Journal of Econometrics 6 (1977) 199–224. © North-Holland Publishing Company

ESTIMATION OF A NON-INVERTIBLE MOVING AVERAGE PROCESS

The Case of Overdifferencing*

Charles I. PLOSSER

Graduate School of Business, Stanford University, Stanford, CA 94305, USA

G. William SCHWERT

Graduate School of Management, University of Rochester, Rochester, NY 14627, USA

Received February 1977, final version received March 1977

The effect of differencing all of the variables in a properly specified regression equation is examined. Excessive use of the difference transformation induces a non-invertible moving average (MA) process in the disturbances of the transformed regression. Monte Carlo techniques are used to examine the effects of overdifferencing on the efficiency of regression parameter estimates, inferences based on these estimates, and tests for overdifferencing based on the estimator of the MA parameter for the disturbances of the differences regression. Overall, the problem of overdifferencing is not serious if careful attention is paid to the properties of the disturbances of regression equations.

1. Introduction

Recently Granger and Newbold (1974) have illustrated the potential for observing spurious correlations between time series variables which have common deterministic or stochastic trends. Yule (1926) made this point a half century ago, but it has not been fully incorporated into applied econometric work. One suggestion for avoiding this difficulty is to take the differences of the variables until each has a constant unconditional mean over the sample period, and then estimate the correlation or regression relationship between the transformed variables. This is proposed as a strategy for data analysis by Box and Jenkins (1970), Granger and Newbold (1974), and others. While it is true that differencing all of the variables in a regression function does not generally affect the

*We would like to acknowledge valuable comments from John Abowd, Mukhtar Ali, Kenneth Gaver, Martin Geisel, Charles Nelson, David Pierce, Harry Roberts, Christopher Sims, William Wecker, and Arnold Zellner, although we are responsible for remaining errors. Plosser's participation in this research was partially supported by National Science Foundation Grant SOC 7305547 and the H.G.B. Alexander Foundation at the University of Chicago. An earlier version of this paper was presented before the Econometric Society in September 1976 at Atlantic City, New Jersey.

values of the regression coefficients,[1] the properties of the disturbance term for the regression equation are affected by the transformation. In particular, it is possible to induce a non-invertible moving average process in the disturbances for the regression equation between the transformed variables by differencing the regressand and the regressors prior to estimating the regression parameters. In this paper we examine the problems associated with estimating such a non-invertible moving average process, and the implications of these problems for the model-building strategy of Box, Jenkins, Granger, and Newbold.

A time series $\{\tilde{z}_t\}$ is said to have a representation as a first-order moving average process [MA(1)] if it can be written as

$$\begin{aligned}\tilde{z}_t &= \tilde{a}_t - \theta_1 \tilde{a}_{t-1} \\ &= (1-\theta_1 L)\tilde{a}_t, \end{aligned} \tag{1.1}$$

where $\{\tilde{a}_t\}$ is a sequence of independent identically distributed (i.i.d.) random variables with mean zero and constant variance σ^2, θ_1 is the constant moving average parameter, and L is the lag operator: $L^k \tilde{x}_t \equiv \tilde{x}_{t-k}$. If $|\theta_1| < 1$ in eq. (1.1), the process is said to be *invertible*; that is, $\tilde{z}_t$ has an autoregressive representation,

$$\tilde{z}_t = \sum_{i=1}^{\infty} \Pi_i \tilde{z}_{t-i} + \tilde{a}_t. \tag{1.2}$$

Since the autocorrelation structure of the MA(1) process is unchanged if θ_1 is replaced by $1/\theta_1$ in eq. (1.1), the admissible range of θ_1 is usually restricted to the region $|\theta_1| \leqq 1$ in order to identify estimates of the MA parameter. Although a number of authors have studied the properties of different estimators of θ_1 for invertible MA processes, little attention has been directed towards properties of estimators of θ_1 in the case where $|\theta_1| = 1$.

