(r) = \begin{bmatrix} \int_0^r F(s)\,dw_1(s) \\ \int_0^r w_k(s)\,dw_1(s) \end{bmatrix} - \begin{bmatrix} \int_0^r F(s)\,F(s)'\,ds & \int_0^r F(s)\,w_k(s)'\,ds \\ \int_0^r w_k(s)\,F(s)'\,ds & \int_0^r w_k(s)\,w_k(s)'\,ds \end{bmatrix} \times \begin{bmatrix} \left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 F^X(s)\,dw_1(s) \\ \left(\int_0^1 w_k^F(s)\,w_k^F(s)'\,ds\right)^{-1}\int_0^1 w_k^F(s)\,dw_1(s) \end{bmatrix}.$$

Note that the asymptotic distribution of $\hat{S}_{[rT]}$ and the distribution of the parameters share a common property: both are proportional to Σ. This property is used to create a transformation of the parameters, which ensures that the asymptotic distribution of the transformed parameters do not depend on Σ. To this end, define $\hat{C} \equiv T^{-2}\sum_{t=1}^{T}\hat{S}_t\hat{S}_t'$. From Lemma 3 and the continuous mapping theorem, it follows that

$$\begin{aligned} D_T^{-1}\hat{C}D_T^{-1} &= T^{-1}\sum_{t=1}^{T}\left[T^{-\frac{1}{2}}D_T^{-1}\hat{S}_t\right]\left[T^{-\frac{1}{2}}D_T^{-1}\hat{S}_t\right]' \\ &\Rightarrow \left(\sigma^2\Sigma\right)\int_0^1 Q_k^F(r)\left(Q_k^F(r)\right)' dr\Sigma' \\ &= \sigma^2\Sigma P_k^F\Sigma', \end{aligned}$$

where

$$P_k^F \equiv \int_0^1 Q_k^F(r)\left(Q_k^F(r)\right)' dr.$$

A matrix whose asymptotic distribution is proportional to Σ has now been found. Letting $\hat{C} = \hat{C}^{\frac{1}{2}}\left(\hat{C}^{\frac{1}{2}}\right)'$ be the Cholesky decomposition of $\hat{C}$, define the transformation matrix $\hat{M} = \left(\frac{1}{T}\begin{bmatrix} \mathbf{f}(T)'\mathbf{f}(T) & \mathbf{f}(T)'X \\ X'\mathbf{f}(T) & X'X \end{bmatrix}\right)^{-1}\hat{C}^{\frac{1}{2}}$. Then the following lemma can be stated:

Lemma 4 *The asymptotic distribution of* $\hat{M}^{-1}T^{\frac{1}{2}}\left(\hat{\theta}-\theta\right)$ *does not depend on* σ *or on* Σ.

The proof of Lemma 4 along with the expression for the asymptotic distribution can be found in Appendix D.3. The fact that the asymptotic distribution of $\hat{M}^{-1}T^{\frac{1}{2}}\left(\hat{\theta}-\theta\right)$ does not depend on the nuisance parameters could be utilized to construct a $t-$test for simple hypotheses. This is not done here, as these hypotheses are equally well handled by the test statistic which encompasses more general hypotheses. This more general test statistic is developed below.

To construct a Wald-type test, define

$$\hat{B} = \left(\frac{1}{T}\begin{bmatrix} \mathbf{f}(T)'\mathbf{f}(T) & \mathbf{f}(T)'X \\ X'\mathbf{f}(T) & X'X \end{bmatrix}\right)^{-1}\hat{C}\left(\frac{1}{T}\begin{bmatrix} \mathbf{f}(T)'\mathbf{f}(T) & \mathbf{f}(T)'X \\ X'\mathbf{f}(T) & X'X \end{bmatrix}\right)^{-1}.$$

Most hypotheses of the form $H : R\hat{\theta} = r$, can now be tested using the following test statistic:

$$F^{*} = T\left(R\hat{\theta} - r\right)'\left[R\hat{B}R'\right]^{-1}\left(R\hat{\theta} - r\right)/q.$$

Notice that $\hat{B}$ enters the test statistic at exactly the place where a covariance matrix estimate would typically be inserted. It is important to note, however, that $\hat{B}$ is *not* an estimate of the covariance matrix. In fact $\hat{B}$ converges to a non-degenerate distribution.

In what follows, two different types of hypotheses will be considered: $H_0^F : R^F\alpha = r$ and $H_0^X : R^X\beta = r$, where R^F and R^X are non-stochastic restriction matrices of dimension $q \times k_1$ and $q \times k$ respectively. Both R^F and R^X are assumed to have rank q. The separation of hypotheses on the two types of parameters is required to eliminate the nuisance parameters and simplify the asymptotic distribution.

Some additional notation is required to state the asymptotic distribution of F^{*} under H_0^F. As is well-known, estimators of coefficients on different trends will often converge at different rates. Specifically, the coefficients entering the constraint which converge the slowest will dominate the asymptotic distribution. In order to formalize this, let μ_i be the largest non-positive power of t in the nonzero elements in the i'th row of $R\tau_T$. Then define the $q \times q$ diagonal matrix A in such a way that $A_{ii} = T^{\mu_i}$, and let $R^{*} = \lim_{T\to\infty} A^{-1}R\tau_T$. Finally define $\tilde{w}_q^F(s)$ as the residual from the projection of the first q coordinates of $w_k^F(s)$ onto the last $(k-q)$ coordinates of $w_k^F(s)$. The following theorem states the asymptotic distribution of F^{*}.

Theorem 12 *Suppose Assumptions 5 and 6 hold. Then*

(a) Under H_0^X,

$$F^* \Rightarrow \left[\int_0^1 \hat{w}_q^F(s)\,dw_1(s)\right]' \left[\int_0^1 V(r)\,V(r)'\,dr\right]^{-1} \left[\int_0^1 \hat{w}_q^F(s)\,dw_1(s)\right]/q$$

where

$$\begin{aligned} V(r) &= \int_0^r \hat{w}_q^F(s)\,dw_1(s) \\ &\quad - \left(\int_0^r \hat{w}_q^F(s)\,\hat{w}_q^F(s)'\,ds\right)\left(\int_0^1 \hat{w}_q^F(s)\,\hat{w}_q^F(s)'\,ds\right)^{-1}\int_0^1 \hat{w}_q^F(s)\,dw_1(s). \end{aligned}$$

(b) Under H_0^F, *if* $k > 0$,

$$\begin{aligned} F^* \;\Rightarrow\; & \left(R^*\left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 F^X(s)\,dw_1(s)\right)' \\ & \Bigg[R^*\left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 V^{F^X}(r)\left(V^{F^X}(r)\right)'\,dr \\ & \left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}(R^*)'\Bigg]^{-1} \\ & \left(R^*\left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 F^X(s)\,dw_1(s)\right) \end{aligned}$$

where

$$\begin{aligned} V^{F^X}(r) &= \int_0^r F^X(s)\,dw_1(s) - \left(\int_0^r F^X(s)\,F^X(s)'\,ds\right)\cdot \\ & \quad \left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 F^X(s)\,dw_1(s). \end{aligned}$$

(c) Under H_0^F, if $k=0$,

$$
\begin{aligned}
F^* \Rightarrow\ & \left(R^*\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 F(s)\,dw_1(s)\right)' \\
& \left[R^*\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 V^F(r)\left(V^F(r)\right)'dr\right. \\
& \left.\left(\int_0^1 F(s)F(s)'ds\right)^{-1}(R^*)'\right]^{-1} \\
& \left(R^*\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 F(s)\,dw_1(s)\right)
\end{aligned}
$$

where

$$
\begin{aligned}
V^F(r) =\ & \int_0^r F(s)\,dw_1(s) \\
& -\left(\int_0^r F(s)F(s)'ds\right)\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 F(s)\,dw_1(s).
\end{aligned}
$$

Part (a) of the theorem concerns testing of hypotheses on the cointegrating vector. Through $\hat{w}_q^F(s)$, this distribution will depend on i) the number of restrictions being tested, ii) the number of regressors in the model, and iii) the trends included. The critical values of this distribution have been simulated for k and q up to 8 in the cases with a constant, a constant and a linear trend, and a constant, a linear trend and a quadratic trend. The critical values are provided in Tables 6.1-6.3. The critical values for k and q up to 15 are available upon request. All the critical value presented in this chapter have been simulated using sums of $N(0,1)$ i.i.d. random variables to approximate the Wiener processes in the distributions. In each case, 10,000 replications were used, and the integrals were computed as averages over 1000 simulated observation points. This is the case for all simulations in this chapter, unless otherwise stated. The program GAUSS was used for all simulations in this chapter.

Parts (b) and (c) of the theorem concern tests of hypotheses on the trend functions when regressors are, and are not, included in the model respectively. The asymptotic distribution here depends on exactly the same parameters as in part (a), and in addition, it depends on R^*. The presence of R^* in the asymptotic distribution reflects the different rates of convergence of the coefficient estimates of the trends. Tables 6.4-6.9 provide the critical values for all hypotheses with q equal to 1 on the trends in the model. The tables cover the model with just a constant, the model with a constant and a linear trend and the model with a constant, a linear trend, and a quadratic trend, allowing for up to 8 regressors in these models.

Table 6.1: Critical Values, Constant, No Trend

k	q	90%	95%	97.50%	99%
1	1	39.941	64.463	95.198	141.379
2	1	42.531	67.043	101.554	161.063
2	2	51.430	75.914	107.050	147.865
3	1	42.842	70.699	102.425	165.225
3	2	51.482	75.152	102.746	141.582
3	3	59.443	80.974	108.475	152.800
4	1	42.996	66.919	98.473	158.323
4	2	52.419	76.521	105.926	152.213
4	3	61.214	87.044	111.111	147.060
4	4	68.841	91.445	114.780	154.083
5	1	41.403	66.435	102.562	160.043
5	2	53.005	77.154	104.592	146.212
5	3	61.827	84.704	111.876	142.583
5	4	71.013	95.087	120.736	152.009
5	5	78.076	101.373	126.923	164.426
6	1	41.059	63.989	97.631	150.444
6	2	52.665	78.567	108.830	148.055
6	3	62.948	88.009	111.270	151.113
6	4	71.512	95.028	121.887	156.659
6	5	80.503	106.103	129.534	164.301
6	6	89.009	112.603	137.332	168.117
7	1	40.744	64.536	94.292	143.314
7	2	52.461	75.540	101.681	139.659
7	3	62.786	85.263	110.869	146.516
7	4	70.879	96.156	119.198	155.765
7	5	78.813	103.290	128.257	164.253
7	6	87.056	111.456	136.388	170.960
7	7	93.511	116.464	143.263	176.476
8	1	41.328	66.412	98.689	140.320
8	2	52.401	75.654	102.149	139.288
8	3	61.930	86.174	111.639	148.823
8	4	71.745	96.014	121.467	162.687
8	5	80.671	105.717	133.598	173.835
8	6	89.189	115.504	141.195	180.695
8	7	97.376	122.492	150.311	187.835
8	8	104.039	131.187	156.800	190.164

Model: $y_t = \alpha_1 + \beta X_t + u_t$

H_0: $R\beta = r$, rank(R) = q

Table 6.2: Critical Values, Constant, Linear Trend

k	q	90%	95%	97.50%	99%
1	1	40.302	67.331	100.574	150.131
2	1	41.048	64.983	95.489	147.936
2	2	51.350	76.736	103.065	148.711
3	1	40.704	66.470	99.157	151.090
3	2	52.836	78.077	105.669	152.292
3	3	62.749	87.868	113.329	153.347
4	1	42.663	68.085	97.591	155.178
4	2	53.540	79.826	107.101	149.458
4	3	63.961	87.373	111.070	143.204
4	4	70.892	94.058	116.727	153.595
5	1	41.821	67.376	99.607	152.917
5	2	52.332	76.406	102.789	142.259
5	3	62.047	85.743	108.786	142.031
5	4	71.411	93.383	117.568	147.852
5	5	78.657	102.028	125.914	158.851
6	1	40.778	65.865	97.849	144.572
6	2	52.819	77.826	103.948	147.378
6	3	62.691	88.168	112.699	152.327
6	4	72.771	96.024	124.426	157.167
6	5	82.178	106.089	129.627	167.112
6	6	89.187	112.979	138.177	173.989
7	1	41.379	66.153	97.654	147.515
7	2	52.396	77.377	104.253	145.342
7	3	61.967	86.013	112.262	148.147
7	4	71.365	94.125	118.493	155.587
7	5	80.031	102.343	126.040	161.689
7	6	88.732	111.442	138.584	168.871
7	7	95.325	119.017	144.373	175.987
8	1	41.855	65.646	96.733	138.585
8	2	52.880	76.271	102.861	139.301
8	3	62.837	87.332	114.177	152.365
8	4	71.726	98.535	123.487	163.620
8	5	81.571	106.123	131.457	163.516
8	6	90.326	115.252	142.593	177.756
8	7	97.385	123.846	149.708	183.471
8	8	104.046	130.620	157.268	193.199

Model: $y_t = \alpha_1 + \alpha_2 t + \beta X_t + u_t$

H_0: $R\beta = r$, rank(R) = q

Table 6.3: Critical Values, Constant, Linear and Quadratic Trend

k	q	90%	95%	97.50%	99%
1	1	40.901	65.625	92.960	147.419
2	1	40.906	65.977	97.374	149.117
2	2	52.741	77.904	106.186	144.130
3	1	42.997	66.642	93.577	137.544
3	2	52.679	75.533	104.835	146.465
3	3	60.929	85.321	114.794	151.887
4	1	40.859	68.639	102.170	153.968
4	2	53.693	78.056	102.439	140.205
4	3	62.367	84.512	107.757	144.293
4	4	70.425	92.268	116.921	155.833
5	1	41.743	66.446	100.747	150.121
5	2	52.307	74.980	99.840	132.988
5	3	62.664	83.807	107.075	143.149
5	4	70.393	93.046	117.660	152.543
5	5	78.199	101.610	126.040	161.086
6	1	40.196	66.791	97.462	153.851
6	2	52.901	75.757	104.041	142.922
6	3	63.727	88.842	112.393	155.020
6	4	74.071	95.982	121.698	159.960
6	5	80.535	105.627	131.286	166.083
6	6	89.287	114.607	142.655	179.816
7	1	42.491	67.789	100.358	147.496
7	2	51.889	77.450	104.480	147.797
7	3	62.766	87.165	114.896	151.412
7	4	72.963	98.068	123.676	164.721
7	5	82.437	108.603	134.586	173.427
7	6	89.145	113.264	140.902	177.243
7	7	97.081	122.450	150.617	185.986
8	1	40.392	66.337	95.583	143.188
8	2	52.549	74.199	99.595	134.496
8	3	63.265	89.436	116.867	157.467
8	4	73.755	99.281	125.514	160.995
8	5	82.359	107.812	132.491	166.428
8	6	90.756	116.617	145.619	181.820
8	7	99.187	126.714	153.042	185.649
8	8	105.653	131.474	157.223	194.111

Model: $y_t = \alpha_1 + \alpha_2 t + \alpha_3 t^2 + \beta X_t + u_t$

H_0: $R\beta = r$, rank(R) = q

Table 6.4: Critical Values, Constant, Test on Constant

k	90%	95%	97.50%	99%
0	29.066	46.704	65.761	101.115
1	41.272	66.981	101.335	150.917
2	45.566	71.193	105.510	152.674
3	49.611	81.567	115.226	185.294
4	48.752	78.268	113.942	174.757
5	48.998	77.779	116.975	185.329
6	50.007	82.780	123.154	196.556
7	52.564	84.088	122.333	195.123
8	51.053	81.869	122.796	181.355

Model: $y_t = \alpha_1 + \beta X_t + u_t$

H_0: $\alpha_1 = r$

Table 6.5: Critical values, Constant and Trend, Test on Constant

k	90%	95%	97.50%	99%
0	44.500	74.159	106.332	160.229
1	50.331	82.912	116.475	167.548
2	52.628	87.534	126.794	194.629
3	54.937	88.148	132.070	199.587
4	52.207	82.831	123.600	196.940
5	52.054	86.823	127.757	196.587
6	52.110	83.437	127.109	196.183
7	54.583	88.108	128.936	201.540
8	52.021	88.203	125.961	198.052

Model: $y_t = \alpha_1 + \alpha_2 t + \beta X_t + u_t$

H_0: $\alpha_1 = r$

Table 6.6: Critical Values, Constant and Trend, Test on Trend

k	90%	95%	97.50%	99%
0	40.586	65.500	97.940	149.300
1	45.824	73.820	105.282	166.101
2	45.650	74.752	108.416	160.058
3	45.177	72.369	104.935	163.517
4	43.335	69.712	104.218	155.573
5	43.643	70.274	105.105	165.427
6	43.143	69.613	102.879	158.146
7	41.867	69.155	99.137	148.255
8	41.419	69.005	102.759	153.782

Model: $y_t = \alpha_1 + \alpha_2 t + \beta X_t + u_t$

H_0: $\alpha_2 = r$

Table 6.7: Critical Values, Constant, Linear and Quadratic Trend, Test on Constant

k	90%	95%	97.50%	99%
0	64.594	102.436	150.447	222.201
1	64.182	104.363	149.005	230.798
2	67.945	107.109	150.538	228.866
3	67.407	106.989	152.426	235.882
4	60.055	96.479	137.906	219.423
5	61.426	101.818	152.135	228.772
6	59.833	97.551	148.058	229.032
7	60.128	101.135	149.611	219.516
8	59.925	96.901	144.817	215.843

Model: $y_t = \alpha_1 + \alpha_2 t + \alpha_3 t^2 + \beta X_t + u_t$

H_0: $\alpha_1 = r$

Table 6.8: Critical Values, Constant, Linear and Quadratic Trend, Test on Linear Trend

k	90%	95%	97.50%	99%
0	45.990	73.994	108.815	161.666
1	47.211	75.400	110.295	165.738
2	48.647	78.216	112.615	171.502
3	47.890	74.293	107.468	164.720
4	45.580	75.160	112.917	171.393
5	44.712	71.353	108.486	165.347
6	44.563	71.918	103.713	156.789
7	42.588	68.375	95.997	156.518
8	43.231	69.003	102.201	158.722

Model: $y_t = \alpha_1 + \alpha_2 t + \alpha_3 t^2 + \beta X_t + u_t$

H_0: $\alpha_2 = r$

Table 6.9: Critical Values, Constant, Linear and Quadratic Trend, Test on Quadratic Trend

k	90%	95%	97.50%	99%
0	44.315	71.743	107.814	158.306
1	46.048	74.218	104.328	165.030
2	48.228	76.563	111.731	171.607
3	46.234	72.582	106.676	154.149
4	45.793	75.116	108.106	162.784
5	45.206	71.312	106.484	169.002
6	43.687	69.890	101.159	145.975
7	44.393	70.377	104.009	158.481
8	44.282	72.791	106.057	162.112

Model: $y_t = \alpha_1 + \alpha_2 t + \alpha_3 t^2 + \beta X_t + u_t$

H_0: $\alpha_3 = r$

6.4 Monte Carlo Experiments

In this section, results of simulation experiments designed to analyze the finite sample performance of the test statistic are reported. To obtain a reference for comparison, the simulations are repeated for tests currently employed in the literature.

The experiments concentrate on two different models, one with four regressors, and one with no regressors. Both models have the same trend function, $f(t) = [\ 1 \quad t\ ]'$. The first model may be formally stated as:

$$\begin{aligned} y_t &= \alpha_1 + \alpha_2 t + X_t'\beta + u_t \\ X_t &= X_{t-1} + v_t, \end{aligned} \tag{6.3}$$

where X_t is a (4×1) matrix of regressors, and the second model is described by

$$y_t = \alpha_1 + \alpha_2 t + u_t. \tag{6.4}$$

All the simulation results reported in this section have been performed using 10,000 replications.

The first set of simulations are based on (6.3). The hypothesis being tested is $H_0 : \beta_1 = 0$ against the alternative $\beta_1 \neq 0$. The errors are generated according to $u_t = \rho u_{t-1} + e_t + \eta e_{t-1}$, where e_t is *i.i.d.* $N(0,1)$ and $\{v_t\}$ is *i.i.d.* $N(0,1)$. Simulations are reported for $\rho = 0,\ 0.9,\ 0.95,\ 1$, and for $\eta = 0,\ -0.4,\ 0.4$. The sample size is 40.

