Chapter 6

Asymptotics for Unit Root Tests Based on M Estimators

This chapter is largely based on Lucas (1995b). It considers unit root tests based on M estimators. The asymptotic theory for these tests is developed. It is shown how the asymptotic distributions of the tests depend on nuisance parameters and how tests can be constructed that are invariant to these parameters. It is also shown that a particular linear combination of a unit root test based on the OLS estimator and on an M estimator converges to a normal random variate. The interpretation of this result is discussed. A simulation experiment is provided, illustrating the level and power of different unit root tests for several sample sizes and data generating processes. The tests based on M estimators turn out to be more powerful than the OLS based tests if the innovations are fat-tailed.

The chapter is set up as follows. Section 6.1 briefly describes how the present chapter fits into the existing literature. Section 6.2 presents the main theoretical results. Section 6.3 comments on the choice of the M estimator for the applied researcher. Section 6.4 presents the results of a simulation experiment. Section 6.5 concludes the chapter. The proofs of the the statements in Section 6.2 can be found in the appendix to this chapter.

6.1 Introduction

In the previous two chapters, I proposed outlier robust alternatives to the OLS-based unit root testing procedure of Dickey and Fuller (1979). The Dickey-Fuller t-test (DF-t) is well known and widely used. Extensions of the test that allow for serially correlated innovations are given in Phillips (1987) and Phillips and Perron (1988).

The idea that OLS based testing procedures are nonrobust, is not new. It has since long been known in the literature that the OLS estimator is sensitive to the occurrence of outliers in the data, both in the setting of independently and identically distributed (i.i.d.) random variables (see Hampel et al. (1986)) and in the context of stationary time series (Martin and Yohai (1986)). In the setting of nonstationary time series, however, the OLS estimator can have some desirable robustness properties. As is shown in Phillips (1987) and Phillips and Perron (1988), the innovations driving a random walk may demonstrate a considerable amount of heterogeneity and temporal dependence without seriously changing the asymptotic properties of the OLS estimator. Tests can be designed that have the same asymptotic distribution under the hypothesis of a unit root for a wide variety of data generating processes. Several well known outlier types (see, e.g., Martin and Yohai (1986)) are also covered by the assumptions in Phillips (1987). Moreover, the influence curve of the OLS estimator in the random walk context is identically equal to zero (see Chapter 4). These findings would suggest that in the context of testing the unit root hypothesis there is no pressing need to replace OLS by an outlier robust estimation technique. As the previous chapters have demonstrated, however, certain types of outliers (especially additive outliers) can significantly distort the finite sample behavior of the DF-t and its extensions, whereas outlier robust variants of these tests are less affected. Moreover, the use of outlier robust estimators can positively affect the power properties of the tests, see Cox and Llatas (1991), Hampel et al. (1986), and Herce (1993). These points lead to the consideration of the behavior of unit root tests based on outlier robust estimators.

The main idea of this chapter is to develop an asymptotic theory for unit root tests based on M estimators. Furthermore, the effectiveness of these tests in finite samples is evaluated by means of a simulation experiment. M estimators are known to possess a certain degree of insensitivity to outliers. Their behavior for integrated processes with infinite variance innovations has been studied by Knight (1989, 1991). Herce (1993) discussed the asymptotic properties of unit root tests based on the least absolute deviations (LAD) estimator when the innovations have finite variance. The present chapter fits between these articles. It considers processes with finite variance innovations and estimators that are defined by smoother objective functions than the LAD estimator. The behavior of M estimators in nearly nonstationary models is treated in Cox and Llatas (1991). Their results can be used to study the asymptotic power properties of the present tests (compare Chapter 7).

The notation used in this chapter for expressing the limiting distributions of unit root test statistics was explained in Subsection 1.4.4.

6.2 An Asymptotic Analysis

Consider the process $\{y_t\}$, generated by the stochastic difference equation

$$y_t = y_{t-1} + \varepsilon_t. \tag{6.1}$$

For notational convenience, the initial value y_0 is assumed to be zero. The results of this section continue to hold if y_0 has a given distribution. A realization of the process generated by (6.1) is observed for $t = 0, \ldots, T$. The innovations ε_t are assumed to satisfy the following assumption.

Assumption 6.1 *The process* $\{\varepsilon_t\}$ *is strictly stationary and strongly mixing with mixing coefficients* α_m *that satisfy* $\alpha_m = O(m^{-\lambda})$ *for all* $\lambda > p\eta/(p-\eta)$ *for some* $p > \eta > 4$; $E(\varepsilon_t) = 0$ *and* $E|\varepsilon_t|^p < \infty$; $\bar{\sigma}_\varepsilon^2 = \lim_{T\to\infty} E(y_T^2)/T > 0$.

Because of the stationarity requirement, Assumption 6.1 is more restrictive than the assumption in Phillips (1987). However, a considerable amount of heterogeneity and temporal dependence is still tolerated. Quite a few of the processes encountered in the theoretical econometric literature are covered by Assumption 6.1, for example, the class of Gaussian stationary autoregressive moving average processes. The mixing part of Assumption 6.1 restricts the dependence of ε_{t+k} on ε_t and states that the mixing coefficients should be of size $p\eta/(p-\eta)$, White (1984). The moment condition $p > \eta > 4$ can be relaxed to $p > \eta > 2$ if the variance and autocorrelations of ε_t are known. If they are unknown, the stronger moment condition is needed in order to ensure the consistency of the estimators for the variance terms. The assumption of strict stationarity combined with the existence of the moments up to order $p > 2$ implies the weak stationarity of the ε_t process. Finally, the last assumption serves to exclude some degenerate cases.

Using the observed values $y_0, \ldots, y_T$, a regression model is fitted. It is well known that including deterministic regressors, like polynomial time trends, can affect the limiting distribution of the estimators. Therefore, the following vectors of explanatory variables are defined. Let $x_t^r = (1, t, \ldots, t^{r-1}, y_{t-1})^\top$ and $\beta^r = (\gamma_0, \ldots, \gamma_{r-1}, \phi)^\top$ for $r > 0$. If $r = 0$, then $x_t^r = y_{t-1}$ and $\beta^r = \phi$. These definitions are similar to the ones used in Park and Phillips (1988), where only the values $r = 0, 1, 2$ are considered. The regression model is given by

$$y_t = x_t^{r\top}\beta^r + \varepsilon_t. \tag{6.2}$$

The parameter vector β^r of (6.2) is estimated using an M estimator. Let $\hat{\varepsilon}_t = y_t - x_t^{r\top}\hat{\beta}_\psi^r$, then an M estimator is defined as the vector $\hat{\beta}_\psi^r$ that solves the first order condition

$$\sum_{t=1}^{T} \psi(\hat{\varepsilon}_t)x_t^r = 0, \tag{6.3}$$

with $\psi(\cdot)$ a real valued function that satisfies the following assumption (compare Hampel et al. (1986), Huber (1981)).