In this paper we analyze the *strictly non-invertible* moving average process which occurs when $|\theta_1| = 1$, so neither θ_1 nor $1/\theta_1$ yield a convergent autoregressive representation for $\{\tilde{z}_t\}$. In section 2 we show that this case could occur frequently in the analysis of autoregressive-integrated-moving average (ARIMA) models or time series regression models as a result of excessive use of the difference transformation: $\Delta \tilde{x}_t \equiv \tilde{x}_t - \tilde{x}_{t-1}$. In section 3 previous work on the large and small sample properties of estimators of invertible MA parameters is reviewed. We emphasize the considerations which are likely to be problematic in the strictly noninvertible case. In section 4 we present Monte

[1] This is true if every variable in the regression equation is differenced the same number of times, and the exogenous variables are not polynomials in time. When we say that all variables are differenced until each has a constant unconditional mean over the same period, we assume that all variables are differenced the same number of times.

Carlo experiments for two plausible situations where a strictly non-invertible moving average process might occur as a result of excessive differencing. First, we consider the case

$$\tilde{z}_t = \alpha+\beta t+\tilde{a}_t, \tag{1.3}$$

so that $\{\tilde{z}_t\}$ deviates randomly around a linear trend. If $\tilde{z}_t$ is differenced in order to create a transformed series with a constant unconditional mean and variance, the changes in $\tilde{z}_t$ follow a non-invertible MA(1) process,

$$\Delta\tilde{z}_t = \beta+\tilde{a}_t-\theta_1\tilde{a}_{t-1}, \qquad \theta_1 = 1. \tag{1.4}$$

Second, we consider the case

$$\tilde{y}_t = \alpha+\beta\tilde{x}_t+\tilde{a}_t, \tag{1.5a}$$

and

$$\tilde{x}_t = \tilde{x}_{t-1}+\tilde{u}_t, \tag{1.5b}$$

where $\{\tilde{a}_t\}$ and $\{\tilde{u}_t\}$ are independent sequences of i.i.d. random variables, so $\tilde{x}_t$ follows a random walk and $\tilde{y}_t$ is linearly related to $\tilde{x}_t$ with a stationary disturbance. If both $\{\tilde{y}_t\}$ and $\{\tilde{x}_t\}$ are differenced, so each has a constant unconditional mean and variance, the regression equation between the differenced variables is

$$\Delta\tilde{y}_t = \alpha'+\beta\Delta\tilde{x}_t+(1-\theta_1 L)\tilde{a}_t, \qquad \theta_1 = 1. \tag{1.6}$$

The disturbance in eq. (1.6) follows a non-invertible MA(1) process. The Monte Carlo experiments concentrate on three issues: (a) the sampling distribution of an estimator of θ_1; (b) the properties of estimated standard errors of $\hat{\theta}_1$ and the test statistics associated with the hypothesis $\theta_1 = 1$; and (c) the sampling distribution of the estimators of the regression parameter β when $\theta_1 = 1$ in eqs. (1.4) or (1.6).

Finally, in the last section we summarize the findings of the Monte Carlo experiments and relate them to the problem facing the data analyst of whether or not to use the difference transformation. We conclude that overdifferencing does not create serious problems for estimating linear regression models, although estimates of the moving average parameter are likely to be biased toward zero if $|\theta_1| = 1$. Thus, the strategy of Box and Jenkins (1970) or Granger and Newbold (1974) should not lead to serious errors in the event that differencing is not necessary to induce stationarity in the disturbance for the regression relationship. Our sampling experiments should provide applied econometricians with a basis for deciding whether overdifferencing has occurred.