Table 6.10: Simulations: Finite Sample Size and Power of F*, Wald and Wald PW

		η = 0.0					η = −0.4					η = 0.4		
ρ	β_1	F*	Wald	Wald PW	ρ	β_1	F*	Wald	Wald PW	ρ	β_1	F*	Wald	Wald PW
	0.0	0.053	0.126	0.132		0.0	0.029	0.174	0.140		0.0	0.076	0.121	0.101
	0.5	0.768	0.886	0.887		0.5	0.857	0.932	0.947		0.5	0.604	0.772	0.733
0.00	1.0	0.983	0.999	0.999	0.00	1.0	0.992	0.999	1.000	0.00	1.0	0.939	0.995	0.990
	1.5	0.998	1.000	1.000		1.5	0.999	1.000	1.000		1.5	0.993	1.000	1.000
	0.0	0.168	0.241	0.174		0.0	0.137	0.232	0.211		0.0	0.177	0.240	0.136
	1.0	0.565	0.693	0.632		1.0	0.827	0.940	0.920		1.0	0.375	0.461	0.385
0.90	2.0	0.917	0.973	0.944	0.90	2.0	0.991	1.000	0.998	0.90	2.0	0.781	0.884	0.802
	3.0	0.988	0.998	0.991		3.0	1.000	1.000	1.000		3.0	0.937	0.980	0.940
	4.0	0.998	1.000	0.998		4.0	1.000	1.000	1.000		4.0	0.983	0.996	0.978
	5.0	1.000	1.000	0.999		5.0	1.000	1.000	1.000		5.0	0.995	0.999	0.992
	0.0	0.180	0.256	0.183		0.0	0.148	0.250	0.225		0.0	0.190	0.256	0.146
	1.0	0.517	0.624	0.561		1.0	0.794	0.909	0.885		1.0	0.336	0.412	0.346
0.95	2.0	0.889	0.954	0.916	0.95	2.0	0.988	0.999	0.996		2.0	0.735	0.846	0.754
	3.0	0.979	0.994	0.981		3.0	0.999	1.000	1.000	0.95	3.0	0.914	0.966	0.917
	4.0	0.996	0.999	0.995		4.0	1.000	1.000	1.000		4.0	0.973	0.992	0.967
	5.0	0.999	1.000	0.998		5.0	1.000	1.000	1.000		5.0	0.992	0.998	0.985
	0.0	0.190	0.263	0.187		0.0	0.155	0.255	0.229		0.0	0.195	0.265	0.152
	1.0	0.501	0.611	0.556		1.0	0.766	0.898	0.874		1.0	0.327	0.399	0.330
	2.0	0.879	0.945	0.905		2.0	0.981	0.997	0.994		2.0	0.725	0.828	0.732
1.00	3.0	0.975	0.993	0.977	1.00	3.0	0.999	1.000	0.999	1.00	3.0	0.906	0.957	0.903
	4.0	0.994	0.998	0.993		4.0	1.000	1.000	1.000		4.0	0.970	0.989	0.960
	5.0	0.999	1.000	0.997		5.0	1.000	1.000	1.000		5.0	0.989	0.997	0.981

Model: $y_t = \alpha_1 + \alpha_2 t + X_t\beta + u_t$ N = 10,000, T = 40 H_0: $\beta_1 = 0$

Alongside the results for F^*, results for two Wald tests are also reported. The column labeled WALD reports the results for the standard Wald test which is constructed by replacing the usual OLS estimate of the variance by a HAC estimate of the long run variance constructed from the OLS residuals. The specific HAC estimator used is the one recommended by Andrews (1991), which utilizes the quadratic spectral kernel; the automatic data-dependent procedure proposed by Andrews (1991) was used to select the bandwidth. The column labeled WALD-PW is computed using the HAC estimator with pre-whitening based on a AR(1) model, as suggested by Andrews and Monahan (1992).

The results of the simulations described above are reported in Table 6.10. It is clear that even when there is no serial correlation, the Wald tests are oversized; F^* does substantially better. In unreported simulations, it was seen that the size of the Wald tests does come closer to 5% as the number of observations increase. The size-adjusted power of F^* is always slightly lower than that of the two Wald tests. Even when serial correlation is high, the size of F^* is generally less distorted than that of the Wald tests. The exception to this statement is the Wald-PW test, which is less distorted, when the moving average coefficient of the data generating process is positive. For this test, unreported simulations indicate that size approaches 0 as the MA coefficient and/or sample size increases. Further investigation is required to determine the exact circumstances in which the Wald-PW test will perform this well. In terms of size, the Wald test which does not use prewhitening, is clearly dominated by the other two test statistics,[11] but in terms of size-adjusted power, the Wald statistic which does not use prewhitening does better than the other two.

[11] Even when the size distortions of the new test are lower than those of the Wald tests, the test is still oversized when serial correlation is high. Section 6.5 below suggests a way of reducing the size distortion.

The second set of simulations are designed to compare the new test statistic to two test statistics developed in Vogelsang (1998). To define those test statistics, it is necessary to introduce the model:

$$\breve{y}_t = \delta_1 t + \delta_2 \left[\frac{1}{2}\left(t^2 + t\right)\right] + E_t, \tag{6.5}$$

where $\breve{y}_t = \sum_{i=1}^{t} y_i$ and $E_t = \sum_{i=1}^{t} u_i$. Let $\hat{\delta}$ be the OLS estimate of $\delta' = [\delta_1' \ \delta_2']$, and $\breve{s}^2$ be the OLS estimate of the error variance. Finally, let $\breve{\mathbf{f}}(T)$ be the matrix of the stacked trend functions in (6.5). Then the two relevant test statistics from Vogelsang (1998) can be defined as follows:

$$PS_T = T^{-1}\left(\hat{\delta}_2\right)'\left[R\left(\breve{\mathbf{f}}(T)'\breve{\mathbf{f}}(T)'\right)^{-1}R'\right]^{-1}\hat{\delta}_2/\breve{s}^2,$$

and

$$PSW_T = \frac{(\hat{\alpha}_2)'\left[R\left(\mathbf{f}(T)'\mathbf{f}(T)\right)^{-1}R'\right]^{-1}\hat{\alpha}_2}{T^{-1}100\breve{s}^2}.$$

These do not exactly correspond to the test statistics presented in Vogelsang (1998); specifically, the correction for high serial correlation is not employed. This is done so as to enable a more meaningful comparison of the three tests.[12] Note that a comparison of F^* with PS_T and PSW_T is especially interesting because, like F^*, neither PS_T nor PSW_T rely on a direct estimate of the covariance matrix and the asymptotic distributions of these test statistics do not depend on parameters relating to the error structure. In fact, PSW_T can be viewed as belonging to the same class of test statistics as F^*; it too utilizes a data-dependent transformation of the OLS parameter estimates for the purpose of eliminating the nuisance parameters.

[12]Section 6.5 presents a comparison where the correction for high serial correlation is implemented for all three test statistics.

Table 6.11: Simulations: Finite Samplesize and Power of F*, PS and PSW

	$\eta = 0.0$					$\eta = -0.4$					$\eta = 0.4$			
ρ	α_2	F*	PS	PSW	ρ	α_2	F*	PS	PSW	ρ	α_2	F*	PS	PSW
	0.00	0.171	0.238	0.226		0.00	0.170	0.235	0.223		0.00	0.173	0.239	0.225
	0.10	0.189	0.200	0.225		0.10	0.363	0.407	0.445		0.10	0.127	0.129	0.141
	0.20	0.479	0.535	0.592		0.20	0.771	0.848	0.891		0.20	0.313	0.341	0.383
0.90	0.30	0.721	0.796	0.852	0.90	0.30	0.944	0.981	0.990	0.90	0.30	0.525	0.584	0.638
	0.40	0.873	0.936	0.961		0.40	0.992	0.999	1.000		0.40	0.694	0.768	0.821
	0.50	0.951	0.983	0.991		0.50	0.999	1.000	1.000		0.50	0.819	0.889	0.927
ρ	α_2	F*	PS	PSW	ρ	α_2	F*	PS	PSW	ρ	α_2	F*	PS	PSW
	0.00	0.225	0.318	0.301		0.00	0.224	0.304	0.291		0.00	0.225	0.319	0.305
	0.10	0.121	0.126	0.133		0.10	0.230	0.256	0.275		0.10	0.086	0.089	0.093
	0.20	0.299	0.338	0.361		0.20	0.565	0.648	0.693		0.20	0.186	0.206	0.220
0.95	0.30	0.502	0.577	0.615	0.95	0.30	0.801	0.891	0.922	0.95	0.30	0.329	0.371	0.404
	0.40	0.673	0.765	0.805		0.40	0.927	0.977	0.985		0.40	0.474	0.546	0.585
	0.50	0.794	0.885	0.914		0.50	0.976	0.997	0.998		0.50	0.601	0.692	0.737
ρ	α_2	F*	PS	PSW	ρ	α_2	F*	PS	PSW	ρ	α_2	F*	PS	PSW
	0.00	0.401	0.508	0.509		0.00	0.391	0.492	0.492		0.00	0.402	0.509	0.511
	0.10	0.071	0.074	0.075		0.10	0.107	0.118	0.117		0.10	0.062	0.062	0.061
	0.20	0.131	0.138	0.142		0.20	0.263	0.305	0.310		0.20	0.093	0.095	0.097
1.00	0.30	0.221	0.249	0.255	1.00	0.30	0.447	0.547	0.554	1.00	0.30	0.142	0.150	0.153
	0.40	0.331	0.385	0.398		0.40	0.617	0.739	0.755		0.40	0.206	0.227	0.233
	0.50	0.442	0.525	0.536		0.50	0.744	0.870	0.879		0.50	0.282	0.320	0.333

Model: $y_t = \alpha_1 + \alpha_2 t + u_t$

N = 10,000, T = 40

H_0: $\alpha_2 = 0$

In the simulation experiment, the hypothesis being tested is $H_0 : \alpha_2 = 0$ against the alternative $\alpha_2 \neq 0$. Again, the errors are generated according to $u_t = \rho u_{t-1} + e_t + \eta e_{t-1}$, where e_t is *i.i.d.* $N(0,1)$. Finite sample size and size-adjusted power is reported for sample size $T = 40$, and for $\rho = 0.8, 0.9, 0.95, 1$ and $\eta = -0.4$, 0, 0.4. Simulations for smaller ρ were performed, but for these parameter values the three statistics are indistinguishable; hence the simulations are not reported here. As is clear from Table 6.11, size for all three tests is substantially higher than 5% for high values of ρ; moreover, the size of F^* is significantly less than that of PS_T and PSW_T. In terms of size-adjusted power, however, PS_T and PSW_T perform uniformly better than F^*. In unreported simulations, it was observed that these results prevail for sample sizes greater than 40 as well.

Given the substantially less-distorted size of F^*, there is reason to believe that the power of F^* relative to PS_T and PSW_T may improve when the adjustment for high serial correlation is implemented. The intuition behind this observation is that the adjustment is primarily designed to get size right; as usual the "cost" is lower power.

6.5 Correction for High Serial Correlation

From the simulation experiments in the previous section, it is clear that size becomes distorted when serial correlation is high. An examination of the asymptotic distribution of the test statistic when the errors contain unit roots may provide insight into the cause of this systematic size distortion. This is because the asymptotic distribution of F^* when the errors have unit roots, is "closer" to the actual distribution than the distribution of F^* described in Theorem 12.

Lemma 5 *Suppose u is $I(1)$. Then*

(a) Under H_0^X,

$$F^* \Rightarrow \left[\int_0^1 \hat{w}_q^F(s)\, w_1(s)\, ds\right]' \left[\int_0^1 V_u(r)\, V_u(r)'\, dr\right]^{-1} \left[\int_0^1 \hat{w}_q^F(s)\, w_1(s)\, ds\right] / q$$

where

$$\begin{aligned} V_u(r) &= \int_0^r \hat{w}_q^F(s)\, w_1(s)\, ds - \left(\int_0^r \hat{w}_q^F(s)'\, \hat{w}_q^F(s)\, ds\right) \times \\ &\quad \left(\int_0^1 \hat{w}_q^F(s)'\, \hat{w}_q^F(s)\, ds\right)^{-1} \int_0^1 \hat{w}_q^F(s)\, w_1(s)\, ds. \end{aligned}$$

(b) Under H_0^F, if $k > 0$,

$$\begin{aligned} F^* &\Rightarrow \left(R^* \left(\int_0^1 F^X(s)\, F^X(s)'\, ds\right)^{-1} \int_0^1 F^X(s)\, w_1(s)\, ds\right)' \\ &\quad \left[R^* \left(\int_0^1 F^X(s)\, F^X(s)'\, ds\right)^{-1} \int_0^1 V_u^{F^X}(r) \left(V_u^{F^X}(r)\right)' dr \right. \\ &\quad \left. \left(\int_0^1 F^X(s)\, F^X(s)'\, ds\right)^{-1} (R^*)'\right]^{-1} \\ &\quad \left(R^* \left(\int_0^1 F^X(s)\, F^X(s)'\, ds\right)^{-1} \int_0^1 F^X(s)\, w_1(s)\, ds\right) \end{aligned}$$

where

$$\begin{aligned} V_u^{F^X}(r) &= \int_0^r F^X(s)\, w_1(s)\, ds - \left(\int_0^r F^X(s)\, F^X(s)'\, ds\right) \times \\ &\quad \left(\int_0^1 F^X(s)\, F^X(s)'\, ds\right)^{-1} \int_0^1 F^X(s)\, w_1(s)\, ds. \end{aligned}$$

(c) Under H_0^F, if $k=0$,

$$F^* \Rightarrow \left(R^*\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 F(s)w_1(s)ds\right)' \left[R^*\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 V_u^F(r)\left(V_u^F(r)\right)'dr \left(\int_0^1 F(s)F(s)'ds\right)^{-1}(R^*)'\right]^{-1} \left(R^*\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 F(s)w_1(s)ds\right)$$

where

$$V_u^F(r) = \int_0^r F(s)w_1(s)ds - \left(\int_0^r F(s)F(s)'ds\right)\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\times \int_0^1 F(s)w_1(s)ds.$$

As before, the asymptotic distribution of the test statistic does not depend on nuisance parameters. The fact that F^* (when the errors contain unit roots) converges to a distribution is important. If F^* *diverged*, it would not be possible to implement this type of size correction.

The analytical expressions for the asymptotic distributions themselves do not provide much insight; nonetheless, they can be used to simulate the critical values when the errors are $I(1)$. This has been done for the sets of models and hypotheses used for the simulation experiments in the previous section. Table 6.12 tabulates the critical values for model (6.4) when the errors are stationary and when they contain a unit root. Table 6.13 tabulates the corresponding critical values for model (6.3).

Clearly the critical values for the unit root case are much higher for both models. This implies that the true hypothesis will be rejected too frequently when the $I(0)$

Table 6.12: Critical Values, Unit Root Errors, Model (6.4)

u / level	90%	95%	99%
$I(0)$	55.7	85.3	164.0
$I(1)$	2,520	4,390	11,500

Table 6.13: Critical Values, Unit Root Errors, Model (6.3)

u / level	90%	95%	99%
$I(0), q = 1$	42.66	68.09	155.18
$I(1), q = 1$	136.41	230.66	634.57

critical values are used for data which is approximately $I(1)$; hence the inflated size.

It is possible to enhance the performance of the test statistic when serial correlation is high. The principle is to modify the test in such a way that the relevant critical values remain the same, even when the errors are $I(1)$. Note that the procedure does not alter the asymptotic performance of the test when the errors are *not* highly correlated. As such, it may be interpreted as smoothing the discontinuity in the asymptotic distribution that arises when the error term moves from a stationary to a unit root process. This method is similar to that employed in Vogelsang (1998).

In this chapter the alteration of the test statistic will be implemented for the special case where there is just a constant and a linear trend in the model. The alteration utilizes a unit root test introduced in Park and Choi (1988) and Park (1990). Some additional notation is required to give a precise definition of the modified test statistic. Consider the regression

$$y_t = \alpha_1 + \alpha_2 t + \sum_{i=3}^{9} \alpha_i t^i + u_t. \tag{6.6}$$

Let J_T denote the standard OLS Wald statistic, normalized by T^{-1}, used to test the hypothesis $\alpha_3 = ... = \alpha_9 = 0$.[13] Then $J_T = (SSR_R - SSR_U)/SSR_U$, where SSR_U is the sum of squared residuals obtained from the estimation of (6.6) by OLS, and SSR_R be the sum of squared residuals from the OLS estimation of (6.4). Let F^J denote the modified test statistic. Then

$$F^J = T\,(R\hat{\alpha} - r)' \left[R\hat{B}R'\right]^{-1} (R\hat{\alpha} - r) \,/\, (q \exp(bJ_T)\,,$$

where b is a constant. In Appendix D.6 it is shown that the asymptotic distribution of F^J is the same as that of F^* when the errors are stationary. When they are not, the asymptotic distribution of F^J is different from that of F^*. While the asymptotic distribution of F^J does not depend on parameters relating to the correlation structure, it *does* depend on b. For each nominal confidence level, b can be calculated such that the critical value is the same, irrespective of whether the errors in the cointegration relationship are $I\,(0)$ or $I\,(1)$.

The simulation experiment comparing F^* with PS_T and PSW_T from the previous section is now repeated, but with the correction for high serial correlation implemented for all three statistics. The results are reported in Table 6.14. Not surprisingly, the size of all three tests is now substantially closer to 5%, and unreported simulations verify that as sample size grows, size is less distorted. The difference in size between the three tests is now much less than before the correction was employed. As expected, the differences in size-adjusted power are now much less, and for some parameter values, F^* is more powerful than PS_T and PSW_T.

[13] Appendix D.6 describes the procedure used to determine the fact that six additional regressors should be added in (6.6).

Table 6.14: Finite Samplesize and Power, J Correction Employed

	$\eta = 0.0$					$\eta = -0.4$					$\eta = 0.4$			
ρ	α_2	F*	PS	PSW	ρ	α_2	F*	PS	PSW	ρ	α_2	F*	PS	PSW
	0.00	0.014	0.014	0.013		0.00	0.042	0.042	0.041		0.00	0.009	0.008	0.007
	0.10	0.124	0.120	0.119		0.10	0.255	0.267	0.272		0.10	0.087	0.082	0.083
	0.20	0.231	0.231	0.230		0.20	0.486	0.501	0.504		0.20	0.152	0.147	0.146
0.9	0.30	0.310	0.310	0.301	0.90	0.30	0.615	0.635	0.632	0.90	0.30	0.208	0.199	0.197
	0.40	0.363	0.356	0.349		0.40	0.694	0.709	0.704		0.40	0.244	0.240	0.235
	0.50	0.404	0.397	0.388		0.50	0.747	0.759	0.752		0.50	0.274	0.268	0.261
ρ	α_2	F*	PS	PSW	ρ	α_2	F*	PS	PSW	ρ	α_2	F*	PS	PSW
	0.00	0.020	0.019	0.019		0.00	0.055	0.056	0.055		0.00	0.012	0.012	0.011
	0.10	0.086	0.081	0.084		0.10	0.171	0.173	0.177		0.10	0.066	0.067	0.067
	0.20	0.161	0.156	0.158		0.20	0.346	0.362	0.365		0.20	0.104	0.104	0.106
0.95	0.30	0.226	0.216	0.215	0.95	0.30	0.464	0.481	0.482	0.95	0.30	0.147	0.143	0.145
	0.40	0.271	0.264	0.264		0.40	0.550	0.564	0.559		0.40	0.179	0.174	0.174
	0.50	0.306	0.297	0.294		0.50	0.604	0.624	0.618		0.50	0.205	0.201	0.199
ρ	α_2	F*	PS	PSW	ρ	α_2	F*	PS	PSW	ρ	α_2	F*	PS	PSW
	0.00	0.045	0.044	0.043		0.00	0.115	0.122	0.120		0.00	0.028	0.025	0.025
	0.10	0.061	0.059	0.058		0.10	0.088	0.091	0.088		0.10	0.055	0.053	0.053
	0.20	0.087	0.088	0.086		0.20	0.180	0.192	0.189		0.20	0.065	0.066	0.066
1.00	0.30	0.126	0.125	0.123	1.00	0.30	0.275	0.296	0.287	1.00	0.30	0.082	0.086	0.084
	0.40	0.160	0.161	0.157		0.40	0.358	0.379	0.369		0.40	0.106	0.108	0.105
	0.50	0.192	0.194	0.187		0.50	0.420	0.442	0.430		0.50	0.126	0.126	0.124

Model: $y_t = \alpha_1 + \alpha_2 t + u_t$

N = 10,000, T = 40

H_0: $\alpha_2 = 0$

6.6 Conclusion

In this chapter a new test statistic that is robust to serial correlation/heteroskedasticity of unknown form is developed. The environment is a single-equation model of cointegration that incorporates linear polynomial trend functions. The standard approach used to deal with heteroskedasticity and serial correlation in models of this type has been to estimate the correlation structure of the error terms. While such a technique generates consistent estimates of the correlation structure, the possibility of substantial size distortions in finite samples remain. The test proposed in this chapter eliminates the need to estimate the correlation structure, and hence removes an important source of size distortion.

To evaluate the finite sample performance of the new test, simulation experiments were performed. It was shown that in general, size distortions are much less severe than those of tests currently employed in the literature, and that the size-adjusted power of the new test is only marginally lower than that of tests currently employed. Furthermore, a size correction procedure was implemented for a simple model, which illustrated that size distortion can be virtually eliminated, even when the extent of serial correlation is substantial.

A fruitful extension of the approach outlined in this chapter would be the extension to models of cointegration, where the regressors are not exogenous. The F^* statistic cannot be used directly, as the asymptotic distribution in this case depends on the parameters pertaining to the correlation structure. One possible way of getting around this problem, is to use the dynamic OLS regression suggested by Saikkonen (1991), Phillips and Loretan (1991), Stock and Watson (1993), and Woolridge (1991). It seems likely that F^* can be applied directly to this regression, although a proof is required. This method would still require a choice of the number of lags and leads of the regressors to be added to the regression, but the benefits from circumventing the estimation of the serial correlation parameters would probably still be present.