Assumption 6.2 $\psi(\cdot)$ *is a differentiable function with derivative* $\psi'(\cdot)$; $\psi'(\cdot)$ *is first order Lipschitz; the function* $\psi(\varepsilon)/\varepsilon$ *is bounded;* $E(\psi(\varepsilon_t)) = 0$; $0 < \mu_\psi = E(\psi'(\varepsilon_t)) < \infty$; $\bar{\sigma}_\psi^2 = \lim_{T\to\infty} T^{-1}E\{(\sum_{t=1}^T \psi(\varepsilon_t))^2\} > 0$.

The first two parts of this assumption are rather common smoothness conditions for $\psi(\cdot)$, which are also found in Knight (1989, 1991). The main results are expected to hold even if $\psi(\cdot)$ is allowed to be discontinuous at a finite number of points, as is exemplified by comparing the results below with those of Herce (1993). In that case, however, different methods of proof must be used. This is not pursued here. In the third part of Assumption 6.2, attention is restricted to functions that are at most of the same order of magnitude for large ε as the function $\psi(\varepsilon) = \varepsilon$, which defines the OLS estimator. This causes the focus to be on estimators that are less sensitive to outliers than the OLS estimator (see Hampel et al. (1986) and Huber (1981)). The next two parts of the assumption contain a centering condition for the transformed innovations $\psi(\varepsilon_t)$, and a moment condition for $\psi'(\varepsilon_t)$, respectively. Finally, the last condition of Assumption 6.2 rules out some singularities, as in Assumption 6.1.

In order to test the unit root hypothesis, an estimate of the covariance matrix V_T of $\hat{\beta}^r_\psi$ is needed. Out of the several asymptotically equivalent formulas that are available in the literature, I use the standard one given in Hampel et al. (1986, p. 316),

$$V_T = \left[\sum_{t=1}^T \psi'(\hat{\varepsilon}_t) x_t^r x_t^{r\top}\right]^{-1} \left[\sum_{t=1}^T \psi(\hat{\varepsilon}_t)^2 x_t^r x_t^{r\top}\right] \left[\sum_{t=1}^T \psi'(\hat{\varepsilon}_t) x_t^r x_t^{r\top}\right]^{-1}, \quad (6.4)$$

(compare Chapters 4 and 5). Note that V_T is a heteroskedasticity consistent type of covariance matrix estimator, as in Huber (1981) and White (1980). This can easily be seen by inserting the $\psi(\cdot)$ function of the OLS estimator, $\psi(\hat{\varepsilon}_t) = \hat{\varepsilon}_t$. The robustness of V_T to heteroskedasticity was also noted in Simpson, Ruppert, and Carroll (1992). As already mentioned in Chapter 4, it is not customary to use this type of standard errors for calculating unit root tests. However, their use mitigates some of the finite sample problems caused by the occurrence of outliers in the data (see Chapter 4). This emerges from the simulation experiment presented in Section 6.3. For reasons of comparison, the heteroskedasticity consistent standard errors are used for both the M estimator and the OLS based unit root tests. It is important to note that the OLS based tests now differ from their original versions, compare Fuller (1976), Phillips (1987), and Phillips and Perron (1988). Although this difference vanishes in the limit, it is important in finite samples.

It will prove useful to define the r-dimensional vector $e^r = (0, \ldots, 0, 1)^\top$. Following Fuller (1976), the statistics of interest for testing the unit root hypothesis are $T(e^{r\top}\hat{\beta}^r_\psi - 1)$ and $t_\psi = (e^{r\top}\hat{\beta}^r_\psi - 1)/(e^{r\top}V_T e^r)^{1/2}$. In order to give expressions for the asymptotic distributions of these two test statistics, introduce the bivariate partial sum process $B_T(s) = \sum_{t=1}^{\lfloor sT \rfloor}(\varepsilon_t, \psi(\varepsilon_t))^\top$, with s a real number between zero and one. Using Assumptions 6.1, 6.2, and the results of Phillips and Durlauf (1986), it follows that $B_T(s)/T^{1/2}$ converges weakly to the bivariate Brownian motion $B(s)$, which has covariance matrix

$$\Omega = \begin{pmatrix} \bar{\sigma}_\varepsilon^2 & \bar{\sigma}_{\varepsilon\psi} \\ \bar{\sigma}_{\varepsilon\psi} & \bar{\sigma}_\psi^2 \end{pmatrix} = \lim_{T\to\infty} E[B_T(1)B_T(1)^\top]/T < \infty. \quad (6.5)$$

The two elements of B are denoted by $\bar{\sigma}_\varepsilon b_1$ and $\bar{\sigma}_\psi b_2$, respectively, with b_1 and b_2 two (correlated) standard Brownian motions. Theorem 6.1 now states the asymptotic distribution of the two test statistics.

Theorem 6.1 *Given Assumptions 6.1 and 6.2, regression model (6.2), and the estimator defined in (6.3), then if* $(\hat{\varepsilon}_t - \varepsilon_t) = o_p(1)$ *uniformly for* $t = 1, \ldots, T$,

$$T(e^{r\top}\hat{\beta}^r_\psi - 1) \Rightarrow \frac{\bar{\sigma}_\psi}{\mu_\psi \bar{\sigma}_\varepsilon} e^{r\top} \left(\int x^r x^{r\top} \right)^{-1} \left(\int x^r db_2 + \frac{(\bar{\sigma}_{\varepsilon\psi} - \sigma_{\varepsilon\psi}) e^r}{2\bar{\sigma}_\psi \bar{\sigma}_\varepsilon} \right) \quad (6.6)$$

and

$$t_\psi \Rightarrow \frac{\bar{\sigma}_\psi}{\sigma_\psi} \cdot \frac{e^{r\top} \left(\int x^r x^{r\top} \right)^{-1} \left(\int x^r db_2 + \frac{1}{2}(\bar{\sigma}_{\varepsilon\psi} - \sigma_{\varepsilon\psi}) e^r / \bar{\sigma}_\psi \bar{\sigma}_\varepsilon \right)}{\left[e^{r\top} \left(\int x^r x^{r\top} \right)^{-1} e^r \right]^{1/2}}, \quad (6.7)$$

with $\sigma^2_\psi = E(\psi(\varepsilon_t)^2)$, $\sigma_{\varepsilon\psi} = E(\varepsilon_t \psi(\varepsilon_t))$, $x^r(s) = (1, \ldots, s^{r-1}, b_1(s))^\top$ *for* $r > 0$ *and* $x^r(s) = b_1(s)$ *for* $r = 0$.