2. Differencing and non-invertibility

2.1. Univariate time series models

The difference transformation, $\Delta\tilde{x}_t \equiv \tilde{x}_t - \tilde{x}_{t-1}$, has been advocated as a technique for eliminating deterministic or stochastic trends from time series data so that the transformed variable has a constant unconditional mean.[2] Tintner's (1940, 1955, 1962) 'variate difference' method uses the difference transformation to remove a polynomial time trend from the time series,

$$\tilde{z}_t = f(t) + \tilde{a}_t, \tag{2.1a}$$

$$\Delta^d\tilde{z}_t = (1-L)^d\tilde{z}_t = \alpha + \tilde{a}'_t, \tag{2.1b}$$

by analyzing the sample variance of successive differences of the time series. In principle $f(t)$ could be a general deterministic function of time which might be approximated by a dth order polynomial equation.[3]

Other analysts have advocated the use of the difference transformation to eliminate stochastic trends in time series data. Wold (1938), Yaglom (1955), Whittle (1963, pp. 92–96), Quenouille (1968, pp. 50–57), and Box and Jenkins (1970, pp. 85–125), among others, suggest the use of the difference transformation to eliminate unitary roots in the autoregressive polynomial for an autoregressive-integrated-moving average (ARIMA) process. For example, suppose a time series has a representation as an ARIMA process,

$$\phi_p(L)\tilde{z}_t = \theta_q(L)\tilde{a}_t, \tag{2.2}$$

where $\phi_p(L)$ is the pth order AR polynomial in the lag operator and $\theta_q(L)$ is the qth order MA polynomial in the lag operator, or

$$\tilde{z}_t = (\theta_q(L)/\phi_p(L))\tilde{a}_t, \tag{2.3}$$

where $\tilde{a}_t$ is a serially independent random variable with a stationary distribution over time. The dth differences of $\tilde{z}_t$ can be represented as

$$(1-L)^d\tilde{z}_t = ((1-L)^d\theta_q(L)/\phi_p(L))\tilde{a}_t. \tag{2.4}$$

[2] It may be necessary to use a logarithmic or power transformation [cf. Box and Cox (1964)], or some other transformation prior to differencing in order to induce a constant unconditional variance for the transformed variable. We say a random variable is stationary if it has a constant unconditional mean and variance, although we are primarily concerned with mean nonstationarity in this paper.

[3] Kendall and Stuart (1968, pp. 384–392) discuss the variate difference method.

Now if there are d unitary roots in the AR polynomial,

$$\phi_p(L) = \prod_{j=1}^{p} (1-\Psi_j L) = (1-L)^d \prod_{j=1}^{p-d} (1-\Psi_j L),$$

the difference factor would cancel from the numerator and the denominator of the right-hand side of (2.4). Thus, the dth differences would have a representation as an ARMA process,

$$(1-L)^d \tilde{z}_t = (\theta_q(L)/\phi'_{p-d}(L))\tilde{a}_t, \tag{2.5}$$

where the AR polynomial, $\phi'_{p-d}(L)$, is of order $(p-d)$ and presumably all of the remaining $p-d$ roots lie outside the unit circle.

A simple example of this type of situation where differencing eliminates linear homogeneous non-stationarity is the random walk model,

$$\tilde{z}_t = \tilde{z}_{t-1} + \tilde{a}_t \tag{2.6a}$$

$$= \tilde{a}_t/(1-L) = \sum_{i=0}^{\infty} \tilde{a}_{t-i}. \tag{2.6b}$$

In this case the first difference of $\tilde{z}_t$ is just the stationary random varia)le $\tilde{a}_t$, but the level of the process $\tilde{z}_t$ does not have a finite unconditional mean or variance. The random walk model exhibits 'stochastic trends' due to the accumulation of random shocks $\tilde{a}_t$, even though $\tilde{z}_t$ does not contain a deterministic function of time.

Pierce (1975) discusses the similarity of the sample autocorrelation functions for realizations from processes which have deterministic trends like (2.1) with those from processes which have unitary roots in the autoregressive polynomial like (2.4). He considers methods for simultaneously identifying deterministic time trends and ARIMA models for the errors associated with the time trend model.