Another possible way of dealing with endogenous regressors is to find a different transformation matrix than the one used in this chapter. This transformation should yield an asymptotic distribution of the transformed parameter estimates which does not depend on the parameters pertaining to the correlation structure. It is not even clear, however, that such a transformation can be found. The reason for this is, that the "nice" multiplicative manner in which the nuisance parameters enter the distribution in this chapter would be lost. For illustrative purposes examine the model with no trends and define

$$\begin{aligned}
\Omega_0 &= \lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} E \begin{bmatrix} u_t u_t' & u_t v_t' \\ v_t u_t' & v_t v_t' \end{bmatrix} \\
\Omega_1 &= \lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{t-1} E \begin{bmatrix} u_j u_t' & u_j v_t' \\ v_j u_t' & v_j v_t' \end{bmatrix} \\
\Omega &= \Omega_0 + \Omega_1 + \Omega_1'.
\end{aligned}$$

Then it can be verified that the following version of Theorem 4.1 in Phillips and Durlauf (1986) holds:

$$\begin{aligned}
T\left(\hat{\theta} - \theta\right) \Rightarrow & \left\{ \left[\Omega^{\frac{1}{2}} \int_0^1 w_{k+1}(s)\, w_{k+1}(s)'\, ds \Omega^{\frac{1}{2}} \right]_{22} \right\} \\
& \left[\Omega^{\frac{1}{2}} \int_0^1 w_{k+1}(s)\, w_{k+1}(s)'\, ds \Omega^{\frac{1}{2}} + \Omega_0 + \Omega_1 \right]_{21}
\end{aligned}$$

where $]_{22}$ and $]_{21}$ refers to the specific block of the matrix in question, when partitioned the natural way. Clearly, because of the way the nuisance parameters enter into the distribution, the problem of transforming the parameters has become much harder, and such a problem may or may not have a solution.

The last issue discussed here is again a direction for future research. The topic at hand is the choice of the transformation matrix. While the transformation discussed in this chapter ensures that the asymptotic distribution of the parameters does not depend on nuisance parameters, and the test derived therefrom seems to perform well, there are potentially a large number of data transformations that could perform the task equally well or better. Specifically, *all* data-dependent matrices that are asymptotically proportional to Σ are candidates. In order to find a way of choosing a specific transformation matrix from this large set, an analytical analysis of optimality is required. The first step in this direction would be to exactly determine the full class of tests to which this statistic belongs. Clearly it would have to exclude tests which rely upon estimation of the correlation structure, since these would always dominate asymptotically. Then, from within this class, the optimal test or tests, if any, would have to found.

Chapter 7

Conclusion

Economists write down simple theoretical models to explain certain economic phenomena. At the point at which these models are subjected to the scrutiny of data, the econometrician typically converts these models into econometric models to facilitate the "testing". At this stage of the scientific process, two sets of parameters typically emerge from an econometric model: parameters that are of interest to the economist and about which the theory has something to say, and those that are not, about which the theory may have very little to say, called nuisance parameters. For example, economic theory may have a lot to say about the male elasticity of labor supply with respect to the wage in a labor market model but may have little to contribute as far as how the standard deviation of sample wage data may behave for a given sample. This creates a problem in the sense that while statistical inference is drawn only on the parameters of interest, the treatment accorded by the investigator to the nuisance parameters can significantly contaminate the results. The chapters in this dissertation have tried to grapple with a common theme: how should one conduct proper statistical inference on the parameters of interest when nuisance parameters are present.

More specifically, means of conducting inference on the parameters of interest, that are robust to the structure implied by the nuisance parameters, were studied. One line of research explored the possibilities of conducting separate inference on the parameters of interest in general likelihood models, which contain nuisance parameters. This work was built upon the theory of local cuts. The other line of research concerns hypothesis testing in models with serial correlation or heteroskedasticity of unknown form (the entire error structure is the nuisance parameter here). A test statistic that is robust to different error structures, (and does not require an actual estimate of the error structure) is developed.

The second chapter explored general likelihood models that have two sets of parameters; the parameters of interest and the nuisance parameters. It defined local cuts and adaptivity as well as marginal and conditional local cuts and examined the properties of models that allow for local cuts or adaptive estimators. It is shown that block diagonality of the Fisher information matrix is a central requirement for both local cuts and adaptivity when dealing with regular models. In order to obtain a local cut in a model where an adaptive estimator is provided it is shown that the information matrix must be insensitive to the nuisance parameters and the maximum likelihood estimates of all parameters need to be well behaved. On the other hand, obtaining an adaptive estimate in a model which allows for a local cut just requires that a well behaved estimate of the parameters can be found.

When local cuts are used to justify separate inference, it is sometimes the case that some estimate of the nuisance parameter is required. This need for an estimate arises because a local cut only ensures that the score function will be free of nuisance parameters asymptotically. The finite sample score used for estimation, however, may depend on the nuisance parameters. Looking ahead then, there is some need to explore the properties this estimate of the nuisance parameter must satisfy. In other words, we need to answer the question: How 'bad' an estimate of the nuisance parameter can we get away with?

Another direction for future research on local cuts is the extension of local cuts to semi-parametric models. This is important because the situation where the nuisance parameter is infinite-dimensional is also one where the gain from avoiding estimation is highest.

The third chapter extended the concept of local cuts to an estimating equation environment. The central result here is of prime importance because it establishes a very clear and direct relationship between local cuts in the estimating equation framework and the ability to conduct separate inference. Because the estimates obtained from estimating equations are invariant to all full-rank multiplicative transformations of the estimating equation, while the asymptotic distribution of the estimating equation is not, we defined a transformation of the estimating equation which eliminated the indeterminacy. It is precisely because of the specific properties of this transformed estimating equation that the above-mentioned relationship emerges. Subsequent chapters exploited this result within the context of a dynamic regression model.

The fourth chapter laid the foundation for a new test statistic that is robust to serial correlation/heteroskedasticity of unknown form. The statistic is developed to test hypotheses in linear regression models of the form introduced in the third chapter. The novelty here is that the tests are simple and do not require heteroskedasticity and autocorrelation consistent (HAC) estimators; hence the size distortion caused by the estimation of the correlation structure is eliminated. The development of the new test relies upon a data-dependent transformation of the ordinary least squares estimates of the parameters, this is the exact transformation used in the third chapter to show that a local cut existed in this framework. Examples were used to illustrate that the size of the new test is usually less distorted and finite sample power greater than tests that utilize HAC estimators.

The fifth chapter extends the test statistics introduced in the fourth chapter to a non-linear weighted regression environment. It is established that the class of tests introduced in the third chapter is applicable in this framework as well. Given that these new tests compare favorably to HAC methods in finite samples and are simpler to compute, one speculates that they may attract a wide body of users.

It also bears emphasis here that the tests presented in the previous two chapters introduced a new class of tests which are pivotal and robust to heteroskedasticity and serial correlation in the errors. While the selection of the specific test statistics within this class is somewhat arbitrary (there are many conceivable transformations that will yield asymptotic pivotal statistics), their properties with respect to invariance to nuisance parameters and minor size distortion are highly desirable. An open research problem is to develop a theory of optimality for this new class of tests.

In the sixth chapter, the techniques introduced in the fourth and fifth chapters are employed to develop a test statistic that is robust to serial correlation/heteroskedasticity of unknown form in a cointegration environment that incorporates linear polynomial trend functions. The test can be employed to conduct inference on the trend function or the cointegration vector in a cointegration relationship, and to test hypotheses about the parameters of the deterministic trend function of a univariate time series. Extensive simulation experiments revealed that size distortions are generally less than those of tests currently employed in the literature with no associated (substantial) reduction in power. A fruitful extension of the approach outlined in this chapter would be the extension to models of cointegration with endogenous regressors or with multiple cointegrating vectors.

Appendix A

Appendices for Chapter 2

A.1 Detailed calculations for Example 1

First we write out the expression of the two relevant terms:

$$\begin{aligned}
\ln p\left(x;\mu,\sigma^2|T(x)\right) &= -\tfrac{(n-1)}{2}\ln\left(2\pi\sigma^2\right)-\tfrac{1}{2}\ln n \\
&\quad -\tfrac{1}{2\sigma^2}\left[\sum x_i^2 - nT(x)^2\right] \\
\ln p\left(x;\mu,\sigma^2+\frac{\varepsilon}{\sqrt{n}}|T(x)\right) &= -\tfrac{(n-1)}{2}\ln\left(2\pi\left(\sigma^2+\tfrac{\varepsilon}{\sqrt{n}}\right)\right)-\tfrac{1}{2}\ln n \\
&\quad -\tfrac{1}{2\left(\sigma^2+\frac{\varepsilon}{\sqrt{n}}\right)}\left[\sum x_i^2 - nT(x)^2\right]
\end{aligned}$$

Examining the difference, we get the following expression:

$$\begin{aligned}
&\ln p\left(x;\mu,\sigma^2+\frac{\varepsilon}{\sqrt{n}}|T(x)\right)-\ln p\left(x;\mu,\sigma^2|T(x)\right) \\
&= -\tfrac{(n-1)}{2}\ln\left(2\pi\left(\sigma^2+\tfrac{\varepsilon}{\sqrt{n}}\right)\right)+\frac{(n-1)}{2}\ln\left(2\pi\sigma^2\right) \\
&\quad -\tfrac{1}{2\left(\sigma^2+\frac{\varepsilon}{\sqrt{n}}\right)}\left[\sum x_i^2 - nT(x)^2\right]+\tfrac{1}{2\sigma^2}\left[\sum x_i^2 - nT(x)^2\right] \\
&= -\tfrac{(n-1)}{2}\left\{\ln\left(2\pi\left(\sigma^2+\tfrac{\varepsilon}{\sqrt{n}}\right)\right)-\ln\left(2\pi\sigma^2\right)\right\} \\
&\quad +\tfrac{1}{2}\left[\sum x_i^2 - nT(x)^2\right]\left[\tfrac{1}{\sigma^2}-\tfrac{1}{\left(\sigma^2+\frac{\varepsilon}{\sqrt{n}}\right)}\right] \\
&= -\tfrac{(n-1)}{2}\ln\left(\frac{2\pi\left(\sigma^2+\frac{\varepsilon}{\sqrt{n}}\right)}{2\pi\sigma^2}\right)+\tfrac{1}{2}\left[\sum x_i^2 - nT(x)^2\right]\left[\frac{\varepsilon/\sqrt{n}}{\sigma^2\left(\sigma^2+\varepsilon/\sqrt{n}\right)}\right] \\
&= -\tfrac{(n-1)}{2}\ln\left(1+\tfrac{\varepsilon}{\sigma^2\sqrt{n}}\right)+\tfrac{1}{2}\left[\sum x_i^2 - nT(x)^2\right]\left[\frac{\varepsilon/\sqrt{n}}{\sigma^2\left(\sigma^2+\varepsilon/\sqrt{n}\right)}\right]
\end{aligned}$$

We will now look at the terms separately before proceeding any further. First we use a standard Taylor expansion on the first term:

$$\begin{aligned}
\ln\left(1+\tfrac{\varepsilon}{\sigma^2\sqrt{n}}\right) &= \tfrac{\varepsilon}{\sigma^2\sqrt{n}} + o\left(n^{-1}\right) \\
-\tfrac{(n-1)}{2}\ln\left(1+\tfrac{\varepsilon}{\sigma^2\sqrt{n}}\right) &= \left(-\frac{n}{2}+\frac{1}{2}\right)\left(\tfrac{\varepsilon}{\sigma^2\sqrt{n}} + o\left(n^{-1}\right)\right) \\
&= -\frac{1}{2}\tfrac{\varepsilon\sqrt{n}}{\sigma^2} + \frac{1}{2}\tfrac{\varepsilon}{\sigma^2\sqrt{n}} - \frac{n}{2}o\left(n^{-1}\right) + o\left(n^{-1}\right) \\
&= -\frac{1}{2}\tfrac{\varepsilon\sqrt{n}}{\sigma^2} + o(1)
\end{aligned}$$

We now proceed to rewrite the stochastic terms, such that all sums are of variables with mean 0, allowing us to apply a central limit theorem:

$$\begin{aligned}
\sum x_i^2 &= \sum\left[x_i^2 - \left(\mu^2+\sigma^2\right)\right] + n\left(\mu^2+\sigma^2\right) \\
&= \sqrt{n}\left\{\frac{1}{\sqrt{n}}\sum\left[x_i^2 - \left(\mu^2+\sigma^2\right)\right]\right\} + n\left(\mu^2+\sigma^2\right)
\end{aligned}$$

$$\begin{aligned}
nT(x)^2 &= n\left[\frac{1}{n}\sum\left(x_i-\mu\right)+\mu\right]^2 \\
&= n\left[\frac{1}{n}\sum\left(x_i-\mu\right)\right]^2 + n\mu^2 + n2\mu\frac{1}{n}\sum\left(x_i-\mu\right) \\
&= n\left[\frac{1}{\sqrt{n}}\left\{\frac{1}{\sqrt{n}}\sum\left(x_i-\mu\right)\right\}\right]^2 + n\mu^2 + 2\mu\sqrt{n}\left\{\frac{1}{\sqrt{n}}\sum\left(x_i-\mu\right)\right\} \\
&= \left[\left\{\frac{1}{\sqrt{n}}\sum\left(x_i-\mu\right)\right\}\right]^2 + n\mu^2 + 2\mu\sqrt{n}\left\{\frac{1}{\sqrt{n}}\sum\left(x_i-\mu\right)\right\}
\end{aligned}$$

$$\begin{aligned}
\sum x_i^2 - nT(x)^2 &= \sqrt{n}\left\{\frac{1}{\sqrt{n}}\sum\left[x_i^2-\left(\mu^2+\sigma^2\right)\right]\right\} + n\left(\mu^2+\sigma^2\right) \\
&\quad -\left[\left\{\frac{1}{\sqrt{n}}\sum\left(x_i-\mu\right)\right\}\right]^2 - n\mu^2 \\
&\quad -2\mu\sqrt{n}\left\{\frac{1}{\sqrt{n}}\sum\left(x_i-\mu\right)\right\} \\
&= \sqrt{n}\left\{\frac{1}{\sqrt{n}}\sum\left[x_i^2-\left(\mu^2+\sigma^2\right)\right]\right\} - \left[\left\{\frac{1}{\sqrt{n}}\sum\left(x_i-\mu\right)\right\}\right]^2 \\
&\quad -2\mu\sqrt{n}\left\{\frac{1}{\sqrt{n}}\sum\left(x_i-\mu\right)\right\} + n\sigma^2
\end{aligned}$$

Finally we put it all together to obtain the rate:

$$
\begin{aligned}
& -\tfrac{(n-1)}{2}\ln\left(1+\tfrac{\varepsilon}{\sigma^2\sqrt{n}}\right)+\tfrac{1}{2}\left[\sum x_i^2-nT(x)^2\right]\left[\frac{\varepsilon/\sqrt{n}}{\sigma^2\left(\sigma^2+\varepsilon/\sqrt{n}\right)}\right] \\
= & -\frac{1}{2}\frac{\varepsilon\sqrt{n}}{\sigma^2}+o(1)+\tfrac{1}{2}\left[\tfrac{\varepsilon/\sqrt{n}}{\sigma^2(\sigma^2+\varepsilon/\sqrt{n})}\right]\sqrt{n}\left\{\frac{1}{\sqrt{n}}\sum\left[x_i^2-\left(\mu^2+\sigma^2\right)\right]\right\} \\
& -\tfrac{1}{2}\left[\tfrac{\varepsilon/\sqrt{n}}{\sigma^2(\sigma^2+\varepsilon/\sqrt{n})}\right]\left[\left\{\frac{1}{\sqrt{n}}\sum(x_i-\mu)\right\}\right]^2 \\
& -\tfrac{1}{2}\left[\tfrac{\varepsilon/\sqrt{n}}{\sigma^2(\sigma^2+\varepsilon/\sqrt{n})}\right]2\mu\sqrt{n}\left\{\frac{1}{\sqrt{n}}\sum(x_i-\mu)\right\}+\tfrac{1}{2}\left[\tfrac{\varepsilon/\sqrt{n}}{\sigma^2(\sigma^2+\varepsilon/\sqrt{n})}\right]n\sigma^2 \\
= & -\frac{1}{2}\frac{\varepsilon\sqrt{n}}{\sigma^2}+\tfrac{1}{2}\left[\tfrac{\varepsilon\sqrt{n}}{(\sigma^2+\varepsilon/\sqrt{n})}\right]+o(1)+\tfrac{1}{2}\left[\tfrac{\varepsilon}{\sigma^2(\sigma^2+\varepsilon/\sqrt{n})}\right]\left\{\tfrac{1}{\sqrt{n}}\sum\left[x_i^2-\left(\mu^2+\sigma^2\right)\right]\right\} \\
& -\tfrac{1}{2}\left[\tfrac{\varepsilon/\sqrt{n}}{\sigma^2(\sigma^2+\varepsilon/\sqrt{n})}\right]\left[\left\{\tfrac{1}{\sqrt{n}}\sum(x_i-\mu)\right\}\right]^2-\left[\tfrac{\varepsilon\mu}{\sigma^2(\sigma^2+\varepsilon/\sqrt{n})}\right]\left\{\frac{1}{\sqrt{n}}\sum(x_i-\mu)\right\} \\
= & \ \tfrac{\varepsilon\sqrt{n}}{2}\left[\tfrac{\varepsilon/\sqrt{n}}{\sigma^2(\sigma^2+\varepsilon/\sqrt{n})}\right]+o(1)+O_P(1)+O\left(n^{-\frac{1}{2}}\right)\cdot O_P(1)+O_P(1) \\
= & \ O(1)+O_P(1)=O_P(1)
\end{aligned}
$$

To obtain this result, note that all terms in $\{\}$ converge to normal distributions.

A.2 Proof of Theorem 1

Since the model allows for a local cut, $\frac{\partial^2}{\partial\upsilon\partial\eta}\ln p(\mathbf{x};\upsilon,\eta)$, correctly normalized, will converge to 0 as $n\to\infty$. Therefore, any estimates which are asymptotically equivalent to the MLE estimates will have distributions determined as follows:

$$
\begin{aligned}
(\hat{\upsilon}-\upsilon) &= \left(\frac{\partial^2}{\partial\upsilon^2}\ln p(\mathbf{x};\upsilon,\eta|T)\right)^{-1}\frac{\partial}{\partial\upsilon}\ln p(\mathbf{x};\upsilon,\eta|T) \\
(\hat{\eta}-\eta) &= \left(\frac{\partial^2}{\partial\eta^2}\ln p(T;\upsilon,\eta)\right)^{-1}\frac{\partial}{\partial\eta}\ln p(T;\upsilon,\eta).
\end{aligned}
$$

Thus, for separate inference to be justified, it is required that

$$
\frac{\partial}{\partial\eta}\left[\lim_{n\to\infty}\frac{\partial^2}{\partial\upsilon^2}\ln p(\mathbf{x};\upsilon,\eta|T)\right]=\frac{\partial}{\partial\upsilon}\left[\lim_{n\to\infty}\frac{\partial^2}{\partial\eta^2}\ln p(T;\upsilon,\eta)\right]=0.
$$

This is satisfied because of the assumption that the relevant corners of the Fisher information matrix does not depend on nuisance parameters. The final requirements for separate inference to be justified is that

$$\frac{\partial}{\partial\eta}\left[\lim_{n\to\infty}\frac{\partial}{\partial\upsilon}\ln p\left(\mathbf{x};\upsilon,\eta|T\right)\right]=\frac{\partial}{\partial\upsilon}\left[\lim_{n\to\infty}\frac{\partial}{\partial\eta}p\left(T;\upsilon,\eta\right)\right]=0.$$

To prove that this holds, first note that we know

$$\begin{aligned}\frac{\partial}{\partial\upsilon}\ln p\left(\mathbf{x};\upsilon,\eta|T\right)\frac{\delta}{\sqrt{n}} &= O_p\left(n^{-\frac{s}{2}}\right)\Leftrightarrow\\ \frac{\partial}{\partial\upsilon}\ln p\left(\mathbf{x};\upsilon,\eta|T\right) &= O_p\left(n^{-\frac{s}{2}+\frac{1}{2}}\right)\end{aligned}$$

such that

$$n^{\frac{s}{2}-\frac{1}{2}}\frac{\partial}{\partial\upsilon}\ln p\left(\mathbf{x};\upsilon,\eta|T\right)=O_P\left(1\right).$$

This implies that the term of interest is

$$\begin{aligned}\frac{\partial}{\partial\eta}\lim_{n\to\infty}\left[n^{\frac{s}{2}-\frac{1}{2}}\frac{\partial}{\partial\upsilon}\ln p\left(\mathbf{x};\upsilon,\eta|T\right)\right] &= \lim_{n\to\infty}\left[n^{\frac{s}{2}-\frac{1}{2}}\frac{\partial^2}{\partial\upsilon\partial\eta}\ln p\left(\mathbf{x};\upsilon,\eta|T\right)\right]\\ &= \lim_{n\to\infty}\left[n^{\frac{s}{2}-\frac{1}{2}}\frac{\partial}{\partial\eta}\ln p\left(\mathbf{x};\upsilon,\eta|T\right)\frac{\partial}{\partial\upsilon}\ln p\left(\mathbf{x};\upsilon,\eta|T\right)\right].\end{aligned}$$

Then we can use the fact that

$$n^{\frac{s}{2}-\frac{1}{2}}\frac{\partial}{\partial\upsilon}\ln p\left(\mathbf{x};\upsilon,\eta|T\right)=O_P\left(1\right)$$

to conclude that a necessary condition for

$$\frac{\partial}{\partial\eta}\lim_{n\to\infty}\left[n^{\frac{s}{2}-\frac{1}{2}}\frac{\partial}{\partial\upsilon}\ln p\left(\mathbf{x};\upsilon,\eta|T\right)\right]=0$$

to hold is that

$$\lim_{n\to\infty}\frac{\partial}{\partial\eta}\ln p\left(\mathbf{x};\upsilon,\eta|T\right)=0.$$

But

$$\frac{\partial}{\partial\eta}\ln p\left(\mathbf{x};\upsilon,\eta|T\right)=O_p\left(n^{-\frac{f}{2}+\frac{1}{2}}\right)$$

so this limit would be 0 if

$$-\frac{f}{2}+\frac{1}{2}<0\Leftrightarrow f>1,$$

but this is simply saying that the local cut is of order greater than 1.