The condition $(\hat{\varepsilon}_t - \varepsilon_t) = o_p(1)$ is a consistency requirement as in Knight (1989, 1991) and Chapter 5. It is not needed if $\psi(\cdot)$ is the derivative of a convex function with a unique minimum. In that case, different methods of proof can be used to obtain identical results (compare Davis, Knight, and Liu (1992) and Knight (1989, 1991)). In other cases, the condition requires that a consistent estimate is chosen out of the set of possible solutions to (6.3). This is especially important if $\psi(\varepsilon)$ is zero outside some interval, because then (6.3) has infinitely many solutions (see also Hampel et al. (1986) and Huber (1981)). For the case of i.i.d. innovations, $\bar{\sigma}^2_\psi = \sigma^2_\psi$, $\bar{\sigma}^2_\varepsilon = \sigma^2_\varepsilon$, and $\bar{\sigma}^2_{\varepsilon\psi} = \sigma^2_{\varepsilon\psi}$, such that (6.7) reduces to the result of Theorem 5.1 in Chapter 5.

Using the techniques of Park and Phillips (1988), the results of Theorem 6.1 can also be formulated in terms of detrended Brownian motions. Let the detrended Brownian motion $b^r_1(s)$ be the residual from the continuous time least-squares regression of the model $b_1(s) = \gamma_0 + \ldots + \gamma_{r-1} s^{r-1} + b^r_1(s)$, then

$$T(e^{r\top}\hat{\beta}^r_\psi - 1) \Rightarrow \bar{\sigma}_\psi \left(\mu_\psi \bar{\sigma}_\varepsilon \int (b^r_1)^2 \right)^{-1} \left(\int b^r_1 db_2 + (\bar{\sigma}_{\varepsilon\psi} - \sigma_{\varepsilon\psi}) / (2\bar{\sigma}_\psi \bar{\sigma}_\varepsilon) \right)$$

and

$$t_\psi \Rightarrow \bar{\sigma}_\psi \left(\sigma^2_\psi \int (b^r_1)^2 \right)^{-1/2} \left(\int b^r_1 db_2 + (\bar{\sigma}_{\varepsilon\psi} - \sigma_{\varepsilon\psi}) / (2\bar{\sigma}_\psi \bar{\sigma}_\varepsilon) \right).$$

The limiting distributions in Theorem 6.1 depend upon a number of nuisance parameters. Tests that do not depend upon these quantities can, however, easily be obtained. The nuisance parameters just have to be replaced by consistent estimates, as is done in Phillips (1987) and Phillips and Perron (1988). The most difficult quantity to estimate is Ω, the elements of which appear in (6.6) and (6.7). Following Phillips (1987) and Herce (1993), one can use an estimate of the spectral density evaluated at zero, multiplied by 2π. In this chapter the Parzen window is used to obtain this estimate. Denote the estimated elements of Ω by $\bar{s}_\varepsilon^2$, $\bar{s}_\psi^2$ and $\bar{s}_{\varepsilon\psi}$, respectively. For example,

$$\bar{s}_\varepsilon^2 = T^{-1} \sum_{k=-\ell}^{\ell} w_{k\ell} \hat{\varepsilon}_t \hat{\varepsilon}_{t-k}, \tag{6.8}$$

with ℓ a truncation parameter satisfying both $\ell \to \infty$ and $\ell / T^{1/4} \to 0$ if T diverges to infinity. Moreover,

$$w_{k\ell} = \begin{cases} \left(1 - 6k^2(\ell+1)^{-2}\right)\left(1 - |k|(\ell+1)^{-1}\right) & \text{for } |k| \leq (\ell+1)/2, \\ 2\left(1 - |k|(\ell+1)^{-1}\right)^3 & \text{for } |k| > (\ell+1)/2. \end{cases}$$

So $\bar{s}_\varepsilon^2$ uses a limited number of estimated autocovariances in order to approximate the long run covariance $\bar{\sigma}_\varepsilon^2$. The number of autocovariances that is used, increases with the sample size in order to guarantee consistency. Finally, let

$$\begin{array}{rclrcl} m_\psi & = & T^{-1} \sum_{t=1}^T \psi'(\hat{\varepsilon}_t), & s_\varepsilon^2 & = & T^{-1} \sum_{t=1}^T \hat{\varepsilon}_t^2, \\ s_\psi^2 & = & T^{-1} \sum_{t=1}^T \psi(\hat{\varepsilon}_t)^2, & s_{\varepsilon\psi} & = & T^{-1} \sum_{t=1}^T \hat{\varepsilon}_t \psi(\hat{\varepsilon}_t). \end{array} \tag{6.9}$$

The unit root tests that do not have nuisance parameters in their asymptotic distributions are now given by

$$M_{1,\psi}^r = \frac{\bar{s}_\varepsilon}{\bar{s}_\psi} \left[m_\psi T (e^{r\top} \hat{\beta}_\psi^r - 1) - \tfrac{1}{2}(\bar{s}_{\varepsilon\psi} - s_{\varepsilon\psi}) e^{r\top} \left(T^{-2} \sum_{t=1}^T x_t^r x_t^{r\top} \right)^{-1} e^r \right] \tag{6.10}$$

and

$$M_{2,\psi}^r = \frac{s_\psi}{\bar{s}_\psi} t_\psi - \frac{\bar{s}_{\varepsilon\psi} - s_{\varepsilon\psi}}{2\bar{s}_\psi} \left[e^{r\top} \left(T^{-2} \sum_{t=1}^T x_t^r x_t^{r\top} \right)^{-1} e^r \right]^{1/2}. \tag{6.11}$$

Corollary 6.1 *Given the conditions of Theorem 6.1 and the definitions in (6.8) and (6.9),*

$$M_{1,\psi}^r \Rightarrow e^{r\top} \left(\int x^r x^{r\top} \right)^{-1} \int x^r \, db_2$$

and

$$M_{2,\psi}^r \Rightarrow \frac{e^{r\top} \left(\int x^r x^{r\top} \right)^{-1} \int x^r \, db_2}{\left[e^{r\top} \left(\int x^r x^{r\top} \right)^{-1} e^r \right]^{1/2}}.$$

The critical values of the unit root tests $M_{1,\psi}^r$ and $M_{2,\psi}^r$ can be obtained by means of simulation. This is done in Section 6.4. If $\psi(\varepsilon) = \varepsilon$, the Brownian motion b_2 equals b_1 and the critical values for large T can be found in, for example, Fuller (1976). For small or moderate T, however, these critical values are too small, giving rise to an actual size above the nominal significance level. This is caused by the heteroskedasticity consistent standard errors used in the construction of the test. The correct critical values for several sample sizes can be found in Section 6.4.

Remark 6.1 From the proof of Theorem 6.1 it can be seen that the asymptotic distributions do not change if regression model (6.2) is augmented with regressors of the form $y_{t-k} - y_{t-k-1}$, for $k \geq 1$. Therefore, instead of using the semiparametric corrections suggested in the Corollary, one can make parametric corrections by including additional explanatory variables of the mentioned form. This leads to augmented DF-t type unit root tests, which have a similar asymptotic distribution as the ones mentioned in Theorem 6.1. This prewhitening technique can be related to the estimator for Ω suggested by Andrews and Monahan (1992).