If a univariate time series is differenced in order to remove a deterministic trend as in (2.1b), the resulting disturbances $\tilde{a}'_t$ will have a moving average polynomial with a unitary root. Given the similarity of the sample realizations from models with deterministic and stochastic trends, data analysts could frequently difference variables to induce stationarity while inadvertently inducing a non-invertible MA process for the disturbances of the transformed series.

A similar problem can occur if a time series variable exhibits pronounced seasonal behavior due to different means for the periods within the year. Seasonal differencing the series creates a transformed series with a constant mean, but it also induces a non-invertible moving average process at the seasonal lag. For example, suppose $z_{t,s}$ represents a realization from a stochastic process in the month s of year t and the process is defined as

$$\tilde{z}_{t,s} = E(\tilde{z}) + E(\tilde{z}_s) + \tilde{\varepsilon}_{t,s}, \qquad s = 1, \dots, 12, \quad t = 1, \dots, T, \tag{2.7}$$

where $E(\tilde{z}_s)$ represents the deviation of the mean in month s from the overall mean $E(\tilde{z})$, and $\tilde{\varepsilon}_{t,s}$ is an i.i.d. random variable whose distribution does not depend on t or s. The seasonal difference of $\{\tilde{z}_{t,s}\}$ is

$$(1-L^{12})\tilde{z}_{t,s} = \tilde{z}_{t,s} - \tilde{z}_{t-1,s} = \tilde{\varepsilon}_{t,s} - \tilde{\varepsilon}_{t-1,s}, \tag{2.8}$$

which is a strictly non-invertible first-order seasonal MA process.

2.2. *Regression models*

It is well known that serially correlated disturbances can seriously affect the distributional properties of t or F statistics which are computed under the erroneous assumption that the disturbances are serially independent.[4] In particular, if the regression disturbance follows an ARMA process with one or more unitary roots in the AR polynomial it would be necessary to difference all of the variables in the regression equation one or more times in order to eliminate the non-stationarity of the disturbances. In general, there is no theoretical reason to presume that the regression disturbance would be stationary for the levels of the variables, so a data analyst might consider estimating the regression function between the differences as well as the levels of the variables.

The solution tentatively suggested by Granger and Newbold (1974, p. 118), following the lead of Box and Jenkins (1970, p. 378), is to difference all of the variables in the regression model until each has a constant unconditional mean before trying to estimate the coefficients of the model. However, this strategy could lead to a non-invertible MA process in the regression disturbances for the transformed variables if the non-stationary behavior in the regressand $\tilde{y}_t$ is a result of the relationship with a non-stationary regressor $\tilde{x}_t$. An example of this case is given in eqs. (1.5) and (1.6) where $\tilde{x}_t$ follows a random walk, so $\tilde{y}_t$ has a representation as an IMA(1, 1) process [the first differences of $\tilde{y}_t$ follow an invertible MA(1) process],[5]

$$\Delta\tilde{y}_t = \beta\Delta\tilde{x}_t + \Delta a_t \tag{2.9a}$$

$$= \beta\tilde{u}_t + \tilde{a}_t - \tilde{a}_{t-1}. \tag{2.9b}$$

[4] Vinod (1976) illustrates the effects of serially correlated disturbances on the critical values for t and F tests by computing upper and lower bounds for the critical values of the test statistics.

[5] Box and Jenkins (1970, pp. 121–125) illustrate a case similar to this one.

Thus, both $\tilde{y}_t$ and $\tilde{x}_t$ are non-stationary processes and would be incorrectly differenced prior to estimating the regression parameters following the Box–Jenkins strategy. The subsequent Monte Carlo experiments examine the losses which are associated with estimating the regression parameters from the differenced data [as in eq. (1.6)], rather than the correctly specified regression between the levels of the variables [eq. (1.4)].

Thus, contrary to the assertion of Whittle (1963, p. 43) that a strictly non-invertible MA process would only rarely occur in practical situations 'because the fact that a variable can only be observed with a limited accuracy...', the use of the difference transformation can lead to strictly non-invertible moving average processes in several plausible cases. In the subsequent sections, we analyze the problem of estimating the MA parameter in this case and we examine the implications of 'overdifferencing' on the estimators of time trend or regression parameters.