A.3 Proof of Theorem 2

First look at the following expansion:

$$\begin{aligned} & \ln p\left(\mathbf{x}; \upsilon + \frac{\varepsilon_i}{\sqrt{n}}, \eta\right) - \ln p\left(\mathbf{x}; \upsilon, \eta\right) \\ = & \frac{\partial}{\partial \upsilon} \ln p\left(\mathbf{x}; \upsilon, \eta\right) \cdot \frac{\varepsilon_i}{\sqrt{n}} + \frac{\partial^2}{\partial \upsilon \partial \upsilon'} \ln p\left(\mathbf{x}; \tilde{\upsilon}, \eta\right) \cdot \frac{\varepsilon_i^2}{n}, \qquad \tilde{\upsilon} \in [\upsilon, \upsilon + \frac{\varepsilon_i}{\sqrt{n}}] \end{aligned}$$

But, since the MLE is asymptotically normal, we know that $n^{-\frac{1}{2}} \frac{\partial}{\partial \upsilon} \ln p\left(\mathbf{x}; \upsilon, \eta\right)$ converges to a normal distribution and that $\frac{1}{n} \frac{\partial^2}{\partial \upsilon \partial \upsilon'} \ln p\left(\mathbf{x}; \upsilon, \eta\right)$ converges to a negative definite matrix. Therefore it must be the case that

$$\ln p\left(\mathbf{x}; \upsilon + \frac{\varepsilon_i}{\sqrt{n}}, \eta\right) - \ln p\left(\mathbf{x}; \upsilon, \eta\right) = O\left(1\right),$$

but by the definition of a local cut:

$$\begin{aligned} \ln p\left(\mathbf{x}; \upsilon + \frac{\varepsilon_i}{\sqrt{n}}, \eta\right) - \ln p\left(\mathbf{x}; \upsilon, \eta\right) &= \ln p\left(\mathbf{x}; \upsilon + \frac{\varepsilon_i}{\sqrt{n}}, \eta | T\right) - \ln p\left(\mathbf{x}; \upsilon, \eta | T\right) \\ &\quad + \ln p\left(T; \upsilon + \frac{\varepsilon_i}{\sqrt{n}}, \eta\right) - \ln p\left(T; \upsilon, \eta\right) \\ &= O\left(n^{-\frac{1}{2} s_c(\upsilon)}\right) + O\left(n^{-\frac{1}{2} f_m(\upsilon)}\right). \end{aligned}$$

But by the definition of a local cut, $f_m\left(\upsilon\right) \geq s_c\left(\upsilon\right)$ and therefore

$$\ln p\left(\mathbf{x}; \upsilon + \frac{\varepsilon_i}{\sqrt{n}}, \eta\right) - \ln p\left(\mathbf{x}; \upsilon, \eta\right) = O\left(1\right) = O\left(n^{-\frac{1}{2} s_c(\upsilon)}\right)$$

and it must be the case that $s_c\left(\upsilon\right) = 0$. The proof of $s_m\left(\eta\right) = 0$ can be done in exactly the same manner. ■

A.4 Proof of Theorem 3

We need to show that

$$I_{\eta\upsilon}\left(\theta\right) = 0.$$

Now, it is well-known (since the slow rates are 0) that we can write

$$I_{\eta\upsilon}(\theta) = \lim_{n\to\infty}\left[\frac{1}{n}\frac{\partial}{\partial\eta}\frac{\partial}{\partial\upsilon}\ln p(\mathbf{x};\upsilon,\eta)\right].$$

In general,

$$\frac{\partial}{\partial\eta}\frac{\partial}{\partial\upsilon}\ln p(\mathbf{x};\upsilon,\eta) = \frac{\partial}{\partial\eta}\frac{\partial}{\partial\upsilon}\ln p(\mathbf{x};\upsilon,\eta|T) + \frac{\partial}{\partial\eta}\frac{\partial}{\partial\upsilon}\ln p(T;\upsilon,\eta).$$

So it is sufficient to show that

$$\lim_{n\to\infty}\left[\frac{1}{n}\tfrac{\partial}{\partial\eta}\tfrac{\partial}{\partial\upsilon}\ln p(\mathbf{x};\upsilon,\eta|T)\right] = 0 \wedge \lim_{n\to\infty}\left[\frac{1}{n}\tfrac{\partial}{\partial\eta}\tfrac{\partial}{\partial\upsilon}\ln p(T;\upsilon,\eta)\right] = 0.$$

Since the model allows a local cut, the following equation holds:

$$\ln p\left(\mathbf{x};\upsilon,\eta+\tfrac{\varepsilon}{\sqrt{n}}|T\right) - \ln p(\mathbf{x};\upsilon,\eta|T) = O\left(n^{-\frac{1}{2}f_c(\eta)}\right) \tag{A.1}$$

and using the standard Taylor expansion, we also know that

$$\ln p\left(\mathbf{x};\upsilon,\eta+\tfrac{\varepsilon}{\sqrt{n}}|T\right) = \ln p(\mathbf{x};\upsilon,\eta|T) + \frac{\partial}{\partial\eta}\ln p(\mathbf{x};\upsilon,\eta|T)\tfrac{\varepsilon}{\sqrt{n}} + o\left(n^{-1}\right) \Leftrightarrow$$
$$\frac{\partial}{\partial\eta}\ln p(\mathbf{x};\upsilon,\eta|T) = \frac{\sqrt{n}}{\varepsilon}\left\{\ln p\left(\mathbf{x};\upsilon,\eta+\tfrac{\varepsilon}{\sqrt{n}}|T\right) - \ln p(\mathbf{x};\upsilon,\eta|T)\right\} + o\left(n^{-\frac{1}{2}}\right)$$

Using (A.1), we obtain:

$$\frac{\partial}{\partial\eta}\ln p(\mathbf{x};\upsilon,\eta|T) = \frac{\sqrt{n}}{\varepsilon}\left\{O\left(n^{-\frac{1}{2}f_c(\eta)}\right)\right\} + o\left(n^{-\frac{1}{2}}\right) = O\left(n^{-\frac{1}{2}f_c(\eta)+\frac{1}{2}}\right) + o\left(n^{-\frac{1}{2}}\right)$$

Now, since the fast rates are greater than 0, this implies that

$$\frac{\partial}{\partial\eta}\ln p(\mathbf{x};\upsilon,\eta|T) = o\left(n^{-\frac{1}{2}}\right)$$

This now provides us with the desired result:

$$\begin{aligned}\lim_{n\to\infty}\left\{\frac{1}{n}\frac{\partial}{\partial\eta}\ln p(\mathbf{x};\upsilon,\eta|T)\right\} &= 0 \text{ and therefore}\\ \lim_{n\to\infty}\left\{\frac{1}{n}\frac{\partial}{\partial\upsilon}\frac{\partial}{\partial\eta}\ln p(\mathbf{x};\upsilon,\eta|T)\right\} &= 0.\end{aligned}$$

In the same manner it can be shown that

$$\lim_{n\to\infty}\left\{\frac{1}{n}\frac{\partial}{\partial\upsilon}\frac{\partial}{\partial\eta}\ln p\left(T;\upsilon,\eta\right)\right\}=0$$

Hence $I_{\eta\upsilon}\left(\theta\right)=\lim_{n\to\infty}\left\{\frac{1}{n}\frac{\partial}{\partial\upsilon}\frac{\partial}{\partial\eta}\ln p\left(\mathbf{x};\upsilon,\eta\right)\right\}=0.$ ■

A.5 Proof of Theorem 4

By the definition of adaptivity, $\hat{\upsilon}$ is efficient in the model $\mathbf{P}_1\left(\eta\right), \forall\eta$. The efficiency bound in this model is $I_{\upsilon\upsilon}^{-1}$. But in general the variance of $\hat{\upsilon}$, if $\hat{\upsilon}$ is LGR, is $\left(I_{\upsilon\upsilon}-I_{\upsilon\eta}\left(I_{\eta\eta}\right)^{-1}I_{\eta\upsilon}\right)$, so to achieve this bound, it is necessary that $I_{\upsilon\eta}=I_{\eta\upsilon}=0$. ■

A.6 Proof of Theorem 6

By the definition of adaptivity, First note that

$$\begin{aligned}\left(\gamma_1,\gamma_2\right) &= \left(\upsilon,\eta+I_{\upsilon\eta}\left(I_{\eta\eta}\right)^{-1}\upsilon\right)\Leftrightarrow\\ \left(\upsilon,\eta\right) &= \left(\gamma_1,\gamma_2-I_{\upsilon\eta}\left(I_{\eta\eta}\right)^{-1}\gamma_1\right).\end{aligned}$$

Substituting this into the log-likelihood function yields:

$l\left(\upsilon,\eta\right)=l\left(\gamma_1,\gamma_2-I_{\upsilon\eta}\left(I_{\eta\eta}\right)^{-1}\gamma_1\right)$. Calculating that Fisher information matrix:

$$\begin{aligned}\frac{\partial}{\partial\gamma_1}l\left(\gamma_1,\gamma_2-I_{\upsilon\eta}\left(I_{\eta\eta}\right)^{-1}\gamma_1\right) &= \frac{\partial}{\partial\upsilon}l\left(\upsilon,\eta\right)-I_{\upsilon\eta}\left(I_{\eta\eta}\right)^{-1}\frac{\partial}{\partial\eta}l\left(\upsilon,\eta\right)\\ \frac{\partial^2}{\partial\gamma_1\partial\gamma_2}l\left(\gamma_1,\gamma_2-I_{\upsilon\eta}\left(I_{\eta\eta}\right)^{-1}\gamma_1\right) &= \frac{\partial^2}{\partial\upsilon\partial\eta}l\left(\upsilon,\eta\right)-I_{\upsilon\eta}\left(I_{\eta\eta}\right)^{-1}\frac{\partial^2}{\partial\eta\partial\eta}l\left(\upsilon,\eta\right)\Leftrightarrow\\ I_{\gamma_1\gamma_2} &= I_{\upsilon\eta}-I_{\upsilon\eta}\left(I_{\eta\eta}\right)^{-1}I_{\eta\eta}=0\end{aligned}$$ ■

A.7 Proof of Theorem 8

a)

Since $\hat{\upsilon}_{MLE}$ is adaptive, and $I_{\upsilon\upsilon}(\theta)$ does not depend on η, we know that the asymptotic distribution of $\hat{\upsilon}_{MLE}$ does not depend on η. This, in turn, implies that

$$\lim_{n\to\infty}\left\{\ln p\left(\hat{\upsilon};\upsilon,\eta+\frac{\delta_i}{\sqrt{n}}\right)-\ln p\left(\hat{\upsilon};\upsilon,\eta\right)\right\}=0, \tag{A.2}$$

but recalling the definition of local cuts,

$$\ln p\left(\hat{\upsilon};\upsilon,\eta+\frac{\delta_i}{\sqrt{n}}\right)-\ln p\left(\hat{\upsilon};\upsilon,\eta\right)=O\left(n^{-\frac{1}{2}f_m(\eta)}\right). \tag{A.3}$$

Comparing (A.2) and (A.3) it is clear that

$$\lim_{n\to\infty}\left\{n^{-\frac{1}{2}f_m(\eta)}\right\}=0$$

and hence that $f_m(\eta)>0$.

b) Since $\hat{\upsilon}$ is asymptotically normal, we can approximate $\ln p(\hat{\upsilon};\upsilon,\eta)$ using an Edgeworth expansion. But the leading term in this expansion is $N(\upsilon, I_{\upsilon\upsilon}(\theta))$. Since this term does depend on υ, it will be the term which determines the rate. So if we let $h(\hat{\upsilon};\upsilon)$ be the logarithm of this distribution,

$$\ln p\left(\hat{\upsilon};\upsilon+\tfrac{\varepsilon_i}{\sqrt{n}},\eta\right)-\ln p\left(\hat{\upsilon};\upsilon,\eta\right)\approx h\left(\hat{\upsilon};\upsilon+\tfrac{\varepsilon_i}{\sqrt{n}}\right)-h\left(\hat{\upsilon};\upsilon\right)=O\left(n^{-\frac{1}{2}s_m(\upsilon)}\right).$$

It is easy to check, however, that

$$h\left(\hat{\upsilon};\upsilon+\tfrac{\varepsilon_i}{\sqrt{n}}\right)-h\left(\hat{\upsilon};\upsilon\right)=O_P(1)$$

and therefore $s_m(\upsilon)=0$. ■

A.8 Proof of Theorem 9

We will prove the theorem in two parts:

a)$\min\left(s_m(\upsilon), f_c(\upsilon)\right) = 0$

b)$s_c(\eta) = 0$

Furthermore, we know from Theorem 8 that $s_m(\upsilon) = 0$ and that $f_m(\eta) > 0$. Holding these together we obtain:

$$s_m(\upsilon) = s_c(\eta) = 0,\ f_m(\eta) > 0 \text{ and } f_c(\upsilon) \geq 0$$

implying that there is a marginal local cut in the model. Now for the proofs of **a)** and **b)**:

a)

Since $\hat{\upsilon}_{MLE}$ is adaptive, we know that

$$\frac{\partial}{\partial \upsilon} \ln p(x; \upsilon, \eta) = O\left(\sqrt{n}\right). \tag{A.4}$$

We also know that

$$\begin{aligned} & \ln p\left(\mathbf{x}; \upsilon + \tfrac{\varepsilon_i}{\sqrt{n}}, \eta\right) - \ln p(\mathbf{x}; \upsilon, \eta) \qquad \text{(A.5)} \\ = \; & \ln p\left(\mathbf{x}; \upsilon + \tfrac{\varepsilon_i}{\sqrt{n}}, \eta | \hat{\upsilon}_{MLE}\right) - \ln p(\mathbf{x}; \upsilon, \eta | \hat{\upsilon}_{MLE}) \\ & + \ln p\left(\hat{\upsilon}_{MLE}; \upsilon + \tfrac{\varepsilon_i}{\sqrt{n}}, \eta\right) - \ln p(\hat{\upsilon}_{MLE}; \upsilon, \eta) \\ = \; & O\left(n^{-\frac{1}{2} s_m(\upsilon)}\right) + O\left(n^{-\frac{1}{2} f_c(\upsilon)}\right) \end{aligned}$$

and, from the standard Taylor expansion,

$$\begin{aligned} \ln p\left(x; \upsilon + \tfrac{\varepsilon_i}{\sqrt{n}}, \eta\right) - \ln p(x; \upsilon, \eta) &= \frac{\partial}{\partial \upsilon} \ln p(x; \upsilon, \eta) \tfrac{\varepsilon_i}{\sqrt{n}} + o\left(n^{-1}\right) \qquad \text{(A.6)} \\ &= O_P\left(\sqrt{n}\right) \tfrac{\varepsilon_i}{\sqrt{n}} + o\left(n^{-1}\right) = O_P(1). \end{aligned}$$

What remains is to connect this result to the rates. Comparing (A.5) and (A.6), we get the following equality:

$$O_P(1) = O\left(n^{-\frac{1}{2}s_m(\upsilon)}\right) + O\left(n^{-\frac{1}{2}f_c(\upsilon)}\right) \tag{A.7}$$

implying that $\min\left(s_m(\upsilon), f_c(\upsilon)\right) = 0$.

b) Since $\hat{\eta}_{MLE}$ is LGR, we know that

$$\frac{\partial}{\partial\eta}\ln p(x;\upsilon,\eta) = O_P\left(\sqrt{n}\right) \tag{A.8}$$

Using the standard Taylor expansion together with (A.8), we get the rate for a shift in η in the entire likelihood function:

$$\begin{aligned}\ln p\left(x;\upsilon,\eta+\tfrac{\delta_i}{\sqrt{n}}\right) - \ln p(x;\upsilon,\eta) &= \frac{\partial}{\partial\eta}\ln p(x;\upsilon,\eta)\tfrac{\delta_i}{\sqrt{n}} + o\left(n^{-1}\right) \\ &= O_P(1),\end{aligned}$$

but we also know that

$$\begin{aligned}&\ln p\left(\mathbf{x};\upsilon,\eta+\tfrac{\delta_i}{\sqrt{n}}\right) - \ln p(\mathbf{x};\upsilon,\eta) \\ =\ &\ln p\left(\mathbf{x};\upsilon,\eta+\tfrac{\delta_i}{\sqrt{n}}|\hat{\upsilon}_{MLE}\right) - \ln p(\mathbf{x};\upsilon,\eta|\hat{\upsilon}_{MLE}) \\ &+\ln p\left(\hat{\upsilon}_{MLE};\upsilon,\eta+\tfrac{\delta_i}{\sqrt{n}}\right) - \ln p(\hat{\upsilon}_{MLE};\upsilon,\eta) \\ =\ &O\left(n^{-\frac{1}{2}s_c(\eta)}\right) + O\left(n^{-\frac{1}{2}f_m(\eta)}\right)\end{aligned}$$

Holding these together, we obtain

$$O_P(1) = O\left(n^{-\frac{1}{2}s_c(\eta)}\right) + O\left(n^{-\frac{1}{2}f_m(\eta)}\right).$$

From Theorem 8, we know that $f_m(\eta) > 0$, and therefore it must be the case that $s_c(\eta) = 0$. ■

Appendix B

Appendices for Chapter 4

B.1 Proof of Theorem 10

Under the null hypothesis and Assumptions 1 and 2 it follows from (4.2) and (4.5) that

$$\begin{aligned} F^* &= T(R(\hat{\beta}-\beta))'[R\hat{B}R']^{-1}(R(\hat{\beta}-\beta))/q \\ &= (RT^{\frac{1}{2}}(\hat{\beta}-\beta))'[R\hat{B}R']^{-1}(RT^{\frac{1}{2}}(\hat{\beta}-\beta))/q \\ &\Rightarrow (RQ^{-1}\Lambda W_k(1))'[RQ^{-1}\Lambda P_k \Lambda' Q^{-1}R']^{-1}RQ^{-1}\Lambda W_k(1)/q. \end{aligned}$$

Because the matrix $RQ^{-1}\Lambda$ has rank q and $W_k(r)$ is a vector of independent Wiener processes and is Gaussian, $RQ^{-1}\Lambda W_k(r)$ can be written as $\Lambda^* W_q^*(r)$ where $W_q^*(r)$ is a $(q \times 1)$ vector of independent Wiener processes, and Λ^* is the $(q \times q)$ matrix square root of $RQ^{-1}\Lambda\Lambda' Q^{-1}R'$. Λ^* exists and is invertible because $RQ^{-1}\Lambda\Lambda' Q^{-1}R'$ is symmetric and full rank. Therefore, $RQ^{-1}\Lambda W_k(1)$ is equivalent in distribution to $\Lambda^* W_q^*(r)$. In addition $RQ^{-1}\Lambda P_k \Lambda' Q^{-1}R'$ is equivalent in distribution to $\Lambda^* P_q^* \Lambda^{*\prime}$ where $P_q^* = \int_0^1 (W_q^*(r) - rW_q^*(1))(W_q^*(r) - rW_q^*(1))' dr$.

This follows because

$$
\begin{aligned}
& RQ^{-1}\Lambda P_k \Lambda' Q^{-1} R' \\
= \ & RQ^{-1}\Lambda\{\int_0^1 [W_k(r) - rW_k(1)][W_k(r) - rW_k(1)]' dr\}\Lambda' Q^{-1} R' \\
= \ & \int_0^1 (RQ^{-1}\Lambda W_k(r) - rRQ^{-1}\Lambda W_k(1))(RQ^{-1}\Lambda W_k(r) - rRQ^{-1}\Lambda W_k(1))' dr
\end{aligned}
$$

which is equivalent in distribution to

$$\int_0^1 (\Lambda^* W_q^*(r) - r\Lambda^* W_q^*(1))(\Lambda^* W_q^*(r) - r\Lambda^* W_q^*(1))' dr = \Lambda^* P_q^* \Lambda^{*\prime}.$$

Thus, $(RQ^{-1}\Lambda W_k(1))'[RQ^{-1}\Lambda P_k \Lambda' Q^{-1} R']^{-1} RQ^{-1}\Lambda W_k(1)/q$ is equivalent in distribution to

$$(\Lambda^* W_q^*(1))'[\Lambda^* P_q^* \Lambda^{*\prime}]^{-1}\Lambda^* W_q^*(1)/q = W_q^*(1)' P_q^{*-1} W_q^*(1)/q$$

as required.