Remark 6.2 By the definition of x_t^r, attention was restricted to time polynomials as additional deterministic regressors. Other deterministic functions of time can be easily dealt with as well. The main condition is that the regressors have to converge in the space of CADLAG functions (see Billingsley (1968)). Therefore, the above theorem can easily be extended to situations with trend breaks at known dates (see, e.g., Perron (1989)).

It is interesting to quantify the change in the asymptotic distributions of the unit root tests based on M estimators and on the OLS estimator. Let $M_{1,\psi}^r$ and $M_{2,\psi}^r$ for $\psi(\varepsilon) = \varepsilon$ be denoted by $M_{1,\varepsilon}^r$ and $M_{2,\varepsilon}^r$, respectively. Then Theorem 6.2 states that a particular linear combination of $M_{2,\psi}^r$ and $M_{2,\varepsilon}^r$ is asymptotically normally distributed. This result can be used in an intuitive way to evaluate the effect of the choice of $\psi(\cdot)$ on the asymptotic distribution of $M_{1,\psi}^r$ relative to that of $M_{1,\varepsilon}^r$.

Theorem 6.2 *Let $\sigma_\varepsilon^2 = E(\varepsilon_t^2)$, $\rho = \bar{\sigma}_{\varepsilon\psi}/\bar{\sigma}_\psi\bar{\sigma}_\varepsilon$ (see (6.5)), and $\hat{\rho} = \bar{s}_{\varepsilon\psi}/\bar{s}_\psi\bar{s}_\varepsilon$ (see (6.8)). Then given the conditions of Theorem 6.1,*

$$M_{1,\psi}^r - \hat{\rho}M_{1,\varepsilon}^r \Rightarrow N(0, 1-\rho^2)\sqrt{e^{r\top}(\textstyle\int x^r x^{r\top})^{-1}e^r}$$

and

$$M_{2,\psi}^r - \hat{\rho}M_{2,\varepsilon}^r \Rightarrow N(0, 1-\rho^2),$$

with $N(m, s^2)$ a normal random variate with mean m and variance s^2.

The parameter ρ gives an indication of the discrepancy between the asymptotic distributions of the unit root tests based on M estimators and on the OLS estimator. It can be interpreted as the long run correlation between the innovations ε_t and their transformation $\psi(\varepsilon_t)$. If $\psi(\cdot)$ is 'close to' the identity function, ρ will be close to unity and the variance of the normal variate of Theorem 6.2 will be approximately zero. This formalizes the intuitive idea that the asymptotic distributions of two unit root tests are approximately the same if their defining $\psi(\cdot)$ functions are close. The necessary closeness criterion is provided by the quantity ρ.

Remark 6.3 Theorem 6.2 can also be used to test the unit root hypothesis directly, as is done in Herce (1993). This results in two other unit root tests, namely $L^r_{1,\psi} = (M^r_{1,\psi} - \hat{\rho} M^r_{1,\varepsilon})/c_\rho$ and $L^r_{2,\psi} = (M^r_{2,\psi} - \hat{\rho} M^r_{2,\varepsilon})/c_\rho$, with $\psi(\varepsilon) \neq \varepsilon$ and $c_\rho = (1 - \hat{\rho}^2)^{1/2}$. Note that the asymptotic distribution of $L^r_{2,\psi}$ does not depend on the order r of the time polynomial included in the regression model.

6.3 The Choice of the M Estimator

M estimators, as defined in (6.3), are not scale invariant. This is an undesirable property. In order to make them scale invariant, the function $\psi(\hat{\varepsilon}_t)$ can be replaced by $\sigma_\varepsilon \psi(\hat{\varepsilon}_t/\sigma_\varepsilon)$. The unknown parameter σ_ε has to be estimated. If the estimator for σ_ε converges in probability to some positive constant, in general σ_ε, then all the results of the previous section still hold (see Davis et al. (1992) and Knight (1989)). Moreover, if a scale equivariant estimator is used to estimate σ_ε, then the modified M estimator becomes scale invariant.

There are at least two different strategies for choosing the functional form of $\psi(\cdot)$. First, following Cox and Llatas (1991), one can choose $\psi(\cdot)$ such that the asymptotic mean squared error of the estimator for nearly nonstationary alternatives is minimized. Given the distribution of the innovations, the optimal $\psi(\cdot)$ in this context is a linear combination of the maximum likelihood estimator and the OLS estimator (compare the results of Chapter 7). Second, one can take a robustness point of view and choose $\psi(\cdot)$ in such a way that the estimator becomes less sensitive to outlying observations. Note that this insensitivity is only important in finite samples. Asymptotically, the effect of the outliers under the null hypothesis is eliminated by the presence of the correction terms in $M^r_{1,\psi}$ and $M^r_{2,\psi}$. Under the alternative hypothesis, a suitable choice for $\psi(\cdot)$ may increase the finite sample power of the tests. This is illustrated in Herce (1993) and in the next section.

This chapter concentrates on the second approach. This choice is motivated by the simulations in Franses and Haldrup (1994) and in the previous two chapters. There it is shown that certain types of outliers have large distortionary effects on the behavior of the traditional unit root tests. In contrast to Chapter 5, the present chapter considers M estimators with a low breakdown point. These estimators are easier to compute than the high breakdown ones and provide at least some protection against discordant observations (compare Chapter 4).

6.4 Simulation Experiment

In this section the level and power properties of the unit root tests presented in Section 6.2 are explored by means of a Monte Carlo simulation experiment. The setup of the experiment closely follows Herce (1993). An extensive tabulation of the results is available in Lucas (1994). Here, only a summary of the experiment and the main conclusions are presented.

The first choice concerns the specification of $\psi(\cdot)$. For illustrative purposes, I consider the OLS estimator, given by $\psi(\varepsilon) = \varepsilon$, and the maximum likelihood estimator for Student t distributed observations (MLT), given by $\psi(\varepsilon) = (c + 1)\varepsilon/(c + \varepsilon^2)$. The degrees of freedom parameter c is kept fixed and equal to 3. The sensitivity of this estimator to outliers increases with the parameter c. I also consider the Huber ψ function, which is often used in the robustness literature (see Hampel et al. (1986), Huber (1981)). It is given by $\psi(\varepsilon) = \min(c, \max(-c, \varepsilon))$, with $c = 1.345$. The value of c is such that the estimator has a relative efficiency of 95% with respect to the mean if one estimates a location parameter for a set of i.i.d. Gaussian disturbances (compare, e.g., Hampel et al. (1986), Yohai (1987)). The use of the MLT estimator can be defended by noting that it has a bounded influence function in the stationary time series context (see Chapter 4 for details). Moreover, the use of the Student t distribution as a way of relaxing the normality assumption is fairly common in econometrics. In all cases I used the scale invariant version of the M estimator, as described in Section 6.3. I chose the median absolute deviation divided by 0.6745 for estimating the scale. This estimator is a consistent estimator for the standard deviation of Gaussian innovations.