3. Estimation of MA parameters

3.1. The likelihood function for the MA(1) process

Suppose a time series $\{\tilde{z}_t\}$ has a representation as an MA(1) process,

$$\tilde{z}_t = \tilde{a}_t - \theta_1 \tilde{a}_{t-1} = (1-\theta_1 L)\tilde{a}_t, \qquad t = 1, \ldots, T, \tag{3.1}$$

where $\{\tilde{a}_t\}$ is a sequence of i.i.d. Normal random variables with mean zero and constant variance, σ_a^2. Through recursive calculation equation (3.1) can be expressed in terms of the past values of the series,

$$\tilde{z}_t = -\theta_1 \tilde{z}_{t-1} - \theta_1^2 \tilde{z}_{t-2} - \ldots - \theta_1^{t-1} \tilde{z}_1 + \tilde{a}_t - \theta_1^t \tilde{a}_0, \tag{3.2}$$

where $\tilde{a}_0$ is the unobservable disturbance which occurred prior to observing the sample. Eq. (3.2) highlights the importance of invertibility for estimating the MA parameter θ_1 since the last term $\theta_1^t \tilde{a}_0$ has diminishing effect on $\tilde{z}_t$ as t gets large as long as $|\theta_1| < 1$. Regardless of invertibility, the likelihood function for the MA(1) process in eq. (3.1) is

$$L(\theta_1, \sigma_a^2 | z) = (2\Pi\sigma_a^2)^{-T/2} |K'K|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_a^2} \; z' M_T z\right\}, \tag{3.3}$$

where

$$M_T = \begin{bmatrix} (1+\theta_1^2) & -\theta_1 & 0 & \cdot\ \cdot\ \cdot & 0 \\ -\theta_1 & (1+\theta_1^2) & -\theta_1 & \cdot & \cdot \\ & & \cdot & \cdot & \cdot\ \cdot \\ 0 & -\theta_1 & \cdot & \cdot & 0 \\ \cdot\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & -\theta_1 \\ \cdot & \cdot & \cdot & \cdot & \\ 0\cdot\ \cdot\ \cdot\ \cdot 0 & & -\theta_1 & (1+\theta_1^2) & \end{bmatrix}^{-1}, \tag{3.4a}$$

and

$$|K'K| = \sum_{j=0}^{T} \theta_1^{2j}. \tag{3.4b}$$

Note that the sum of squared residuals computed recursively as a function of θ_1 is the expression in the exponent,

$$S(\theta_1) = \sum_{t=0}^{T} \tilde{a}_t(\theta_1)^2 = z'M_Tz. \tag{3.5}$$

Shaman (1969) derives an expression for the typical element m_{ij} of M_T which does not require invertibility,

$$m_{ij} = \theta_1^{j-i}\frac{\{1+\theta_1^2+\ldots+\theta_1^{2(i-1)}\}\{1+\theta_1^2+\ldots+\theta_1^{2(T-j)}\}}{\{1+\theta_1^2+\ldots+\theta_1^{2T}\}}, \quad \text{for} \quad i \leqq j. \tag{3.6}$$

If $|\theta_1| < 1$, m_{ij} can be written as

$$m_{ij} = \theta_1^{j-i}\frac{(1-\theta_1^{2i})(1-\theta_1^{2(T+1-j)})}{(1-\theta_1^2)(1-\theta_1^{2(T+1)})}, \quad \text{for} \quad i \leqq j, \tag{3.7}$$

and the determinant (3.4b) becomes

$$|K'K| = \frac{(1-\theta_1^{2(T+1)})}{(1-\theta_1^2)}. \tag{3.8}$$

Thus, the likelihood function is a very complicated function of the MA parameter θ_1, especially if $|\theta_1| \geqq 1$.