B.2 Proof that F^* Satisfies the Frisch-Waugh-Lovell Theorem

This proof is simplified by writing the model and F^* in matrix notation. Write regression (4.1) as $Y = X\beta + u$. Let G denote a $(T \times T)$ lower triangular matrix with elements along the diagonal and below all equal to one. Let $\hat{U}$ denote a $(T \times T)$ diagonal matrix with diagonal elements $(\hat{u}_1', \hat{u}_2', \ldots, \hat{u}_T)$. Define $\hat{S} = G\hat{U}X$. Simple algebra gives $\sum_{t=1}^{T} \hat{S}_t \hat{S}_t' = \hat{S}'\hat{S} = X'\hat{U}G'G\hat{U}X = X'HX$ where $H = \hat{U}G'G\hat{U}$ so that $\hat{C} = T^{-2}X'HX$. Without loss of generality, partition X into X_1' and X_2' where X_1' contains the first k_1' columns of X, and X_2' contains the last $k - k_1'$ columns of X. Partition β into β_1' and β_2' where β_1' is a $(k_1' \times 1)$ vector containing the first k_1' elements of β, and β_2' is a $((k - k_1') \times 1)$ vector containing the last $k - k_1'$ elements of β. To show that F^* satisfies the FWL Theorem we must show that F^* for testing $\beta_1' = 0$ in regression (1) is computationally the same as F^* for testing $\beta_1' = 0$ in the regression

$$\tilde{Y} = \tilde{X}_1\beta_1 + \tilde{u}, \tag{B.1}$$

where $\tilde{Y} = M_2'Y$, $\tilde{X}_1' = M_2'X_1'$, $\tilde{u} = M_2'u$ and $M_2' = I_T - X_2'(X_2'X_2')^{-1}X_2'$. Let $\tilde{\beta}_1'$ denote the OLS estimate of β_1' from regression (B.1), and let $R = [I_{k_1'} 0_{k-k_1'}]$. By the FWL Theorem the OLS residuals from regressions (4.1) and (B.1) are the same. Therefore the H matrix in regression (B.1) is the same as in regression (4.1). Thus, the F^* statistics from regressions (4.1) and (B.1) can be written respectively as

$$T(R\hat{\beta})'[R(X'X)^{-1}X'HX(X'X)^{-1}R']^{-1}(R\hat{\beta})/q, \tag{B.2}$$

$$T\tilde{\beta}_1'[(\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'H\tilde{X}_1(\tilde{X}_1'\tilde{X}_1)^{-1}]^{-1}\tilde{\beta}_1/q. \tag{B.3}$$

By the FWL Theorem $R\hat{\beta} = \tilde{\beta}_1'$; therefore, (B.2) and (B.3) are computationally equivalent if $R(X'X)^{-1}X'HX(X'X)^{-1}R' = (\tilde{X}_1'\tilde{X}_1')^{-1}\tilde{X}_1'HX_1'(\tilde{X}_1'\tilde{X}_1')^{-1}$. It is sufficient to show that $R(X'X)^{-1}X' = (\tilde{X}_1'\tilde{X}_1')^{-1}\tilde{X}_1'$. From the partitioned matrix formula it follows that

$$\begin{aligned} R(X'X)^{-1}X' &= [(\tilde{X}_1'\tilde{X}_1)^{-1}, -(\tilde{X}_1'\tilde{X}_1)^{-1}X_1'X_2(X_2X_2)^{-1}] \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} \\ &= (\tilde{X}_1'\tilde{X}_1)^{-1}X_1' - (\tilde{X}_1'\tilde{X}_1)^{-1}X_1'X_2(X_2'X_2)^{-1}X_2' \\ &= (\tilde{X}_1'\tilde{X}_1)^{-1}X_1'(I_T - X_2(X_2'X_2)^{-1}X_2') \\ &= (\tilde{X}_1'\tilde{X}_1)^{-1}X_1'M_2 = (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1' \end{aligned}$$

which completes the proof.

B.3 Proof of (4.11) and (4.12)

The denominators of t_{HAC} and t^* are invariant to c since they are functions of $\hat{u}_t$ which is invariant to c by construction. It directly follows that

$$\text{plim}\,(\hat{\sigma}_x^{-2}\hat{\sigma}^2\hat{\sigma}_x^{-2})^{\frac{1}{2}} = (\sigma_x^{-2}\sigma^2\sigma_x^{-2})^{\frac{1}{2}} = \sigma/\sigma_x^2,$$

and using (5) with simplifying algebra we have

$$(\hat{\sigma}_x^{-2}\hat{C}\hat{\sigma}_x^{-2})^{\frac{1}{2}} \Rightarrow (\sigma/\sigma_x^2)[\int_0^1 (W_1'(r) - rW_1(1))^2 dr]^{\frac{1}{2}}.$$

Given these results, all that is needed to prove (4.11) and (4.12) is the limit of $T^{\frac{1}{2}}(\hat{\beta} - \beta_0)$ under the local alternative. Using Assumptions 1 and 2 we have

$$T^{\frac{1}{2}}(\hat{\beta} - \beta_0) = c + (T^{-1}\sum_{t=1}^{T} x_t^2)^{-1} T^{-\frac{1}{2}} \sum_{t=1}^{T} x_t u_t \Rightarrow c + \sigma W_1(1)/\sigma_x^2.$$

Simple algebra completes the proof.

Appendix C

Appendices for Chapter 5

C.1 Proof of Lemma 1

First we will prove a small proposition.

Proposition 1 $T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\left[\tilde{f}_t(\beta_0)-\tilde{f}_t\left(\hat{\beta}\right)\right]=O_P\left(T^{-\frac{1}{2}}\right)$

C.1.1 Proof of Proposition 1

First we will make use of the following Taylor expansion of $\tilde{f}_t\left(\hat{\beta}\right)$ around β_0 :

$$\begin{aligned}\tilde{f}_t\left(\hat{\beta}\right) &= \tilde{f}_t(\beta_0)+\tilde{F}_t(\beta_0)\left(\hat{\beta}-\beta_0\right)+r_T \Leftrightarrow\\ \tilde{f}_t(\beta_0)-\tilde{f}_t\left(\hat{\beta}\right) &= -\tilde{F}_t'(\beta_0)\left(\hat{\beta}-\beta_0\right)+r_T\end{aligned}$$

where, since the third derivative of $\tilde{f}_t$ is bounded by an integrable function, r_T is of order T^{-1}. This implies that

$$\begin{aligned}&T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\left[\tilde{f}_t(\beta_0)-\tilde{f}_t\left(\hat{\beta}\right)\right]\\ =\; &-T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\left[\tilde{F}_t'(\beta_0)\left(\hat{\beta}-\beta_0\right)+r_T\right]\\ =\; &-T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\,\tilde{F}_t'(\beta_0)\left(\hat{\beta}-\beta_0\right)\\ &+r_T T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\end{aligned}$$

Since $T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\tilde{F}_t'(\beta_0)$ is of order 1 and $\left(\hat{\beta}-\beta_0\right)$ is of order $T^{-\frac{1}{2}}$, the first term is of order $T^{-\frac{1}{2}}$. Furthermore, since r_T is of order T^{-1} while $T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)$ is of order 1, the second term is of order T^{-1}. Collecting these terms, we obtain the desired result:

$$T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\left[\tilde{f}_t(\beta_0)-\tilde{f}_t\left(\hat{\beta}\right)\right]=O_P\left(T^{-\frac{1}{2}}\right).\blacksquare$$

C.2 Proof of Lemma 3

First we expand $\tilde{F}_t\left(\hat{\beta}\right)$ around β_0 in the following manner

$$\tilde{F}_t\left(\hat{\beta}\right)=\tilde{F}_t(\beta_0)+\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\left(\hat{\beta}-\beta_0\right)+s_T$$

Where s_T consists of all the higher order terms of the expansion. Now, since the third derivative of $\tilde{f}_t$ is bounded and $\hat{\beta}$ is $\sqrt{T}$-consistent, s_T is of order T^{-1}. Making use of this, we obtain

$$\begin{aligned}T^{-\frac{1}{2}}\hat{S}_{[rT]} &= T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\hat{v}_t=T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\tilde{F}_t\left(\hat{\beta}\right)\hat{u}_t \qquad (C.1)\\ &= T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\tilde{F}_t(\beta_0)\hat{u}_t+T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\left(\hat{\beta}-\beta_0\right)\hat{u}_t\\ &\quad+\left(T^{\frac{1}{2}}s_T\right)T^{-1}\sum_{t=1}^{[rT]}\hat{u}\end{aligned}$$

From (C.1), we have

$$l_{[rT]}=T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\left(\hat{\beta}-\beta_0\right)\hat{u}_t+\left(T^{\frac{1}{2}}s_T\right)T^{-1}\sum_{t=1}^{[rT]}\hat{u}_t.$$

Examining the first term of $l_{[rT]}$, we see that

$$\begin{aligned}
& T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\left(\hat{\beta}-\beta_0\right)\hat{u}_t \\
= \; & T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\hat{u}_t\left[T^{\frac{1}{2}}\left(\hat{\beta}-\beta_0\right)\right] \\
= \; & T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\left[\tilde{f}_t(\beta_0)+\tilde{u}_t-\tilde{f}_t\left(\hat{\beta}\right)\right]\left[T^{\frac{1}{2}}\left(\hat{\beta}-\beta_0\right)\right] \\
= \; & T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\left[\tilde{f}_t(\beta_0)-\tilde{f}_t\left(\hat{\beta}\right)\right]\left[T^{\frac{1}{2}}\left(\hat{\beta}-\beta_0\right)\right] \\
& -T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\tilde{u}_t\left[T^{\frac{1}{2}}\left(\hat{\beta}-\beta_0\right)\right] \\
= \; & T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\tilde{u}_t\left[T^{\frac{1}{2}}\left(\hat{\beta}-\beta_0\right)\right]+o_P(1),
\end{aligned} \tag{C.2}$$

Where the last equality follows from Proposition (1). Since moment and mixing conditions have been placed on $\tilde{u}_t$, we can use a law of large numbers to ensure that

$$\text{plim}\left(T^{-1}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\tilde{u}_t\right)=0.$$

Therefore

$$T^{-\frac{1}{2}}\sum_{t=1}^{[rT]}\frac{\partial}{\partial\beta'}\tilde{F}_t(\beta_0)\left(\hat{\beta}-\beta_0\right)\hat{u}_t=o_P(1). \tag{C.3}$$

Now consider the second term of $l_{[rT]}$. Using arguments similar to the ones used above, we obtain

$$\begin{aligned}
T^{-1}\sum_{t=1}^{[rT]}\hat{u}_t &= T^{-1}\sum_{t=1}^{[rT]}\left[\tilde{f}_t(\beta_0)-\tilde{f}_t\left(\hat{\beta}\right)\right]+T^{-1}\sum_{t=1}^{[rT]}\tilde{u}_t \\
&= -T^{-1}\sum_{t=1}^{[rT]}\tilde{F}_t(\beta_0)\left(\hat{\beta}-\beta_0\right)+T^{-1}\sum_{t=1}^{[rT]}\tilde{u}_t+O_P\left(T^{-1}\right),
\end{aligned}$$

and by applying a law of large number to $\tilde{u}_t$, we establish that

$$\text{plim}\left(T^{-1}\sum_{t=1}^{[rT]}\tilde{u}_t\right)=0$$

and therefore, recalling that s_T is of order T^{-1},

$$\text{plim}\left[\left(T^{\frac{1}{2}}s_T\right)T^{-1}\sum_{t=1}^{[rT]}\hat{u}_t\right]=\text{plim}\left(T^{\frac{1}{2}}s_T\right)\text{plim}\left(T^{-1}\sum_{t=1}^{[rT]}\hat{u}_t\right)=0 \qquad \text{(C.4)}$$

From (C.3) and (C.4) it immediately follows that $\text{plim}(l_T)=0$. ■

C.3 Proof of Theorem 11

We wish to find the asymptotic distribution of:

$$F^*=T\,r\left(\hat{\beta}\right)'\left[\hat{R}\hat{B}\hat{R}'\right]^{-1}r\left(\hat{\beta}\right)/q.$$

First note that

$$\begin{aligned} r\left(\beta_0\right) &= r\left(\hat{\beta}\right)+\hat{R}\left(\beta_0-\hat{\beta}\right)+O_P\left(T^{-1}\right)\Leftrightarrow \\ r\left(\hat{\beta}\right) &= r\left(\beta_0\right)+\hat{R}\left(\hat{\beta}-\beta_0\right)+O_P\left(T^{-1}\right) \end{aligned}$$

So under H_0

$$r\left(\hat{\beta}\right)=\hat{R}\left(\hat{\beta}-\beta_0\right)+O_P\left(T^{-1}\right)$$

implying that

$$\begin{aligned} F^* &= T\,r\left(\hat{\beta}\right)'\left[\hat{R}\hat{B}\hat{R}'\right]^{-1}r\left(\hat{\beta}\right)/q \\ &= T\,\left[\hat{R}\left(\hat{\beta}-\beta_0\right)\right]'\left[\hat{R}\hat{B}\hat{R}'\right]^{-1}\left[\hat{R}\left(\hat{\beta}-\beta_0\right)\right]/q+O_P\left(T^{-1}\right) \\ &\Rightarrow \left[R_0Q^{-1}\Lambda W_k(1)\right]'\left[R_0Q^{-1}\Lambda P_k\Lambda'Q^{-1}R_0'\right]^{-1}\left[R_0Q^{-1}\Lambda W_k(1)\right]/q \end{aligned}$$

But since $R_0Q^{-1}\Lambda$ has rank q and $W_k(1)$ is a vector of independent Wiener processes that are Gaussian, we can rewrite $R_0Q^{-1}\Lambda W_k(1)$ as $\Lambda^*W_q^*(1)$, where

$W_q^*(1)$ is a $q-$dimensional vector of independent Wiener processes, and Λ^* is the $q \times q$ matrix square root of $R_0 Q^{-1} \Lambda \Lambda' Q^{-1} R_0'$. Note that this square root exist because $R_0 Q^{-1} \Lambda \Lambda' Q^{-1} R_0'$ is a full rank $q \times q$ matrix. Using the same arguments, note that

$$
\begin{aligned}
& R_0 Q^{-1} \Lambda P_k \Lambda' Q^{-1} R_0' \\
= \; & R_0 Q^{-1} \Lambda \int_0^1 \left(W_k(r) - r W_k(1)\right)\left(W_k(r) - r W_k(1)\right)' dr \Lambda' Q^{-1} R_0' \\
= \; & \int_0^1 \left(R_0 Q^{-1} \Lambda W_k(r) - r R_0 Q^{-1} \Lambda W_k(1)\right) \cdot \\
& \left(R_0 Q^{-1} \Lambda W_k(r) - r R_0 Q^{-1} \Lambda W_k(1)\right)' dr \\
= \; & \int_0^1 \left(\Lambda^* W_q^*(1) - r \Lambda^* W_q^*(1)\right)\left(\Lambda^* W_q^*(1) - r \Lambda^* W_q^*(1)\right)' dr \\
= \; & \Lambda^* \int_0^1 \left(W_q^*(1) - r W_q^*(1)\right)\left(W_q^*(1) - r W_q^*(1)\right)' dr \, (\Lambda^*)' \\
= \; & \Lambda^* P_q^* (\Lambda^*)'
\end{aligned}
$$

implying that

$$
\begin{aligned}
F^* & \Rightarrow \left[\Lambda^* W_q^*(1)\right]' \left[\Lambda^* P_q^* (\Lambda^*)'\right]^{-1} \left[\Lambda^* W_q^*(1)\right]/q \\
& = W_q^*(1)' \left(P_q^*\right)^{-1} W_q^*(1)/q
\end{aligned}
$$

■

Appendix D

Appendices for Chapter 6

D.1 Proof of Lemma 2

To ease the notation, let M_f be the matrix projecting onto the space orthogonal to the space spanned by $\mathbf{f}(T)$, defined by $M_f = \left(I - \mathbf{f}(T)\left(\mathbf{f}(T)'\mathbf{f}(T)\right)^{-1}\mathbf{f}(T)'\right)$, and let the matrix projecting onto the space orthogonal to the space spanned by X be denoted by $M_X = I - X(X'X)^{-1}X'$. Straightforward matrix manipulation yields

$$D_T T^{\frac{1}{2}}\left(\hat{\theta} - \theta\right) = \begin{bmatrix} \left(T^{-1}\tau_T \mathbf{f}(T)' M_X \mathbf{f}(T)\tau_T\right)^{-1}\left(T^{-\frac{1}{2}}\tau_T \mathbf{f}(T)' M_X u\right) \\ \left(T^{-2}X'M_f X\right)^{-1}\left(T^{-1}X'M_f u\right) \end{bmatrix}. \quad \text{(A1)}$$

Central limit theorems are now applied to the individual terms:

$$T^{-1}\tau_T \mathbf{f}(T)' M_X \mathbf{f}(T)\tau_T \Rightarrow \int_0^1 F^X(s)\, F^X(s)'\, ds$$

$$T^{-\frac{1}{2}}\tau_T \mathbf{f}(T)' M_X u \Rightarrow \sigma \int_0^1 F^X(s)\, dw_1(s)$$

$$T^{-2}X'M_f X \Rightarrow \Lambda \int_0^1 w_k^F(s)\, w_k^F(s)'\, ds\Lambda'$$

$$T^{-1}X'M_f u \Rightarrow \sigma\Lambda \int_0^1 w_k^F(s)\, dw_1(s).$$

Together with (A1), this implies that

$$D_T T^{\frac{1}{2}}\left(\hat{\theta}-\theta\right) \Rightarrow \begin{bmatrix} \left(\int_0^1 F^X(s) F^X(s)' ds\right)^{-1}\left(\sigma \int_0^1 F^X(s)\, dw_1(s)\right) \\ \left(\Lambda \int_0^1 w_k^F(s)\, w_k^F(s)' ds \Lambda'\right)^{-1}\left(\sigma\Lambda \int_0^1 w_k^F(s)\, dw_1(s)\right) \end{bmatrix}$$

$$= \sigma\left(\Sigma'\right)^{-1}\begin{bmatrix} \left(\int_0^1 F^X(s) F^X(s)' ds\right)^{-1}\int_0^1 F^X(s)\, dw_1(s) \\ \left(\int_0^1 w_k^F(s)\, w_k^F(s)' ds\right)^{-1}\int_0^1 w_k^F(s)\, dw_1(s) \end{bmatrix}. \quad \blacksquare$$

D.2 Proof of Lemma 3

Simple matrix manipulations yield:

$$T^{-\frac{1}{2}} D_T^{-1} \hat{S}_{[rT]} = \begin{bmatrix} T^{-\frac{1}{2}} \sum_{t=1}^{[rT]} \tau_T f(t)\, u_t \\ T^{-1} \sum_{t=1}^{[rT]} X_t u_t \end{bmatrix}$$

$$- \begin{bmatrix} T^{-1} \sum_{t=1}^{[rT]} \tau_T f(t)\, f(t)' \tau_T & T^{-\frac{3}{2}} \sum_{t=1}^{[rT]} \tau_T f(t)\, X_t' \\ T^{-\frac{3}{2}} \sum_{t=1}^{[rT]} X_t f(t)' \tau_T & T^{-2} \sum_{t=1}^{[rT]} X_t X_t' \end{bmatrix} \times \left(T^{\frac{1}{2}} D_T \left(\hat{\theta}-\theta\right)\right).$$

Central limit theorems can now be applied to show that

$$
\begin{aligned}
& T^{-\frac{1}{2}} D_T^{-1} \hat{S}_{[rT]} \\
\Rightarrow & \begin{bmatrix} \sigma \int_0^r F(s)\, dw_1(s) \\ \sigma \Lambda \int_0^r w_k(s)\, dw_1(s) \end{bmatrix} \\
& - \begin{bmatrix} \int_0^r F(s) F(s)'\, ds & \int_0^r F(s) w_k(s)'\, ds \Lambda' \\ \Lambda \int_0^r w_k(s) F(s)'\, ds & \Lambda \int_0^r w_k(s) w_k(s)'\, ds \Lambda' \end{bmatrix} \times \\
& \sigma \begin{bmatrix} I_{k_1} & 0 \\ 0 & (\Lambda')^{-1} \end{bmatrix} \begin{bmatrix} \left(\int_0^1 F^X(s) F^X(s)'\, ds \right)^{-1} \int_0^1 F^X(s)\, dw_1(s) \\ \left(\int_0^1 w_k^F(s) w_k^F(s)'\, ds \right)^{-1} \int_0^1 w_k^F(s)\, dw_1(s) \end{bmatrix} \\
= & \ \sigma \Sigma Q_k^F(r),
\end{aligned}
$$

where

$$
\begin{aligned}
Q_k^F(r) = & \begin{bmatrix} \int_0^r F(t)\, dw_1(s) \\ \int_0^r w_k(s)\, dw_1(s) \end{bmatrix} - \begin{bmatrix} \int_0^r F(s) F(s)'\, ds & \int_0^r F(s) w_k(s)'\, ds \\ \int_0^r w_k(s) F(s)'\, ds & \int_0^r w_k(s) w_k(s)'\, ds \end{bmatrix} \times \\
& \begin{bmatrix} \left(\int_0^1 F^X(s) F^X(s)'\, ds \right)^{-1} \int_0^1 F^X(s)\, dw_1(s) \\ \left(\int_0^1 w_k^F(s) w_k^F(s)'\, ds \right)^{-1} \int_0^1 w_k^F(s)\, dw_1(s) \end{bmatrix}. \qquad \blacksquare
\end{aligned}
$$