Of course, different choices for $\psi(\cdot)$ may give rise to a different behavior of the unit root tests. However, the specifications provided above appear to perform reasonably well in a variety of circumstances. In particular, they have better size and power properties in finite samples than the traditional OLS based tests. This is revealed in the simulations described below.

The distribution of the test statistics $M^r_{1,\psi}$, $M^r_{2,\psi}$, $L^r_{1,\psi}$ and $L^r_{2,\psi}$ can be simulated in the traditional way, see Fuller (1976) and Herce (1993). First, one generates a time series according to (6.1) with i.i.d. standard Gaussian errors ε_t. Next, one calculates the test statistics for $r = 0, 1, 2$ and the chosen specification for $\psi(\cdot)$. The values of the statistics can be stored and the process repeated a large, say N, number of times. I used $N = 10,000$ and $T = 100, 200$ and 5,000. Using the N simulated values, the α-quantile of the distribution of the unit root tests is estimated by the αNth order statistic. The results are presented in Table 6.1. The distributions of the tests do not vary much over the chosen specifications of $\psi(\cdot)$. This can be expected, because according to Theorem 6.2, $M^r_{2,\psi} - \hat{\rho} M^r_{1,\varepsilon} \Rightarrow N(0, 1 - \rho^2)$. It is easily checked by numerical integration that for the present choices of $\psi(\cdot)$, ρ is greater than 0.94. As a result, when applying the test one can use the critical values for the OLS simulations in Table 6.1, even if $\psi(\varepsilon) \neq \varepsilon$. The only condition is that ρ should not depart too much from unity.

TABLE 6.1
Quantiles for Unit Root Tests

test	$\psi(\cdot)$	$T=100$		$T=200$		$T=5000$	
		0.010	0.050	0.010	0.050	0.010	0.050
	OLS	-13.42	-8.09	-13.73	-7.94	-13.59	-8.15
M_1^0	HUB	-13.32	-8.09	-13.54	-7.82	-13.27	-7.93
	t(3)	-13.04	-7.76	-13.10	-7.65	-13.07	-7.84
	OLS	-2.73	-2.04	-2.63	-1.96	-2.57	-1.96
M_2^0	HUB	-2.79	-2.02	-2.67	-1.95	-2.53	-1.94
	t(3)	-2.83	-2.01	-2.69	-1.97	-2.53	-1.95
L_1^0	HUB	-6.85	-3.79	-6.68	-3.82	-6.53	-3.77
	t(3)	-6.78	-3.80	-6.62	-3.83	-6.46	-3.84
L_2^0	HUB	-2.51	-1.76	-2.37	-1.71	-2.33	-1.68
	t(3)	-2.48	-1.74	-2.41	-1.73	-2.39	-1.63
	OLS	-19.95	-13.98	-19.58	-13.78	-20.05	-13.94
M_1^1	HUB	-19.89	-13.81	-19.40	-13.56	-19.48	-13.60
	t(3)	-19.61	-13.47	-19.11	-13.40	-19.34	-13.40
	OLS	-3.82	-3.06	-3.54	-2.94	-3.39	-2.85
M_2^1	HUB	-3.90	-3.10	-3.61	-2.95	-3.39	-2.82
	t(3)	-3.96	-3.08	-3.66	-2.93	-3.37	-2.80
L_1^1	HUB	-9.62	-5.67	-8.77	-5.46	-8.46	-5.23
	t(3)	-9.83	-5.73	-8.88	-5.48	-8.15	-5.33
L_2^1	HUB	-2.83	-1.91	-2.61	-1.78	-2.32	-1.65
	t(3)	-2.87	-1.89	-2.58	-1.78	-2.31	-1.64
	OLS	-27.40	-20.67	-27.93	-20.85	-28.46	-21.61
M_1^2	HUB	-27.40	-20.56	-27.75	-20.63	-28.14	-21.13
	t(3)	-26.80	-20.39	-27.37	-20.39	-27.28	-20.68
	OLS	-4.40	-3.66	-4.17	-3.53	-3.95	-3.40
M_2^2	HUB	-4.42	-3.68	-4.20	-3.54	-3.93	-3.37
	t(3)	-4.47	-3.66	-4.18	-3.51	-3.88	-3.34
L_1^2	HUB	-13.32	-8.23	-12.63	-7.79	-10.89	-7.33
	t(3)	-13.40	-8.37	-12.02	-7.80	-11.10	-7.36
L_2^2	HUB	-3.14	-2.02	-2.84	-1.92	-2.29	-1.64
	t(3)	-3.23	-2.06	-2.80	-1.87	-2.33	-1.65

The table contains the 5% and 1% critical values of several unit root tests. The entries are based on 10,000 Monte Carlo simulations. The tests $M_{1,\psi}^r$ and $M_{2,\psi}^r$ are presented just above Corollary 6.1, with some necessary additional definitions found in (6.8) and (6.9). The tests $L_{1,\psi}^r$ and $L_{2,\psi}^r$ are presented in Remark 6.3, with $\hat{\rho}$ as defined in Theorem 6.2. OLS means that the unit root test is based on the OLS estimator. Similarly, HUB and t(3) mean that the test is based on the Huber M estimator and on a Student t likelihood with three degrees of freedom, respectively (see also the second paragraph of Section 6.4).

The difference between the values for M_2^r in Table 6.1 and those in Fuller (1976) is caused by the heteroskedasticity consistent standard errors, used for calculating the tests. Although the effect vanishes in the limit, it appears to be important in finite samples.

The power simulations are set up as follows. The data generating process is $y_t = \phi y_{t-1} + \varepsilon_t$, with ϕ =0.90, 0.95, 0.99 and 1.00. I use two values for the sample size T, namely 100 and 200. In order to study the sensitivity of the results, I use three distributions for the errors: the standard normal ($N(0,1)$), the double exponential (DExp) and a truncated Cauchy (TCauchy). The double exponential has exponentially decreasing tails that are fatter than those of the normal distribution. The truncated Cauchy is a badly behaved distribution in the sense that the probability of obtaining large drawings is relatively high. In order to ensure the existence of moments, I truncate the original Cauchy distribution to the interval $[-c, c]$, with $c = 12.7$. In this way 95% of the probability mass of the original Cauchy distribution is captured.