D.3 Proof of Lemma 4

Since

$$
\hat{M} = \left(\frac{1}{T} \begin{bmatrix} \mathbf{f}(T)' \mathbf{f}(T) & \mathbf{f}(T)' X \\ X' \mathbf{f}(T) & X'X \end{bmatrix} \right)^{-1} \hat{C}^{\frac{1}{2}},
$$

it is clear that

$$
\hat{M}^{-1} = \hat{C}^{-\frac{1}{2}} \left(\frac{1}{T} \begin{bmatrix} \mathbf{f}(T)' \mathbf{f}(T) & \mathbf{f}(T)' X \\ X' \mathbf{f}(T) & X'X \end{bmatrix} \right)
$$

and

$$D_T^{-1}\hat{C}D_T^{-1} \Rightarrow \sigma^2\Sigma P_k^F \Sigma'.$$

Note that P_k^F is positive definite by construction. This permits use of the Cholesky decomposition to write $P_k^F = Z_k^F \left(Z_k^F\right)'$. Since $\hat{C}^{\frac{1}{2}}$ is the Cholesky decomposition of $\hat{C}$,

$$\begin{aligned} D_T^{-1}\hat{C}^{\frac{1}{2}} &\Rightarrow (\sigma\Sigma) Z_k^F \Leftrightarrow \\ \hat{C}^{-\frac{1}{2}} D_T &\Rightarrow \sigma^{-1}\left(Z_k^F\right)^{-1}\Sigma^{-1}. \end{aligned} \tag{C1}$$

(C1) implies that

$$\begin{aligned} \hat{M}^{-1}D_T^{-1} &= \hat{C}^{-\frac{1}{2}} D_T \left(\frac{1}{T} D_T^{-1} \begin{bmatrix} \mathbf{f}(T)'\mathbf{f}(T) & \mathbf{f}(T)'X \\ X'\mathbf{f}(T) & X'X \end{bmatrix} D_T^{-1} \right) \\ &\Rightarrow \sigma^{-1}\left(Z_k^F\right)^{-1}\Sigma^{-1}\Sigma \begin{bmatrix} \int_0^1 F(s)F(s)'\,ds & \int_0^1 F(s)w_k(s)'\,ds \\ \int_0^1 w_k(s)F(s)'\,ds & \int_0^1 w_k(s)w_k(s)'\,ds \end{bmatrix} \Sigma' \\ &= \sigma^{-1}\left(Z_k^F\right)^{-1} \begin{bmatrix} \int_0^1 F(s)F(s)'\,ds & \int_0^1 F(s)w_k(s)'\,ds \\ \int_0^1 w_k(s)F(s)'\,ds & \int_0^1 w_k(s)w_k(s)'\,ds \end{bmatrix} \Sigma'. \end{aligned}$$

The distribution of the transformed parameters can then be obtained

$$\hat{M}^{-1}T^{\frac{1}{2}}\left(\hat{\theta}-\theta\right) = \left(\hat{M}^{-1}D_T^{-1}\right)\left(D_T T^{\frac{1}{2}}\left(\theta-\hat{\theta}\right)\right)$$

$$\Rightarrow \sigma^{-1}\left(Z_k^F\right)^{-1}\begin{bmatrix} \int_0^1 F(s)\,F(s)'\,ds & \int_0^1 F(s)\,w_k(s)'\,ds \\ \int_0^1 w_k(s)\,F(s)'\,ds & \int_0^1 w_k(s)\,w_k(s)'\,ds \end{bmatrix}\Sigma' \times$$

$$\left(\sigma\left(\Sigma'\right)^{-1}\right)\begin{bmatrix} \left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 F^X(s)\,dw_1(s) \\ \left(\int_0^1 w_k^F(s)\,w_k^F(s)'\,ds\right)^{-1}\int_0^1 w_k^F(s)\,dw_1(s) \end{bmatrix}$$

$$= \left(Z_k^F\right)^{-1}\begin{bmatrix} \int_0^1 F(s)\,F(s)'\,ds & \int_0^1 F(s)\,w_k(s)'\,ds \\ \int_0^1 w_k(s)\,F(s)'\,ds & \int_0^1 w_k(s)\,w_k(s)'\,ds \end{bmatrix} \times$$

$$\begin{bmatrix} \left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 F^X(s)\,dw_1(s) \\ \left(\int_0^1 w_k^F(s)\,w_k^F(s)'\,ds\right)^{-1}\int_0^1 w_k^F(s)\,dw_1(s) \end{bmatrix}.$$

It is then clear that the asymptotic distribution does not depend on nuisance parameters. ■

D.4 Proof of Theorem 12

Lemma D1 below will be used extensively in the proof of Theorem 6.3. Before stating it, some additional notation is required. Consider the model

$$y = X\beta + u,$$

where the hypothesis of interest is $R\beta = r$. Then let $L' = \begin{bmatrix} R' & D' \end{bmatrix}$, where D is chosen such that L has full rank. The model can then be rewritten as

$$\begin{aligned} y &= \left(XL^{-1}\right)(L\beta) + u \\ &= \begin{bmatrix}\tilde{X}_1 & \tilde{X}_2\end{bmatrix} \begin{bmatrix} \beta_1^* \\ \beta_2^* \end{bmatrix} + u \\ &= \tilde{X}_1\beta_1^* + \tilde{X}_2\beta_2^* + u \end{aligned}$$

Since $\beta_1^* = R\beta$, the hypothesis of interest is now $\tilde{H}_0 : \beta_1^* = r$. The model can now be transformed such that all parameters other than β_1^* are eliminated. For this purpose, use $M_{\tilde{X}_2} = I - \tilde{X}_2\left(\tilde{X}_2'\tilde{X}_2\right)^{-1}\tilde{X}_2'$ projecting off the space spanned by $\tilde{X}_2$, and rewrite the model to its final form:

$$\begin{aligned} M_{\tilde{X}_2}y &= M_{\tilde{X}_2}\tilde{X}_1\beta_1^* + M_{\tilde{X}_2}\tilde{X}_2\beta_2^* + M_{\tilde{X}_2}u \Leftrightarrow \\ y^* &= X_1^*\beta_1^* + u^*. \end{aligned}$$

Lemma 6 *F^* for testing $R\beta = r$ in the model*

$$y = X\beta + u$$

is computationally equivalent to testing $\beta_1^* = r$ in the model

$$y^* = X_1^*\beta_1^* + u^*.$$

The proof of this lemma can be found in Chapter 2. Lemma 6 is now utilized to prove Theorem 6.3.

D.4.1 Proof of (a)

Let

$$L = \begin{bmatrix} R^X \\ D \end{bmatrix},$$

where D is chosen such that L has full rank (k_1), and define

$$\begin{aligned}
\left[\tilde{X}_1\ \tilde{X}_2\right] &= XL^{-1} \\
X_1^* &= M_f M_{\tilde{X}_2} \tilde{X}_1, \\
u^* &= M_f M_{\tilde{X}_2} u, \text{ and} \\
y^* &= M_f M_{\tilde{X}_2} y.
\end{aligned}$$

Using these definitions, Model (6.1) can be rewritten in the following manner:

$$\begin{aligned}
y &= f(T)\alpha + \left(XL^{-1}\right)(L\beta) + u \\
&= f(T)\alpha + \left[\tilde{X}_1\ \tilde{X}_2\right] \begin{bmatrix} \beta_1^* \\ \beta_2^* \end{bmatrix} + u \\
&= f(T)\alpha + \tilde{X}_1\beta_1^* + \tilde{X}_2\beta_2^* + u
\end{aligned}$$

Since $\tilde{X}_1$ and $\tilde{X}_2$ are linear combinations of X, they too contain unit root processes as long as the original assumption of just one cointegration relationship is maintained; moreover, they will still be uncorrelated with u. This implies that Assumptions 5 and 6 hold for the reparametrized model. Then by Lemma D1, it is known that the hypothesis of interest, $H_0^X : R^X\beta = r$, is equivalent to testing the hypothesis $\tilde{H}_0^X : \beta_1^* = r$ in the model

$$y^* = X_1^*\beta_1^* + u^*. \tag{D1.1}$$

All that remains is then to derive the asymptotic distribution of

$$F^* = T\left(\beta_1^* - r\right)'\left[B^*\right]^{-1}\left(\beta_1^* - r\right),$$

where B^* is defined as $\hat{B}$, but for (D1.1). Now note that

$$\begin{aligned}
T\left(\hat{\beta}_1^* - \beta_1^*\right) &= \left[T^{-2}\left(X_1^*\right)' X_1^*\right]^{-1} T^{-1}\left(X_1^*\right)' u^* \\
&\Rightarrow \sigma\left(\Lambda^{*'}\right)^{-1}\left(\int_0^1 \hat{w}_q^F(s)\,\hat{w}_q^F(s)'\,ds\right)^{-1}\int_0^1 \hat{w}_q^F(s)\,dw_1(s)
\end{aligned}$$

where $\Lambda^* (\Lambda^*)'$ is 2π times the spectral density of X_{1t}^* at frequency 0. Also, defining $\hat{S}_{1t}$ in the natural way,

$$\begin{aligned}
T^{-1}\hat{S}_{1[rT]}^* &= \sum_{t=1}^{[rT]} X_{1t}^* \hat{u}_t^* = \sum_{t=1}^{[rT]} X_{1t}^* \left(u_t^* - X_{1t}^* \left(\hat{\beta}_1^* - \beta_1^* \right) \right) \\
&= \sum_{t=1}^{[rT]} X_{1t}^* u_t^* - \sum_{t=1}^{[rT]} X_{1t}^* X_{1t}^* \left(\hat{\beta}_1^* - \beta_1^* \right) \\
&\Rightarrow \sigma \Lambda^* \int_0^r \hat{w}_q^F(s)\, dw_1(s) \\
&\quad - \sigma \Lambda^* \left(\int_0^r \hat{w}_q^F(s)\, \hat{w}_q^F(s)'\, ds \right) \left(\int_0^1 \hat{w}_q^F(s)'\, \hat{w}_q^F(s)\, ds \right)^{-1} \\
&\quad \int_0^1 \hat{w}_q^F(s)\, dw_1(s) \\
&= \sigma \Lambda^* V(r).
\end{aligned}$$

Since $\hat{C}_1^* = T^{-2} \sum_{t=1}^{T} \hat{S}_{1t}^* \hat{S}_{1t}^{*\prime}$,

$$T^{-1}\hat{C}_1^* \Rightarrow \sigma^2 \Lambda^* \int_0^1 V(s)' V(s)\, ds\, (\Lambda^*)'. \tag{D1.2}$$

Using (D1.2), the asymptotic distribution of B^* can be found:

$$
\begin{aligned}
T^{-1}B^* &= \left[T^{-2}\left(X_1^*\right)' X_1^*\right]^{-1} T^{-1}\hat{C}_1^* \left[T^{-2}\left(X_1^*\right)' X_1^*\right]^{-1} \\
&\Rightarrow \left(\Lambda^* \int_0^1 \hat{w}_q^F(s)\, \hat{w}_q^F(s)'\, ds\, (\Lambda^*)'\right)^{-1} \sigma^2 \Lambda^* \int_0^1 V(s)' V(s)\, ds\, (\Lambda^*)' \times \\
&\quad \left(\Lambda^* \int_0^1 \hat{w}_q^F(s)\, \hat{w}_q^F(s)'\, ds\, (\Lambda^*)'\right)^{-1} \\
&= \sigma^2 \left(\Lambda^{*\prime}\right)^{-1} \left(\int_0^1 \hat{w}_q^F(s)\, \hat{w}_q^F(s)'\, ds\right)^{-1} \int_0^1 V(s)' V(s)\, ds \times \\
&\quad \left(\int_0^1 \hat{w}_q^F(s)\, \hat{w}_q^F(s)'\, ds\right)^{-1} (\Lambda^*)^{-1}.
\end{aligned}
$$

The distribution of F^* can now be obtained:

$$
\begin{aligned}
F^* &= T\left(\hat{\beta}_1^* - r\right)' [B^*]^{-1} \left(\hat{\beta}_1^* - r\right) = T\left(\hat{\beta}_1^* - \beta_1^*\right)' \left[T^{-1}B^*\right]^{-1} T\left(\hat{\beta}_1^* - \beta_1^*\right) \\
&\Rightarrow \left[\left(\Lambda^{*\prime}\right)^{-1} \left(\int_0^1 \hat{w}_q^F(s)\, \hat{w}_q^F(s)'\, ds\right)^{-1} \int_0^1 \hat{w}_q^F(s)\, dw_1(s)\right]' \\
&\quad \left[\left(\Lambda^{*\prime}\right)^{-1} \left(\int_0^1 \hat{w}_q^F(s)\, \hat{w}_q^F(s)'\, ds\right)^{-1} \int_0^1 V(s)' V(s)\, ds \right. \\
&\quad \left. \left(\int_0^1 \hat{w}_q^F(s)\, \hat{w}_q^F(s)'\, ds\right)^{-1} (\Lambda^*)^{-1}\right]^{-1} \\
&\quad \left[\left(\Lambda^{*\prime}\right)^{-1} \left(\int_0^1 \hat{w}_q^F(s)\, \hat{w}_q^F(s)'\, ds\right)^{-1} \int_0^1 \hat{w}_q^F(s)\, dw_1(s)\right] \\
&= \left[\int_0^1 \hat{w}_q^F(s)\, dw_1(s)\right]' \left[\int_0^1 V(s)' V(s)\, ds\right]^{-1} \left[\int_0^1 \hat{w}_q^F(s)\, dw_1(s)\right]. \qquad \blacksquare
\end{aligned}
$$

D.4.2 Proof of (b)

The hypothesis being tested is $H_0^F : R^F\alpha = r$. Again, the model is transformed and the distribution of F^* can be obtained from a simplified expression.

$$\begin{aligned} y &= \mathbf{f}(T)\alpha + X\beta + u \Leftrightarrow \\ M_X y &= M_X \mathbf{f}(T)\alpha + M_X u \Leftrightarrow \\ \bar{y} &= \mathbf{f}^X(T)\alpha + \bar{u}, \end{aligned} \tag{D2.1}$$

or equivalently

$$\bar{y}_t = f^X(t)\alpha + \bar{u}_t, \; t = 1, ..., T.$$

Using the result of Lemma D1, what remains is to derive the asymptotic distribution of

$$F_1^* = T\left(R^F\hat{\alpha} - r\right)'\left[R^F\hat{B}^X R^{F\prime}\right]^{-1}\left(R^F\hat{\alpha} - r\right)/q.$$

where $\hat{B}^X$ is calculated from the transformed model in (D2.1). Now,

$$\tau_T^{-1}T^{\frac{1}{2}}(\hat{\alpha} - \alpha) \Rightarrow \sigma\left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 F^X(s)\,dw_1(s)$$

and $\hat{S}$ for the transformed model, denoted by $\hat{S}^X$, will be:

$$\begin{aligned}
T^{-\frac{1}{2}}\tau_T\hat{S}^X_{[rT]} &= T^{-\frac{1}{2}}\tau_T\sum_{t=1}^{[rT]} f^X(t)'\widehat{u}_t = T^{-\frac{1}{2}}\tau_T\sum_{t=1}^{[rT]} f^X(t)'\left(\bar{u}_t - f^X(t)(\hat{\alpha}-\alpha)\right) \\
&= T^{-\frac{1}{2}}\tau_T\sum_{t=1}^{[rT]} f^X(t)'\bar{u}_t - T^{-\frac{1}{2}}\tau_T\sum_{t=1}^{[rT]} f^X(t)' f^X(t)(\hat{\alpha}-\alpha) \\
&= T^{-\frac{1}{2}}\tau_T\sum_{t=1}^{[rT]} f^X(t)'\bar{u}_t - \left(T^{-1}\tau_T\sum_{t=1}^{[rT]} f^X(t)' f^X(t)\tau_T\right) \\
&\quad \left(\tau_T^{-1}T^{\frac{1}{2}}(\hat{\alpha}-\alpha)\right) \\
&\Rightarrow \sigma\int_0^r F^X(s)\,dw_1(s) - \sigma\left(\int_0^r F^X(s)F^X(s)'\,ds\right)\cdot \\
&\quad \left(\int_0^1 F^X(s)F^X(s)'\,ds\right)^{-1}\int_0^1 F^X(s)\,dw_1(s) \\
&= \sigma V^F(r).
\end{aligned}$$

$\hat{C}$ for the transformed model will then be

$$\begin{aligned}
\tau_T\hat{C}^X\tau_T &= T^{-1}\sum_{t=1}^{T}\left(T^{-\frac{1}{2}}\tau_T\hat{S}_t^X\right)\left(T^{-\frac{1}{2}}\tau_T\hat{S}_t^X\right)' \\
&\Rightarrow \sigma^2\int_0^1 V^F(r)\left(V^F(r)\right)' dr,
\end{aligned}$$

and $\hat{B}$ for the transformed model can be expressed as

$$\begin{aligned}
\tau_T^{-1}\hat{B}^X\tau_T^{-1} &= \left(T^{-1}\tau_T\mathbf{f}^X(T)'\mathbf{f}^X(T)\tau_T\right)^{-1}\tau_T\hat{C}^X\tau_T\left(T^{-1}\tau_T\mathbf{f}^X(T)'\mathbf{f}^X(T)\tau_T\right)^{-1} \\
&\Rightarrow \sigma^2\left(\int_0^1 F^X(s)F^X(s)'\,ds\right)^{-1}\int_0^1 V^F(r)\left(V^F(r)\right)' dr \\
&\quad \left(\int_0^1 F^X(s)F^X(s)'\,ds\right)^{-1},
\end{aligned}$$

implying that under the null hypothesis

$$
\begin{aligned}
F^{*} &= T\left(R^{F}\left(\hat{\alpha}-\alpha\right)\right)'\left[R^{F}\hat{B}^{X}R^{F\prime}\right]^{-1}\left(R^{F}\left(\hat{\alpha}-\alpha\right)\right)/q \\
&= T\left(\left(A^{-1}R^{F}\tau_{T}\right)\tau_{T}^{-1}\left(\hat{\alpha}-\alpha\right)\right)'\left[\left(A^{-1}R^{F}\tau_{T}\right)\tau_{T}^{-1}\hat{B}^{X}\tau_{T}^{-1}\left(\tau_{T}R^{F\prime}A^{-1}\right)\right]^{-1}\times \\
&\quad \left(\left(A^{-1}R^{F}\tau_{T}\right)\tau_{T}^{-1}\left(\hat{\alpha}-\alpha\right)\right)/q \\
&= \left(\left(A^{-1}R^{F}\tau_{T}\right)T^{\frac{1}{2}}\tau_{T}^{-1}\left(\hat{\alpha}-\alpha\right)\right)'\left[\left(A^{-1}R^{F}\tau_{T}\right)\tau_{T}^{-1}\hat{B}^{X}\tau_{T}^{-1}\left(\tau_{T}R^{F\prime}A^{-1}\right)\right]^{-1}\times \\
&\quad \left(\left(A^{-1}R^{F}\tau_{T}\right)T^{\frac{1}{2}}\tau_{T}^{-1}\left(\hat{\alpha}-\alpha\right)\right)/q \\
&\Rightarrow \left(R^{*}\sigma\left(\int_{0}^{1}F^{X}(s)\,F^{X}(s)'\,ds\right)^{-1}\int_{0}^{1}F^{X}(s)\,dw_{1}(s)\right)' \\
&\quad \left[R^{*}\sigma^{2}\left(\int_{0}^{1}F^{X}(s)\,F^{X}(s)'\,ds\right)^{-1}\int_{0}^{1}V^{F}(r)\left(V^{F}(r)\right)'dr\right. \\
&\quad \left.\left(\int_{0}^{1}F^{X}(s)\,F^{X}(s)'\,ds\right)^{-1}\left(R^{*}\right)'\right]^{-1} \\
&\quad \left(R^{*}\sigma\left(\int_{0}^{1}F^{X}(s)\,F^{X}(s)'\,ds\right)^{-1}\int_{0}^{1}F^{X}(s)\,dw_{1}(s)\right)
\end{aligned}
$$

$$= \left(R^{*}\left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 F^X(s)\,dw_1(s)\right)'$$

$$\left[R^{*}\left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 V^F(r)\left(V^F(r)\right)'dr\right.$$

$$\left.\left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}(R^{*})'\right]^{-1}$$

$$\left(R^{*}\left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 F^X(s)\,dw_1(s)\right). \quad \blacksquare$$

D.4.3 Proof of (c)

The hypothesis under consideration is $H_0^F : R^F\alpha = r$, in models of the type

$$y = \mathbf{f}(T)\,\alpha + u,$$

and in this case

$$F^{*} = T\left(R^F\hat{\alpha} - r\right)'\left[R^F\hat{B}R^{F\prime}\right]^{-1}\left(R^F\hat{\alpha} - r\right)/q.$$

In this model,

$$T^{-\frac{1}{2}}\tau_T\hat{S}_{[rT]} = T^{-\frac{1}{2}}\tau_T\sum_{t=1}^{[rT]} f(t)\,u_t - T^{-1}\tau_T\sum_{t=1}^{[rT]} f(t)\,f(t)'\,\tau_T\left\{\tau_T^{-1}T^{\frac{1}{2}}(\hat{\alpha}-\alpha)\right\}$$

$$\Rightarrow \sigma\int_0^r F(s)\,dw_1(s)$$

$$-\sigma\int_0^r F(s)\,F(s)'\,ds\left(\int_0^1 F(s)\,F(s)'\,ds\right)^{-1}\int_0^1 F(s)\,dw_1(s),$$

and therefore

$$\begin{aligned}\tau_T \hat{C} \tau_T &= T^{-1} \sum_{t=1}^{T} \left(T^{-\frac{1}{2}} \tau_T \hat{S}_t\right) \left(T^{-\frac{1}{2}} \tau_T \hat{S}_t\right)' \\ &\Rightarrow \sigma^2 \int_0^1 V^F(r) \left(V^F(r)\right)' dr,\end{aligned}$$

and

$$\begin{aligned}\hat{B} &= \left(T^{-1} \mathbf{f}(T)' \mathbf{f}(T)\right)^{-1} \hat{C} \left(T^{-1} \mathbf{f}(T)' \mathbf{f}(T)\right)^{-1} \\ &\Rightarrow \sigma^2 \left(\int_0^1 F(s) F(s)' ds\right)^{-1} \int_0^1 V^F(r) \left(V^F(r)\right)' dr \left(\int_0^1 F(s) F(s)' ds\right)^{-1}.\end{aligned}$$

The asymptotic distribution of F^* can now be obtained:

$$\begin{aligned}
F^* &= T\left(R^F\hat{\alpha}-r\right)'\left[R^F\hat{B}R^{F\prime}\right]^{-1}\left(R^F\hat{\alpha}-r\right)/q \\
&= T\left(\left(A^{-1}R^F\tau_T\right)\tau_T^{-1}\left(\hat{\alpha}-\alpha\right)\right)'\left[\left(A^{-1}R^F\tau_T\right)\tau_T^{-1}\hat{B}\tau_T^{-1}\left(\tau_T R^{F\prime}A^{-1}\right)\right]^{-1}\times \\
&\quad \left(\left(A^{-1}R^F\tau_T\right)\tau_T^{-1}\left(\hat{\alpha}-\alpha\right)\right)/q \\
&= \left(\left(A^{-1}R^F\tau_T\right)T^{\frac{1}{2}}\tau_T^{-1}\left(\hat{\alpha}-\alpha\right)\right)'\left[\left(A^{-1}R^F\tau_T\right)\tau_T^{-1}\hat{B}\tau_T^{-1}\left(\tau_T R^{F\prime}A^{-1}\right)\right]^{-1}\times \\
&\quad \left(\left(A^{-1}R^F\tau_T\right)T^{\frac{1}{2}}\tau_T^{-1}\left(\hat{\alpha}-\alpha\right)\right)/q \\
&\Rightarrow \left[R^*\left(\int_0^1 F(s)\,F(s)'\,ds\right)^{-1}\int_0^1 F(s)\,dw_1(s)\right]' \\
&\quad \left[R^*\left(\int_0^1 F(s)\,F(s)'\,ds\right)^{-1}\int_0^1 V^F(r)\left(V^F(r)\right)'dr\right. \\
&\quad \left.\left(\int_0^1 F(s)\,F(s)'\,ds\right)^{-1}\left(R^*\right)'\right]^{-1} \\
&\quad \left(R^*\left(\int_0^1 F(s)\,F(s)'\,ds\right)^{-1}\int_0^1 F(s)\,dw_1(s)\right). \qquad \blacksquare
\end{aligned}$$

D.5 Proof of Lemma 5

This proof is virtually the same as that of Theorem 12. Only the few parts that are different will be redone.

D.5.1 Proof of (a)

As in the proof of Theorem 12, the hypothesis of interest is $H_0 : \beta_1^* = r$ in the model

$$y^* = X_1^* \beta_1^* + u^*,$$

where

$$\begin{aligned} X_1^* &= M_f M_{\tilde{X}_2} \tilde{X}_1, \\ u^* &= M_f M_{\tilde{X}_2} u \text{ and} \\ y^* &= M_f M_{\tilde{X}_2} y. \end{aligned}$$

What remains is then to derive the asymptotic distribution of F^*. To this end, let σ_u^2 denote the long run variance of u. In the transformed model,

$$F^* = T\left(\hat{\beta}_1^* - r\right)' [B^*]^{-1} \left(\hat{\beta}_1^* - r\right).$$

Now, since u^* is a unit root process,

$$\begin{aligned} \left(\hat{\beta}_1^* - \beta_1^*\right) &= \left[T^{-2} (X_1^*)' X_1^*\right]^{-1} T^{-2} (X_1^*)' u^* \\ &\Rightarrow \sigma_u \left(\Lambda^{*'}\right)^{-1} \left(\int_0^1 \hat{w}_q^F(s) \hat{w}_q^F(s)' ds\right)^{-1} \int_0^1 \hat{w}_q^F(s) w_1(s) ds \end{aligned}$$

and

$$\begin{aligned} T^{-2}\hat{S}_{1[rT]} &= T^{-2}\sum_{t=1}^{[rT]} X_{1t}^{*}u^{*} - T^{-2}\sum_{t=1}^{[rT]} X_{1t}^{*}X_{1t}^{*}\left(\hat{\beta}_{1}^{*}-\beta_{1}^{*}\right) \\ &\Rightarrow \sigma_{u}\Lambda^{*}\int_{0}^{r}\hat{w}_{q}^{F}(s)\,w_{1}(s)\,ds \\ &\quad -\sigma_{u}\Lambda^{*}\left(\int_{0}^{r}\hat{w}_{q}^{F}(s)\,\hat{w}_{q}^{F}(s)'\,ds\right)\left(\int_{0}^{1}\hat{w}_{q}^{F}(s)'\,\hat{w}_{q}^{F}(s)\,ds\right)^{-1}\cdot \\ &\quad \int_{0}^{1}\hat{w}_{q}^{F}(s)\,w_{1}(s)\,ds \\ &= \sigma_{u}\Lambda^{*}V_{u}(r) \end{aligned}$$

and therefore

$$T^{-3}\hat{C}_{1}^{*} \Rightarrow \sigma_{u}^{2}\Lambda^{*}\int_{0}^{1}V_{u}(s)'\,V_{u}(s)\,ds\,(\Lambda^{*})'$$

and

$$\begin{aligned} T^{-1}B^{*} &= \left[T^{-2}(X_{1}^{*})'\,X_{1}^{*}\right]^{-1}T^{-3}\hat{C}_{1}^{*}\left[T^{-2}(X_{1}^{*})'\,X_{1}^{*}\right]^{-1} \\ &\Rightarrow \sigma_{u}^{2}(\Lambda^{*\prime})^{-1}\left(\int_{0}^{1}\hat{w}_{q}^{F}(s)'\,\hat{w}_{q}^{F}(s)\,ds\right)^{-1}\int_{0}^{1}V_{u}(s)'\,V_{u}(s)\,ds\,\times \\ &\quad \left(\int_{0}^{1}\hat{w}_{q}^{F}(s)'\,\hat{w}_{q}^{F}(s)\,ds\right)^{-1}(\Lambda^{*})^{-1}. \end{aligned}$$

Finally

$$\begin{aligned}
F^* &= T\left(\hat{\beta}_1^* - r\right)' [B^*]^{-1} \left(\hat{\beta}_1^* - r\right) \\
&= \left(\hat{\beta}_1^* - \beta_1^*\right)' \left[T^{-1}B^*\right]^{-1} \left(\hat{\beta}_1^* - \beta_1^*\right) \\
&\Rightarrow \left[\Lambda^* \left(\int_0^1 \hat{w}_q^F(s)\,\hat{w}_q^F(s)'\,ds\right)^{-1} \int_0^1 \hat{w}_q^F(s)\,w_1(s)\,ds\right]' \\
&\quad \left[\Lambda^* \left(\int_0^1 \hat{w}_q^F(s)\,\hat{w}_q^F(s)'\,ds\right)^{-1} \int_0^1 V_u(s)'\,V_u(s)\,ds\right. \\
&\quad \left.\left(\int_0^1 \hat{w}_q^F(s)\,\hat{w}_q^F(s)'\,ds\right)^{-1} (\Lambda^*)'\right]^{-1} \\
&\quad \left[\Lambda^* \left(\int_0^1 \hat{w}_q^F(s)\,\hat{w}_q^F(s)'\,ds\right)^{-1} \int_0^1 \hat{w}_q^F(s)\,w_1(s)\,ds\right] \\
&= \left[\int_0^1 \hat{w}_q^F(s)\,w_1(s)\,ds\right]' \left[\int_0^1 V_u(s)'\,V_u(s)\,ds\right]^{-1} \left[\int_0^1 \hat{w}_q^F(s)\,w_1(s)\,ds\right]
\end{aligned}$$

and the proof of (a) is complete. ■

D.5.2 Proof of (b)

The hypothesis being tested is $H_0^F : R^F\alpha = r$, in the transformed model

$$\bar{y} = \mathbf{f}^X(T)\,\alpha + \bar{u},$$

and what remains is to derive the asymptotic distribution of

$$F_1^* = T\left(R^F\hat{\alpha} - r\right)' \left[R^F\hat{B}^X R^{F\prime}\right]^{-1} \left(R^F\hat{\alpha} - r\right)/q.$$

Now, since $\bar{u}$ is a unit root process,

$$\tau_T^{-1}T^{-\frac{1}{2}}\left(\hat{\alpha} - \alpha\right) \Rightarrow \sigma_u \left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1} \int_0^1 F^X(s)\,w_1(s)\,ds$$

and

$$
\begin{aligned}
T^{-\frac{3}{2}}\tau_T\hat{S}^X_{[rT]} &= T^{-\frac{3}{2}}\tau_T\sum_{t=1}^{[rT]} f^X(t)'\,\widehat{u}_t \\
&= T^{-\frac{3}{2}}\tau_T\sum_{t=1}^{[rT]} f^X(t)'\,\bar{u}_t - \left(T^{-1}\tau_T\sum_{t=1}^{[rT]} f^X(t)' f^X(t)\,\tau_T\right) \\
&\quad \left(\tau_T^{-1}T^{-\frac{1}{2}}(\hat{\alpha}-\alpha)\right) \\
&\Rightarrow \sigma_u\int_0^r F^X(s)\,w_1(s)\,ds - \sigma_u\left(\int_0^r F^X(s)\,F^X(s)'\,ds\right)\times \\
&\quad \left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 F^X(s)\,w_1(s)\,ds \\
&= \sigma_u V_u^{F^X}(r),
\end{aligned}
$$

and

$$
\begin{aligned}
T^{-2}\tau_T\hat{C}^X\tau_T &= T^{-1}\sum_{t=1}^{T}\left(T^{-\frac{3}{2}}\tau_T\hat{S}^X_t\right)\left(T^{-\frac{3}{2}}\tau_T\hat{S}^X_t\right)' \\
&\Rightarrow \sigma_u^2\int_0^1 V_u^{F^X}(r)\left(V_u^{F^X}(r)\right)'dr
\end{aligned}
$$

which in turn implies that

$$
\begin{aligned}
T^{-2}\tau_T^{-1}\hat{B}^X\tau_T^{-1} &= \left(T^{-1}\tau_T\mathbf{f}^X(T)'\,\mathbf{f}^X(T)\,\tau_T\right)^{-1}T^{-2}\tau_T\hat{C}^X\tau_T \\
&\quad \left(T^{-1}\tau_T\mathbf{f}^X(T)'\,\mathbf{f}^X(T)\,\tau_T\right)^{-1} \\
&\Rightarrow \sigma_u^2\left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1}\int_0^1 V_u^{F^X}(r)\left(V_u^{F^X}(r)\right)'dr \\
&\quad \left(\int_0^1 F^X(s)\,F^X(s)'\,ds\right)^{-1},
\end{aligned}
$$

implying that under the null

$$\begin{aligned}
F^* &= T\left(R^F(\hat{\alpha}-\alpha)\right)'\left[R^F\hat{B}R^{F\prime}\right]^{-1}\left(R^F(\hat{\alpha}-\alpha)\right)/q \\
&= \left(R^FT^{-\frac{1}{2}}(\hat{\alpha}-\alpha)\right)'\left[R^FT^{-2}\hat{B}R^{F\prime}\right]^{-1}\left(R^FT^{-\frac{1}{2}}(\hat{\alpha}-\alpha)\right)/q \\
&= \left(\left(A^{-1}R^F\tau_T\right)T^{-\frac{1}{2}}\tau_T^{-1}(\hat{\alpha}-\alpha)\right)'\left[\left(A^{-1}R^F\tau_T\right)\left(T^{-2}\hat{B}\right)\left(A^{-1}R^F\tau_T\right)'\right]^{-1}\times \\
&\quad \left(\left(A^{-1}R^F\tau_T\right)T^{-\frac{1}{2}}\tau_T^{-1}(\hat{\alpha}-\alpha)\right) \\
&\Rightarrow \left(R^*\sigma_u\left(\int_0^1 F^X(s)F^X(s)'ds\right)^{-1}\int_0^1 F^X(s)dw_1(s)\right)' \\
&\quad \left[R^*\sigma_u^2\left(\int_0^1 F^X(s)F^X(s)'ds\right)^{-1}\int_0^1 V_u^{FX}(r)\left(V_u^{FX}(r)\right)'dr\right. \\
&\quad \left.\left(\int_0^1 F^X(s)F^X(s)'ds\right)^{-1}(R^*)'\right]^{-1} \\
&\quad \left(R^*\sigma_u\left(\int_0^1 F^X(s)F^X(s)'ds\right)^{-1}\int_0^1 F^X(s)dw_1(s)\right) \\
&= \left(R^*\left(\int_0^1 F^X(s)F^X(s)'ds\right)^{-1}\int_0^1 F^X(s)dw_1(s)\right)' \\
&\quad \left[R^*\left(\int_0^1 F^X(s)F^X(s)'ds\right)^{-1}\int_0^1 V_u^{FX}(r)\left(V_u^{FX}(r)\right)'dr\right. \\
&\quad \left.\left(\int_0^1 F^X(s)F^X(s)'ds\right)^{-1}(R^*)'\right]^{-1} \\
&\quad \left(R^*\left(\int_0^1 F^X(s)F^X(s)'ds\right)^{-1}\int_0^1 F^X(s)dw_1(s)\right). \quad \blacksquare
\end{aligned}$$

D.5.3 Proof of (c)

The hypothesis of interest is $H_0^F : R^F\alpha = r$ in models of the type

$$y = \mathbf{f}(T)\alpha + u,$$

and

$$F^* = T\left(R^F\hat{\alpha} - r\right)'\left[R^F\hat{B}R^{F\prime}\right]^{-1}\left(R^F\hat{\alpha} - r\right)/q.$$

Now, since u is a unit root process,

$$\tau_T^{-1}T^{-\frac{1}{2}}(\hat{\alpha} - \alpha) \Rightarrow \sigma_u\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 F(s)w_1(s)ds$$

and

$$\begin{aligned} T^{-\frac{3}{2}}\tau_T\hat{S}_{[rT]} &= T^{-\frac{3}{2}}\tau_T\sum_{t=1}^{[rT]} f(t)u_t - T^{-1}\tau_T\sum_{t=1}^{[rT]} f(t)f(t)'\tau_T\left\{\tau_T^{-1}T^{-\frac{1}{2}}(\hat{\alpha} - \alpha)\right\} \\ &\Rightarrow \sigma_u\int_0^r F(s)dw_1(s) \\ &\quad -\sigma_u\int_0^r F(s)F(s)'ds\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 F(s)dw_1(s), \end{aligned}$$

and therefore

$$\begin{aligned} T^{-2}\tau_T\hat{C}\tau_T &= T^{-1}\sum_{t=1}^{T}\left(T^{-\frac{3}{2}}\tau_T\hat{S}_t\right)\left(T^{-\frac{3}{2}}\tau_T\hat{S}_t\right)' \\ &\Rightarrow \sigma_u^2\int_0^1 V_u^F(r)\left(V_u^F(r)\right)'dr \end{aligned}$$

and

$$T^{-2}\hat{B} = \left(T^{-1}\mathbf{f}(T)'\mathbf{f}(T)\right)^{-1} T^{-2}\hat{C}\left(T^{-1}\mathbf{f}(T)'\mathbf{f}(T)\right)^{-1}$$

$$\Rightarrow \sigma_u^2\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 V_u^F(r)\left(V_u^F(r)\right)'dr \left(\int_0^1 F(s)F(s)'ds\right)^{-1}.$$

Finally

$$F^* = T\left(R^F(\hat{\alpha}-\alpha)\right)'\left[R^F\hat{B}R^{F\prime}\right]^{-1}\left(R^F(\hat{\alpha}-\alpha)\right)/q$$

$$= T\left(\left(A^{-1}R^F\tau_T\right)\tau_T^{-1}(\hat{\alpha}-\alpha)\right)'\left[\left(A^{-1}R^F\tau_T\right)\tau_T^{-1}\hat{B}\tau_T^{-1}\left(\tau_T R^{F\prime}A^{-1}\right)\right]^{-1}\times \left(\left(A^{-1}R^F\tau_T\right)\tau_T^{-1}(\hat{\alpha}-\alpha)\right)/q$$

$$= \left(\left(A^{-1}R^F\tau_T\right)T^{-\frac{1}{2}}\tau_T^{-1}(\hat{\alpha}-\alpha)\right)'\times \left[\left(A^{-1}R^F\tau_T\right)\left(T^{-2}\tau_T^{-1}\hat{B}\tau_T^{-1}\right)\left(\tau_T R^{F\prime}A^{-1}\right)\right]^{-1}\times \left(\left(A^{-1}R^F\tau_T\right)T^{-\frac{1}{2}}\tau_T^{-1}(\hat{\alpha}-\alpha)\right)/q$$

$$\Rightarrow \left[R^*\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 F(s)w_1(s)ds\right]' \left[R^*\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 V_u^F(r)\left(V_u^F(r)\right)'dr \left(\int_0^1 F(s)F(s)'ds\right)^{-1}(R^*)'\right]^{-1} \left(R^*\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 F(s)w_1(s)ds\right). \quad \blacksquare$$

D.6 Determining the Number of Additional Regressors

The issue which is investigated in this appendix is the choice of the number of additional regressors (m) to include in regression (6.6). To this end, it is useful to model the error terms in more detail than has been done before. Specifically, consider the model

$$y_t = \alpha_1 + \alpha_2 t + u_t, \ t = 1, ..., T$$

$$(1 - L\gamma_T) u_t = e_t, \quad t = 2, 3, ..., T, \quad u_1 = \sum_{i=1}^{[\kappa T]} (\gamma_T)^i e_{1-i},$$

$$e_t = d(L) \varepsilon_t, \quad d(L) = \sum_{i=0}^{\infty} d_i L^i, \quad \sum_{i=0}^{\infty} i |d_i| < \infty, \quad d(1)^2 > 0,$$

where $\{e_t\}$ is a martingale difference sequence with $E(e_t^2 | e_{t-1}, e_{t-2}, ...) = 1$ and $\sup_t E(e_t^4) < \infty$. To further simplify matters, κ is set to 0.[1] Under this specification, the errors can be modeled as being stationary by setting $\gamma_T = \bar{\gamma}$. This would imply that $\{u_t\}$ has an AR coefficient of $\bar{\gamma}$. Alternatively, the errors can be modeled as local to a unit root by letting $\gamma_T = \left(1 - \frac{\bar{\gamma}}{T}\right)$. For a fixed sample size T, $\{u_t\}$ will have an AR coefficient of $\left(1 - \frac{\bar{\gamma}}{T}\right)$, but as $T \to \infty$, $\{u_t\}$ will approach a unit root process.