The sensitivity of the tests with respect to the correlation structure of the errors is investigated by considering the additive outlier model, as in Chapter 5. This is a kind of measurement error model, and therefore the use of autocorrelation consistent test statistics seems advisable. The time series y_t is observed with error as $y_t + u_t$. The measurement error u_t is i.i.d. and equals 0 with probability 0.95. Otherwise, it equals three times the drawing from one of the distributions presented earlier.[1]

In this chapter I only consider the regression models with trend ($r = 2$). Results for the other regression models can be found in Lucas (1994). Using the 5% simulated critical values for the different tests and $\psi(\cdot)$ functions, the rejection frequencies of the unit root tests were stored for the different values of the autoregressive parameter ϕ using $1,000$ replications. The truncation parameter needed for estimating the elements of Ω was 10 for $T = 100$ and 15 for $T = 200$. Making use of additional information about the correlation structure of the errors for choosing a smaller truncation parameter did not substantially change the results. The rejection frequencies for i.i.d. errors and for the additive outlier model are given in Tables 6.2 and 6.3, respectively. Because the $L_{1,\psi}^r$ and $L_{2,\psi}^r$ tests appeared to have lower power than $M_{1,\psi}^r$ and $M_{2,\psi}^r$, they are omitted from the discussion.

For the case of i.i.d. Gaussian errors, the performance of all tests is comparable. The $M_{1,\psi}^2$ have an estimated size that is significantly above the nominal level of five per cent. As the critical values of Table 6.1 are used, this discrepancy appears to be due to the specific set of simulated time series underlying the entries in Table 6.2. For fat-tailed errrors, Table 6.2 reveals that power can be gained by using the Huber or MLT estimator. This power gain is clearest for TCauchy errors. This is due to the fact that robust M estimators fully exploit the advantages of having occasional large errors, i.e., innovative outliers (see Davis et al. (1992)). Also note that the estimated sizes of several of the tests are significantly below the nominal five per cent level for fat-tailed errors. As

[1] For the truncated Cauchy distribution, the multiplication by the factor 3 is omitted.

TABLE 6.2
Rejection Frequencies of Several Unit Root Tests Based on M Estimators; the Case of i.i.d. Innovations

		$T = 100$				$T = 200$			
cdf	$\phi=$	0.90	0.95	0.99	1.00	0.90	0.95	0.99	1.00
$N(0,1)$	$M^2_{1,OLS}$	27.3	13.6	7.2	8.0	78.9	30.0	9.2	9.2
	$M^2_{1,HUB}$	27.9	13.5	7.4	9.0	78.7	31.3	9.9	9.0
	$M^2_{1,MLT}$	28.3	14.1	6.9	9.5	77.3	31.6	10.0	9.0
	$M^2_{2,OLS}$	19.9	9.6	5.8	6.7	66.0	20.4	6.9	7.6
	$M^2_{2,HUB}$	20.3	11.0	6.5	6.3	59.2	20.3	7.6	6.9
	$M^2_{2,MLT}$	20.5	11.7	7.7	7.2	54.8	19.2	6.9	7.3
DExp	$M^2_{1,OLS}$	28.8	13.4	5.8	5.7	81.8	34.2	7.3	7.0
	$M^2_{1,HUB}$	33.6	13.5	4.1	4.2	90.5	38.9	6.2	4.6
	$M^2_{1,MLT}$	36.8	14.5	3.9	4.1	92.1	40.2	5.9	4.0
	$M^2_{2,OLS}$	20.9	10.2	3.8	3.5	68.7	24.3	5.3	4.8
	$M^2_{2,HUB}$	30.0	12.9	3.5	3.1	77.9	33.8	4.7	2.9
	$M^2_{2,MLT}$	31.7	12.3	3.8	2.9	80.2	37.5	5.1	2.8
TCauchy	$M^2_{1,OLS}$	17.4	8.2	5.8	5.0	62.4	22.1	6.5	7.2
	$M^2_{1,HUB}$	45.5	12.8	2.6	1.5	89.8	50.7	3.3	1.9
	$M^2_{1,MLT}$	54.9	21.8	4.1	2.6	90.4	61.6	5.3	2.1
	$M^2_{2,OLS}$	17.5	6.2	2.1	1.5	56.2	15.8	3.8	4.5
	$M^2_{2,HUB}$	52.3	19.9	3.9	2.5	90.4	58.5	5.6	2.2
	$M^2_{2,MLT}$	61.2	32.1	5.9	3.8	92.4	68.5	10.3	3.2

The Table presents the rejection frequencies of unit root tests based on M estimators. The model for generating the data is $y_t = \phi y_{t-1} + \varepsilon_t$, with ε_t a set of i.i.d. disturbances from either the standard normal ($N(0,1)$), the double exponential (DExp), or the truncated Cauchy (TCauchy) distribution. For each entry 1,000 Monte Carlo replications are used. Standard errors of the entries range from 0.4 to at most 1.6. The nominal level of the tests is 5%.

the power properties of these tests are generally quite good, I do not consider it to be much of a problem that the tests are somewhat conservative under the null hypothesis.

For the additive outlier case in Table 6.3, the actual sizes of all tests typically seem to exceed the nominal level, especially if one considers the $M^2_{1,\psi}$ tests. This is in line with the simulations in Chapter 5. The effects found here are, however, less dramatic. As mentioned before, the present OLS based tests are not exactly the ones suggested by Phillips (1987) and Phillips and Perron (1988), but rather heteroskedasticity consistent variants of these. The heteroskedasticity consistent estimates of the standard errors seem to mitigate the size problems of the $M^2_{2,\psi}$ tests. Still, the sizes of the Huber and MLT based tests are closer to the nominal level than the sizes of the OLS based tests. Also note that the power of all tests has dropped relative to the setting with i.i.d. errors.

Other simulations using different correlation structures for the errors were also performed. These seemed to confirm the conclusion that the performance of the tests is comparable in the Gaussian case, slightly favoring the use of OLS

TABLE 6.3
Rejection Frequencies of Several Unit Root Tests Based on M Estimators; the Case of Additive Outliers

		T = 100				T = 200			
cdf	ϕ=	0.90	0.95	0.99	1.00	0.90	0.95	0.99	1.00
$N(0,1)$	$M^2_{1,OLS}$	68.3	42.0	27.8	25.4	97.4	70.8	33.5	26.3
	$M^2_{1,HUB}$	57.2	31.6	20.2	18.0	92.4	54.6	20.6	15.8
	$M^2_{1,MLT}$	53.0	29.7	17.4	16.6	88.1	48.6	17.7	14.0
	$M^2_{2,OLS}$	32.2	17.1	9.9	8.3	76.1	34.0	13.0	10.8
	$M^2_{2,HUB}$	22.6	11.5	7.5	6.9	68.1	28.8	9.5	7.6
	$M^2_{2,MLT}$	23.1	12.5	9.1	7.9	64.6	28.2	10.2	8.1
DExp	$M^2_{1,OLS}$	77.8	55.3	43.1	40.4	98.4	79.9	45.5	44.1
	$M^2_{1,HUB}$	64.1	36.0	24.6	24.2	93.0	58.7	22.6	21.0
	$M^2_{1,MLT}$	53.8	30.5	17.7	19.4	88.9	49.6	17.8	16.2
	$M^2_{2,OLS}$	33.8	19.7	12.5	11.7	73.8	36.4	13.7	12.1
	$M^2_{2,HUB}$	16.9	8.7	5.9	5.7	62.0	23.8	7.6	5.4
	$M^2_{2,MLT}$	18.6	11.3	7.7	7.0	63.9	27.0	9.0	7.2
TCauchy	$M^2_{1,OLS}$	47.0	34.3	26.8	27.6	89.9	67.5	39.3	32.2
	$M^2_{1,HUB}$	40.2	22.9	15.9	17.0	79.8	48.9	23.1	18.7
	$M^2_{1,MLT}$	37.7	22.4	15.0	14.7	74.6	44.0	20.6	17.0
	$M^2_{2,OLS}$	26.0	15.2	11.6	10.4	67.8	39.9	18.2	15.2
	$M^2_{2,HUB}$	19.9	11.8	8.7	8.0	58.7	28.8	14.3	10.9
	$M^2_{2,MLT}$	22.6	14.1	9.7	9.4	55.9	30.3	15.3	11.5