The modelling of the error as local to a unit root is useful when choosing m, because the asymptotic distribution of the teststatistic can be found analytically as a function of $\bar{\gamma}$. This can then be used to simulate the asymptotic size and power of the test for different choices of m. The results of those simulations will

[1] Simulations in Vogelsang [1998] indicate that this will not significantly affect the results.

provide a basis for choosing m. Below, this is done for hypotheses of the form $H_0 : \alpha_2 = r$.[2]

Before proceeding to select m, some additional notation is required. Let $f^m(t) = [\ 1 \quad t \quad t^2 \quad \ldots \quad t^m\ ]'$ and define τ_T^m and $F^m(t)$ such that $\tau_T^m f^m(t) = F^m(t/T) + o(1)$. The asymptotic distribution of $J_T(m)$ and therefore F^J can now be computed. The following lemma follows directly from Theorems 1 and 2 in Vogelsang (1998):

Lemma 7

a) If $\gamma_T = \bar{\gamma}$, with $|\bar{\gamma}| < 1$, then $J_T(m) \Rightarrow 0$.

b) If $\gamma_T = \left(1 - \frac{\bar{\gamma}}{T}\right)$, then $J_T(m) \Rightarrow J_{\bar{\gamma}}(m)$, where

$$J_{\bar{\gamma}}(m) = \frac{\int_0^1 \Psi_{\bar{\gamma}}(s)^2 ds - \int_0^1 F(s)'\Psi_{\bar{\gamma}}(s) ds \left(\int_0^1 F(s)F(s)'ds\right)^{-1} \int_0^1 F(s)\Psi_{\bar{\gamma}}(s) ds}{\int_0^1 \Psi_{\bar{\gamma}}(s)^2 ds - \int_0^1 F^m(s)'\Psi_{\bar{\gamma}}(s) ds \left(\int_0^1 F^m(s)F^m(s)'ds\right)^{-1} \int_0^1 F^m(s)\Psi_{\bar{\gamma}}(s) ds} - 1$$

and

$$\Psi_{\bar{\gamma}}(r) = \int_0^r \exp\left(-\bar{\gamma}\left(r-s\right)\right) dw_1(s).$$

From Lemma 7, it is clear that when the errors are $I(1)$, $J_T(m)$ has a non-degenerate distribution; when the errors are $I(0)$, $J_T(m)$ converges to zero. For the sake of completeness, the exact expression for the asymptotic distribution of F^J is given in the following lemma:

Lemma 8 *a) When $\gamma_T = \bar{\gamma}$ and $|\bar{\gamma}| < 1$,*

$$F^J \Rightarrow \left(-6w_1(1) + 12\int_0^1 s\,ds\right)^2 \left[\left[\begin{array}{cc} -6 & 12 \end{array}\right] \int_0^1 V^F(r) V^F(r)' dr \left[\begin{array}{c} -6 \\ 12 \end{array}\right]\right]^{-1},$$

[2]Note that for the model *with* regressors, the number of simulations required to implement the method described above is substantial, since the critical values differ for each k and q. This implies that a different (m, b) pair would have to be found for each (k, q) combination.

where

$$V^{F}(r)=\begin{bmatrix} w_1(r) \\ \int_0^r s\,dw_1(s) \end{bmatrix}-\begin{bmatrix} 4r-3r^2 & -6r+6r^2 \\ 2r^2-2r^3 & -3r^2+4r^3 \end{bmatrix}\begin{bmatrix} w_1(1) \\ \int_0^1 s\,dw_1(s) \end{bmatrix}.$$

b) When $\gamma_T=\left(1-\frac{\bar{\gamma}}{T}\right)$,

$$F^J \Rightarrow \left(-6\int_0^1 \Psi_{\bar{\gamma}}(s)\,ds+12\int_0^1 s\Psi_{\bar{\gamma}}(s)\,ds\right)^2$$
$$\left[\begin{bmatrix} -6 & 12 \end{bmatrix}\int_0^1 V_{\bar{\gamma}}^F(r)\left(V_{\bar{\gamma}}^F(r)\right)'dr\begin{bmatrix} -6 \\ 12 \end{bmatrix}\right]^{-1}\times\exp(-bJ_{\bar{\gamma}}(m)),$$

where

$$V_{\bar{\gamma}}^F(r)=\begin{bmatrix} \int_0^r \Psi_{\bar{\gamma}}(s)\,ds \\ \int_0^r s\Psi_{\bar{\gamma}}(s)\,ds \end{bmatrix}$$
$$-\begin{bmatrix} 4r-3r^2 & -6r+6r^2 \\ 2r^2-2r^3 & -3r^2+4r^3 \end{bmatrix}\begin{bmatrix} \int_0^1 \Psi_{\bar{\gamma}}(s)\,ds \\ \int_0^1 s\Psi_{\bar{\gamma}}(s)\,ds \end{bmatrix}.$$

D.6.1 Proof of Lemma 8

First note that part a) follows trivially from Theorem 6.3. To prove b), the same principles as in Vogelsang (1998) will be employed. For the derivations below, it is necessary to define $\Psi_{\bar{\gamma}}(r)=\int_0^r \exp(-\bar{\gamma}(r-s))\,dw_1(s)$. Then

$$T^{-\frac{1}{2}}\tau_T^{-1}(\hat{\alpha}-\alpha)\Rightarrow d(1)\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 F(s)\Psi_{\bar{\gamma}}(s)\,ds$$

and

$$T^{-2}\tau_T^{-1}\hat{B}\tau_T^{-1} \Rightarrow d(1)^2\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 V_{\bar{\gamma}}^F(r)\left(V_{\bar{\gamma}}^F(r)\right)'dr\cdot\left(\int_0^1 F(s)F(s)'ds\right)^{-1},$$

where

$$V_{\bar{\gamma}}^{F}(r) = \int_0^r F(s)\,\Psi_{\bar{\gamma}}(s)\,ds - \left(\int_0^r F(s)\,F(s)'\,ds\right)\left(\int_0^1 F(s)\,F(s)'\,ds\right)^{-1}\int_0^1 F(s)\,\Psi_{\bar{\gamma}}(s)\,ds.$$

Then, from Theorem 2 in Vogelsang (1998), it is known that

$$J_T(m) = \frac{SSR_R - SSR_U}{SSR_U} \Rightarrow$$

$$J_{\bar{\gamma}}(m) = \frac{\int_0^1 \Psi_{\bar{\gamma}}(s)^2 ds - \int_0^1 F(s)'\Psi_{\bar{\gamma}}(s)ds\left(\int_0^1 F(s)F(s)'ds\right)^{-1}\int_0^1 F(s)\Psi_{\bar{\gamma}}(s)ds}{\int_0^1 \Psi_{\bar{\gamma}}(s)^2 ds - \int_0^1 F^m(s)'\Psi_{\bar{\gamma}}(s)ds\left(\int_0^1 F^m(s)F^m(s)'ds\right)^{-1}\int_0^1 F^m(s)\Psi_{\bar{\gamma}}(s)ds} - 1,$$

and hence, letting $R = [0\ 1]$,

$$F^J \Rightarrow \left(d(1)\, R \left(\int_0^1 F(s)\, F(s)'\, ds\right)^{-1} \int_0^1 F(s)\, \Psi_{\bar{\gamma}}(s)\, ds\right)'$$

$$\left[d(1)^2\, R \left(\int_0^1 F(s)\, F(s)'\, ds\right)^{-1} \int_0^1 V_{\bar{\gamma}}^F(r)\, \left(V_{\bar{\gamma}}^F(r)\right)'\, dr\right.$$

$$\left.\left(\int_0^1 F(s)\, F(s)'\, ds\right)^{-1} R'\right]^{-1}$$

$$\left(d(1)\, R \left(\int_0^1 F(s)\, F(s)'\, ds\right)^{-1} \int_0^1 F(s)\, \Psi_{\bar{\gamma}}(s)\, ds\right) \exp(-bJ_{\bar{\gamma}}(m))$$

$$= \left(R \left(\int_0^1 F(s)\, F(s)'\, ds\right)^{-1} \int_0^1 F(s)\, \Psi_{\bar{\gamma}}(s)\, ds\right)'$$

$$\left[R \left(\int_0^1 F(s)\, F(s)'\, ds\right)^{-1} \int_0^1 V_{\bar{\gamma}}^F(r)\, \left(V_{\bar{\gamma}}^F(r)\right)'\, dr\right.$$

$$\left.\left(\int_0^1 F(s)\, F(s)'\, ds\right)^{-1} R'\right]^{-1}$$

$$\left(R \left(\int_0^1 F(s)\, F(s)'\, ds\right)^{-1} \int_0^1 F(s)\, \Psi_{\bar{\gamma}}(s)\, ds\right) \exp(-bJ_{\bar{\gamma}}(m)).$$

Using $F(s)' = [1\ s]$ simple algebra yields the result. ■

The asymptotic distribution of F^J is the same as that of F^* when the errors are stationary. When the errors are local to a unit root, this is no longer the case. As expected, the distribution does not depend on nuisance parameters, but it does depend on m and on b.

Table D.1: Asymptotic Size of F*, PST and PSWT

	γ						
m	0	2	4	6	8	10	12
3	0.050	0.033	0.028	0.026	0.0286	0.028	0.030
4	0.050	0.029	0.023	0.022	0.0231	0.024	0.027
5	0.050	0.025	0.019	0.018	0.0192	0.021	0.023
6	0.050	0.026	0.019	0.017	0.0181	0.020	0.022
7	0.050	0.026	0.019	0.016	0.0173	0.019	0.021
8	0.050	0.024	0.018	0.015	0.0162	0.018	0.019
9	0.050	0.024	0.016	0.014	0.0145	0.017	0.020
10	0.050	0.022	0.015	0.014	0.0144	0.017	0.018

Nominal size = 0.05

Model: $y_t = \alpha_1 + \alpha_{2t} + u_t$, $u_t = (1-\gamma/T)u_{t-1} + e_t$

H_0: $\alpha_2 = \alpha_{20}$

Table D.2: Asymptotic Power of F*, PST and PSWT

	c/d(1)	m = 3	m = 4	m = 5	m = 6	m = 7	m = 8	m = 9	m = 10
$\gamma = 0$	5	0.151	0.197	0.223	0.239	0.252	0.257	0.262	0.273
$\gamma = 0$	10	0.188	0.261	0.304	0.326	0.351	0.355	0.366	0.377
$\gamma = 0$	15	0.209	0.294	0.344	0.373	0.400	0.405	0.422	0.437
$\gamma = 0$	20	0.224	0.319	0.373	0.407	0.430	0.444	0.461	0.479
$\gamma = 0$	25	0.237	0.341	0.397	0.432	0.460	0.475	0.493	0.506
$\gamma = 5$	5	0.252	0.335	0.379	0.404	0.425	0.431	0.441	0.452
$\gamma = 5$	10	0.304	0.425	0.484	0.515	0.540	0.552	0.573	0.589
$\gamma = 5$	15	0.333	0.474	0.537	0.568	0.599	0.617	0.638	0.650
$\gamma = 5$	20	0.352	0.503	0.571	0.606	0.641	0.655	0.677	0.692
$\gamma = 5$	25	0.367	0.527	0.599	0.632	0.666	0.686	0.708	0.723
$\gamma = 10$	5	0.399	0.549	0.604	0.636	0.663	0.676	0.695	0.707
$\gamma = 10$	10	0.467	0.645	0.710	0.742	0.775	0.792	0.809	0.819
$\gamma = 10$	15	0.506	0.691	0.756	0.795	0.824	0.841	0.854	0.869
$\gamma = 10$	20	0.533	0.720	0.786	0.825	0.855	0.881	0.895	0.904
$\gamma = 10$	25	0.553	0.742	0.806	0.846	0.875	0.886	0.903	0.913

Nominal size = 0.05

Model: $y_t = \alpha_1 + \alpha_{2t} + u_t$, $u_t = (1-\gamma/T)u_{t-1} + e_t$

H_0: $\alpha_2 = \alpha_{20}$, $\alpha_2 = \alpha_{20} + T^{-1/2}c$

Using the distribution of the modified test statistic derived in Lemma 8, the asymptotic size of the test for values of m ranging from 3 to 9 and $\bar{\gamma}$ from 0 to 12 have been simulated and are tabulated in Table D.1. It is clear that the while size is decreasing in m, the changes are very small.

As the final criterion for choosing m, asymptotic power under local alternatives is simulated for different local-to-unit-root processes. The local alternative employed is

$$H_1 : \alpha_2 = r + T^{-\frac{1}{2}}c.$$

The following lemma provides the distribution of the test statistic under H_1 :

Lemma 9 *If* $\gamma_T = \left(1 - \frac{\bar{\gamma}}{T}\right)$, *then under* H_1,

$$\begin{aligned} F^J \Rightarrow & \left(\begin{bmatrix} -6 & 12 \end{bmatrix} \begin{bmatrix} \int_0^1 \Psi_{\bar{\gamma}}(s)\, ds \\ \int_0^1 s\Psi_{\bar{\gamma}}(s)\, ds \end{bmatrix} + \frac{c}{d(1)} \right)' \\ & \left[\begin{bmatrix} -6 & 12 \end{bmatrix} \int_0^1 V_{\bar{\gamma}}^F(r) \left(V_{\bar{\gamma}}^F(r)\right)' dr \begin{bmatrix} -6 \\ 12 \end{bmatrix} \right]^{-1} \\ & \left(\begin{bmatrix} -6 & 12 \end{bmatrix} \begin{bmatrix} \int_0^1 \Psi_{\bar{\gamma}}(s)\, ds \\ \int_0^1 s\Psi_{\bar{\gamma}}(s)\, ds \end{bmatrix} + \frac{c}{d(1)} \right) \exp(-bJ_{\bar{\gamma}}(m)). \end{aligned}$$

D.6.2 Proof of Lemma 9

Under the alternative

$$F^J = T\left(\hat{\alpha}_2 - r - T^{-\frac{1}{2}}c\right)' \left[R\hat{B}R'\right]^{-1} \left(\hat{\alpha}_2 - r - T^{-\frac{1}{2}}c\right) / \left(q \exp(bJ_T(m)\right).$$

Clearly neither the distribution of $R\hat{B}R'$ nor the distribution of $\exp(bJ_T(m)$ are affected by the hypothesis. Rewriting F^J,

$$\begin{aligned}
F^J &= T\left(\hat{\alpha}_2 - r - T^{-\frac{1}{2}}c\right)'\left[R\hat{B}R'\right]^{-1}\left(\hat{\alpha}_2 - r - T^{-\frac{1}{2}}c\right)/\left(q\exp(bJ_T(m)\right) \\
&= \left(T^{\frac{1}{2}}(\hat{\alpha}-\alpha)-c\right)'\left[R\hat{B}R'\right]^{-1}\left(T^{\frac{1}{2}}(\hat{\alpha}-\alpha)-c\right)/\left(q\exp(bJ_T(m)\right) \\
&\Rightarrow \left(R\left(\int_0^1 F(s)\,F(s)'\,ds\right)^{-1}\int_0^1 F(s)\,\Psi_{\bar{\gamma}}(s)\,ds + c/d(1)\right)' \\
&\quad \left[R\left(\int_0^1 F(s)\,F(s)'\,ds\right)^{-1}\int_0^1 V_{\bar{\gamma}}^F(r)\left(V_{\bar{\gamma}}^F(r)\right)'dr\right. \\
&\quad \left.\left(\int_0^1 F(s)\,F(s)'\,ds\right)^{-1}R'\right]^{-1} \\
&\quad \left(R\left(\int_0^1 F(s)\,F(s)'\,ds\right)^{-1}\int_0^1 F(s)\,\Psi_{\bar{\gamma}}(s)\,ds + c/d(1)\right)\exp(-bJ_{\bar{\gamma}}(m)) \\
&= \left(\begin{bmatrix} -6 & 12 \end{bmatrix}\begin{bmatrix} \int_0^1 \Psi_{\bar{\gamma}}(s)\,ds \\ \int_0^1 s\Psi_{\bar{\gamma}}(s)\,ds \end{bmatrix} + c/d(1)\right)' \\
&\quad \left[\begin{bmatrix} -6 & 12 \end{bmatrix}\int_0^1 V_{\bar{\gamma}}^F(r)\left(V_{\bar{\gamma}}^F(r)\right)'dr\begin{bmatrix} -6 \\ 12 \end{bmatrix}\right]^{-1} \\
&\quad \left(\begin{bmatrix} -6 & 12 \end{bmatrix}\begin{bmatrix} \int_0^1 \Psi_{\bar{\gamma}}(s)\,ds \\ \int_0^1 s\Psi_{\bar{\gamma}}(s)\,ds \end{bmatrix} + c/d(1)\right)\exp(-bJ_{\bar{\gamma}}(m)). \qquad \blacksquare
\end{aligned}$$

Using the distribution in Lemma 9, the asymptotic power of F^J has been simulated for $c/d(1) = 5, 10, 15, 20, 25$, $\bar{\gamma} = 0, 5, 10$ and $m = 3, .., 10$. The results can be found in Table D.2. Power increases with $c/d(1)$ and $\bar{\gamma}$. While power is strictly increasing in m, it seems to flatten around $m = 9$. Since size wasn't affected by m, m is chosen to be 9. The corresponding value for b at the 95% level is 1.82.

References

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

[2] Andrews, D.W.K. and Monahan J.C (1992), "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60, 953-966

[3] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Jon A. Wellner (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: The Johns Hopkins University Press

[4] Bandorff-Neilsen, O.E. (1978) *Information and Exponential Families in Statistical Theory.* Chichester: Wiley

[5] Christensen, B. J., and N. M. Kiefer (1994). "Local Cuts and Separate Inference", *Scandinavian Journal of Statistics*, 21, 389-407.

[6] Davidson, R. and J.G. MacKinnon (1993), *Estimation and Inference in Econometrics,* New York: Oxford University Press.

[7] De Brouwer, G. and N.R. Ericsson (1998) "Modeling Inflation in Australia," *Journal of Business and Economic Statistics*, October, 16 (4), 433-449

[8] Den Haan, W.J. and A. Levin (1997), "A Practictioner's Guide to Robust Covariance Matrix Estimation," Chapter 12 in *Handbook of Statistics: Robust Inference*, volume 15, ed. by G.S. Maddala and C.R. Rao, New York: Elsevier.

[9] Gallant, A.R. (1987), *Nonlinear Statistical Models*, New York: Wiley

[10] Godambe, V.P. *Estimating Functions*, Oxford: Claredon Press

[11] Hall, P. and C. C Heyde (1980), *Martingale Limit Theory and Its Applications*, New York: Academic Press.

[12] Hansen, Bruce E. (1992a) "Convergence of Stochastic Integrals for Dependent Heterogeneous Processes", *Econometric Theory*, 8, 489-500

[13] Hansen, B.E. (1992), "Consistent Covariance Matrix Estimation for Dependent Heterogenous Processes," *Econometrica*, 60, 967-972

[14] Hansen, L. (1982), "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50, 1029-1054

[15] Hamilton, James D. (1994) *Time Series Analysis,* Princeton University Press, Princeton, NJ

[16] Heyde, C. C. *Quasi-likelihood and its Applications: A General Approach to Optimal Parameter Estimation.* Springer Series in Statistics. Springer-Verlag, New York, 1997

[17] Hsiao, C (1986) *Analysis of Panel Data.* Econometric Society Monographs Number 11, Cambridge University Press

[18] Johansen, S (1988) "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, 12, 231-254

[19] — (1991) "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, 59, 1551-1580

[20] — and K. Juselius (1990) "Maximum Likelihood Estimation and Inference on Cointegration: With Applications to the Demand for Money," *Oxford Bulletin of Economics and Statistics*, 52, 169-210

[21] — (1992) "Testing Structural Hypotheses in a Multivariate Cointegration Analysis of the PPP and the UIP for UK," *Journal of Econometrics*, 53, 211-44

[22] MaCurdy T.E (1981) "An Empirical Model of Labor Supply in a Life-Cycle Setting", *Journal of Political Economy*, 9, 1059-1085

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

[24] Metin, K. (1998) "The Relationship between Inflation and the Budget Deficit in Turkey," *Journal of Business and Economic Statistics*, October, 16 (4), 412-422

[25] Park, J.Y (1990) "Testing for Unit Roots and Cointegration by Variable Addition," in *Advances in Econometrics: Cointegration, Spurious Regressions and Unit Roots*, ed. T. Fomby and F. Rhodes. London: Jai Press, pp. 107-34

[26] — and B. Choi (1988) " A New Approach to Testing for a Unit Root," CAE Working Paper # 88-23, Cornell University

[27] — and P.C.B Phillips (1988) "Statistical Inference in Regressions with Integrated Processes: I," *Econometric Theory*, 4, 467-497

[28] Phillips, P.C.B. and S. Durlauf (1986), "Multiple Time Series Regression with Integrated Processes," *Review of Economic Studies,* 53, 473-496

[29] — and Mico Loretan (1991), "Estimating Long-Run Economic Equlibria," *Review of Economic Studies*, 58, 165-93

[30] — and J.Y. Park (1988) "Asymptotic Equivalence of Ordinary Least Squares and Generalized Least Squares in Regressions with Integrated Regressors", *Journal of the American Statistical Association*, March, 83, 111-115

[31] Robinson, P.M (1991), "Automatic Frequency Domain Inference on Semi-parametric and Nonparametric Models", *Econometrica*, 59, 1329-63

[32] — (1998), "Inference-Without Smoothing in the Presence of Nonparametric Autocorrelation," *Econometrica*, 66, 1163-1182.

[33] Saikkonen, Pentti (1991) "Asymptotically Efficient Estimation of Cointegrating Regressions," *Econometric Theory,* 7, 1-21

[34] Stock, James H. and Mark W. Watson (1988) "Testing for Common Trends," *Journal of the American Statistical Association*, 83, 1097-1107

[35] Vogelsang, T.J. (1998), "Trend Function Hypothesis Testing in the Presence of Serial Correlation," *Econometrica*, 66, 123-148.

[36] White, H. (1984), *Asymptotic Theory for Econometricians*, New York: Academic Press

[37] Woolridge, Jefferey M. (1991), "Notes on Regression with Difference-Stationary Data" Michigan State University, Mimeo