The Table presents the rejection frequencies of unit root tests based on M estimators. The model for generating the data is $y_t = x_t + u_t$, with $x_t = \phi x_{t-1} + \eta_t$, η_t a set of i.i.d. standard normal disturbances, and u_t a set of i.i.d. disturbances. u_t is equal to zero with probability 0.95, and otherwise it is a drawing from three times the standard normal ($N(0,1)$), three times the double exponential (DExp), or from the truncated Cauchy (TCauchy) distribution. For each entry 1,000 Monte Carlo replications are used. Standard errors of the entries range from 0.4 to at most 1.6. The nominal level of the tests is 5%.

based tests. If the errors are fat-tailed, the use of the Huber or MLT estimator seems preferable. If the errors ε_t in (6.1) have an autoregressive structure, the power of the tests can drop to the nominal size, especially if $r = 2$. This finding holds for all the M estimators considered in the simulations. Alternatively, if the errors have a moving average structure and are negatively correlated, all the tests have severe size problems. This makes them virtually useless for practical situations.

The overall conclusion from these simulations is that the behavior of the tests based on the presented M estimators is similar to that of the OLS based tests. Substantial power gains can be made by using the non-OLS estimators if the errors are fat-tailed. Also note that the computation time of M estimators is approximately the same as that of OLS. This is intuitively clear by writing (6.3) as an (iterative) weighted least-squares problem, as in Hampel et al. (1986) and Huber (1981). Therefore, it seems worthwhile to perform one of the outlier robust unit root tests presented in this paper, instead of or along with the traditional OLS based tests.

6.5 Concluding remarks

In this chapter the asymptotic distribution theory for unit root tests based on M estimators was discussed. The asymptotic distribution theory turned out to be similar to the theory for the traditional OLS based tests as developed in Phillips (1987) and Phillips and Perron (1988). As one would expect, the asymptotic distributions of a unit root tests based on an M and on the OLS estimator are approximately the same if the M estimator is close to the OLS estimator. A natural closeness criterion was povided by the long run correlation between the innovations and their transformation $\psi(\varepsilon_t)$, where $\psi(\cdot)$ defines the M estimator. It was also shown that a linear combination of an outlier robust test and an OLS based test is asymptotically normally distributed. This result was used to construct new unit root tests, as in Herce (1993).

A simulation experiment was provided, illustrating the level and power properties of several tests. Unit root tests based on a linear combination of an OLS based unit root test and an outlier robust test appeared to have low power. For the more traditional unit root tests, however, power could be gained by using outlier robust estimators if the errors were fat-tailed. This power gain had to be paid in terms of a power loss for Gaussian errors. In many settings, however, the power gain in nonnormal situations outweighted the loss in the Gaussian case. As the computation time of M estimators is similar to that of OLS, it seems worthwhile to use these robust tests instead of, or at least along with, traditional OLS based tests.

A second finding that emerged from the simulation study and that was also noted in Chapter 4 is that the use of heteroskedasticity consistent standard error estimates, as in White (1980), mitigates the size problems of the Dickey-Fuller t-test in additive outlier situations. Therefore, the use of these standard errors seems an easy way of repairing some of the problems of the OLS based tests.

6.A Proofs

In order to prove the theorems of Section 6.2, I first introduce some additional notation. Let D^r be a diagonal matrix with $\text{diag}(D^r) = (T^{-0.5}, \ldots, T^{-r+0.5}, (\bar{\sigma}_\varepsilon T)^{-1})$ for $r \geq 1$, and $D^r = (\bar{\sigma}_\varepsilon T)^{-1}$ for $r = 0$. Furthermore, let $\mathcal{F}_t^\varepsilon$ be the sigma algebra generated by $\varepsilon_t, \varepsilon_{t-1}, \ldots$ and let $E_t(\cdot)$ be the conditional expectation with respect to $\mathcal{F}_t^\varepsilon$. Convergence in probability is denoted by $\xrightarrow{p}$. The remaining notation is the same as in Section 6.2. The parameter r is assumed to be nonnegative. Lemma 6.1 facilitates the proofs of the theorems.

Lemma 6.1 *Given Assumptions 6.1 and 6.2,*

(a) $\sum_{t=1}^T D^r x_t^r \psi(\varepsilon_t)^2 x_t^{r\top} D^r \Rightarrow \sigma_\psi^2 \int x^r x^{r\top}$

(b) $\sum_{t=1}^{T} D^r x_t^r \psi'(\varepsilon_t) x_t^{r\top} D^r \Rightarrow \mu_\psi \int x^r x^{r\top}$

(c) $\sum_{t=1}^{T} D^r x_t^r \psi(\varepsilon_t) \Rightarrow \bar{\sigma}_\psi \int x^r db_2 + (\bar{\sigma}_{\varepsilon\psi} - \sigma_{\varepsilon\psi}) e^r / 2\bar{\sigma}_\varepsilon$

Joint convergence of (a) to (c) also applies.

Proof. To prove (a), note that under the present assumptions

$$\lim_{n\to\infty} \sup_t E|E_{t-n}(\psi(\varepsilon_t)^2) - E(\psi(\varepsilon_t)^2)| = 0$$

(see, e.g., Doob (1960), Ibragimov and Linnik (1971)). Hence, Theorem 3.3 of Hansen (1992) can be applied in order to obtain

$$\sup_{0\le s\le 1} \left| \sum_{t=1}^{\lfloor sT \rfloor} D^r x_t^r x_t^{r\top} D^r (\psi(\varepsilon_t)^2 - \sigma_\psi^2) \right| \xrightarrow{p} 0.$$

Part (a) now follows from the fact that $\sum_{t=1}^{T} D^r x_t^r x_t^{r\top} D^r \Rightarrow \int x^r x^{r\top}$, along the lines of Phillips (1987) and Phillips and Durlauf (1986). Part (b) is proved similarly.

Part (c) is proved by using the martingale difference approximation technique described in Hansen (1992). Define $u_t = \sum_{i=0}^{\infty} E_t\{\psi(\varepsilon_{t+i})\}$, then by applying Theorem 3.1 of Hansen (1992), it only remains to be proved that

$$\textstyle\sum_{t=1}^{T} D^r (x_t^r - x_{t-1}^r) u_t - D^r x_T^r u_{T+1} \xrightarrow{p} (\bar{\sigma}_{\varepsilon\psi} - \sigma_{\varepsilon\psi}) e^r / 2\bar{\sigma}_\varepsilon. \tag{6.12}$$

As the function $\psi(\cdot)$ is continuous, by Theorem 3.49 of White (1984) the process $\{\psi(\varepsilon_t)\}$ is strong mixing with mixing coefficients of the same size as the ones in Assumption 6.1. Therefore, using the proof of Theorem 3.1 of Hansen (1992), $\sup_{t\le T} |u_t|/T^{0.5} \xrightarrow{p} 0$. As $T^{0.5} D^r x_t^r = O_p(1)$, we obtain that $D^r x_T^r u_{T+1} \xrightarrow{p} 0$. Let a be the first term on the left-hand side of (6.12). Using the proof of Theorem 4.1 in Hansen (1992), it is easily shown that the $(r+1)$th element of a converges in probability to $\sum_{i=1}^{\infty} E(\varepsilon_t \psi(\varepsilon_{t+i}))/\bar{\sigma}_\varepsilon$, which equals $(\bar{\sigma}_{\varepsilon\psi} - \sigma_{\varepsilon\psi})/2\bar{\sigma}_\varepsilon$. The remaining elements of a converge to 0 in probability by applying the corollary to Theorem 3.3 in Hansen (1992). This establishes part (c) of the lemma.

Joint convergence follows straightforwardly by stacking the (vectorized) left-hand side of (a) through (c) into a single vector. □

Proof of Theorem 6.1. First, a Taylor series expansion is taken of (6.3) with respect to $\hat{\varepsilon}_t$ around the true innovation ε_t. We obtain

$$0 = \sum_{t=1}^{T} \psi(\varepsilon_t) x_t^r - \sum_{t=1}^{T} \psi'(\varepsilon_t) x_t^r x_t^{r\top} (\hat{\beta}_\psi^r - e^r) + R_T, \tag{6.13}$$

because $\varepsilon_t = y_t - y_{t-1} = y_t - x_t^{r\top} e^r$. Premultiplying (6.13) by D^r and using the Lipschitz condition for $\psi'(\cdot)$, we get

$$|R_T| \le K \textstyle\sum_{t=1}^{T} |(\hat{\varepsilon}_t - \varepsilon_t) D^r x_t^r x_t^{r\top} D^r (D^r)^{-1} (\hat{\beta}_\psi^r - e^r)|,$$

for some constant K. The absolute value of a vector is taken elementwise. The condition $(\hat{\varepsilon}_t - \varepsilon_t) = o_p(1)$ uniformly for $t = 1, \ldots, T$ now implies that R_T can be replaced by $o_p(1)(D^r)^{-1}(\hat{\beta}_\psi^r - e^r)$. Therefore, (6.13) can be rewritten as

$$(D^r)^{-1}(\hat{\beta}_\psi^r - e^r) = \left[\sum_{t=1}^{T} D^r x_t^r \psi'(\varepsilon_t) x_t^{r\top} D^r + o_p(1) \right]^{-1} \left[\sum_{t=1}^{T} D^r x_t^r \psi(\varepsilon_t) \right].$$

Using Lemma 6.1, it follows that

$$\begin{aligned} T(e^{r\top}\hat{\beta}^r_\psi - 1) &= e^{r\top}(D^r)^{-1}(\hat{\beta}^r_\psi - e^r)/\bar{\sigma}_\varepsilon \\ &\Rightarrow \frac{\bar{\sigma}_\psi}{\bar{\sigma}_\varepsilon \mu_\psi} e^{r\top}(\textstyle\int x^r x^{r\top})^{-1}(\textstyle\int x^r db_2 + (\bar{\sigma}_{\varepsilon\psi} - \sigma_{\varepsilon\psi})/2\bar{\sigma}_\varepsilon\bar{\sigma}_\psi). \end{aligned}$$

For proving the second part of the theorem, it only has to be shown that $V_T \Rightarrow \sigma^2_\psi \bar{\sigma}_\varepsilon^{-2}\mu_\psi^{-2}(\int x^r x^{r\top})^{-1}$. Notice that under the present conditions

$$\textstyle\sum_{t=1}^T (\psi(\hat{\varepsilon}_t)^2 - \psi(\varepsilon_t)^2) D^r x^r_t x^{r\top}_t D^r \overset{p}{\to} 0.$$

The theorem now follows directly by applying Lemma 6.1. □

Proof of Corollary 6.1. Consistency of $\bar{s}_\varepsilon$, $\bar{s}_\psi$, and $\bar{s}_{\varepsilon\psi}$ follows from, e.g., Newey and West (1987), Phillips (1987) and White (1984), while that of m_ψ, s_ψ, and $s_{\varepsilon\psi}$ follows from the corollary in Hansen (1992). Corollary 6.1 now follows from the consistency of these estimators and Theorem 6.1. □

Proof of Theorem 6.2. First note that $\hat{\rho}$ is a consistent estimator of ρ. Therefore, we only have to prove that $M^r_{1,\psi} - \rho M^r_{1,\varepsilon}$ and $M^r_{2,\psi} - \rho M^r_{2,\varepsilon}$ converge weakly to the random variates mentioned in the theorem. I only present the proof for the first relation. The second one follows in a similar way. Applying Corollary 6.1, we obtain

$$M^r_{1,\psi} - \rho M^r_{1,\varepsilon} \Rightarrow e^{r\top}\left(\textstyle\int x^r x^{r\top}\right)^{-1}\left(\textstyle\int x^r d(b_2 - \rho b_1)\right).$$

Note that the bivariate Brownian motion (b_1, b_2) has covariance matrix

$$\tilde{\Omega} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

If we define the matrix

$$A = \begin{pmatrix} 1 & 0 \\ -\rho & 1 \end{pmatrix},$$

then the bivariate Brownian motion $(b_1, b_3)^\top = A(b_1, b_2)^\top$ has covariance matrix

$$A\tilde{\Omega}A^\top = \begin{pmatrix} 1 & 0 \\ 0 & 1-\rho^2 \end{pmatrix}.$$

Therefore, the elements of x^r are uncorrelated with b_3. Hence, conditional on x^r, $\int x^r db_3$ is a normally distributed random vector with mean zero and variance covariance matrix $\int x^r x^{r\top}$. This proves the theorem. □