Although the only restriction which must be imposed on eq. (3.1) to make the process stationary is that θ_1 be finite, the autocorrelation function for the MA(1) process,

$$\begin{aligned}\rho_k &= \frac{-\theta_1}{1+\theta_1^2}, \quad k = 1,\\ &= 0, \qquad\quad k > 1,\end{aligned} \tag{3.9}$$

illustrates a fundamental identification problem for MA processes [cf. Quenouille (1968, p. 79)]. If θ_1^* is a solution to (3.9), so is $1/\theta_1^*$, which means that for every non-invertible process there is an observationally equivalent invertible process, except for the case where $|\theta_1| = 1$. This consideration has led Box–Jenkins (1970, A7.6, p. 284) and Osborn (1976) to suggest that the likelihood function should be maximized over the invertibility region of the parameter space in order to identify estimators of MA parameters.[6] Such a restriction of the admissible parameter space has the additional advantage of avoiding computational problems involved in estimating the initial condition $\tilde{a}_0$, and it reduces the importance of the initial condition $\tilde{a}_0$ as seen in eq. (3.2). However, the important case of $|\theta_1| = 1$ is ignored by restricting attention to invertible MA processes, so we suggest restricting the admissible parameter space to $|\theta_1| \leqq 1$ for MA(1) processes.

There are several methods for estimating θ_1 based on approximations to the likelihood function (3.3) and most of them differ because of their treatment of the initial condition $\tilde{a}_0$. Box and Jenkins (1970, pp. 215–220) suggest a method of 'backcasting' the presample value $\tilde{a}_0$ based on the stochastic structure of the sample. They argue that for large samples and an invertible MA process, maximization of the likelihood function (3.3) is essentially equivalent to minimizing the sum of squares function $S(\theta_1)$, because the determinant (3.4b) approaches a constant for large samples.

Unfortunately, these arguments are not true when $|\theta_1| = 1$. First, the determinant (3.4b) does not approach a constant as the sample size gets larger, so it can't be ignored in maximizing (3.3). Second, the backcasting technique can no longer effectively be used to estimate $\tilde{a}_0$, since the effect of the assumption about initial conditions (at the end of the sample) will never be dissipated when the backward process is analyzed [cf. Osborn (1976, p. 82)].

Alternatively, a_0 could be added to the parameter set and estimated jointly with the moving average parameter [cf. Osborn (1976, pp. 76–78)]. While this approach does not require invertibility, rounding errors may become important if $|\hat{\theta}_1|$ is larger than unity.

[6] In higher order moving average processes an equivalent assumption is that all of the roots of the moving average polynomial, $\theta(L)$, lie outside the unit circle.

Finally, the approach which is used most frequently in practice is to set the initial condition equal to its marginal expectation of zero, $a_0 = 0$. This 'conditional' maximum likelihood estimator for θ_1 merely requires finding the value of $\hat{\theta}_1$ which minimizes the sum-of-squares function, $S(\theta_1)$. All of the preceding estimators are easily operationalized using standard nonlinear least squares procedures [cf. Chambers (1973)].

A full maximum likelihood estimator of θ_1 would involve evaluating both the sum-of-squares function $S(\theta_1)$ and the determinant in (3.4b). Osborn (1976) and Ali (1976) discuss some of the difficulties involved in such a procedure beyond those which occur with the estimators which only minimize some form of the sum-of-squares function.

3.2. *Distributional properties of MA parameter estimators*

Since most of the estimators of θ_1 discussed above differ from a full maximum likelihood estimator because of their treatment of the initial condition, and since the initial condition has diminishing influence on $\tilde{z}_t$ as t gets large for invertible processes (since $\theta_1^t \tilde{a}_0$ approaches zero), all of the estimators above share the large sample distributional properties of the likelihood estimator. In particular, $T^{\frac{1}{2}}(\hat{\theta}_1 - \theta_1)$ has an asymptotic Normal distribution with mean zero and variance $(1-\theta_1^2)$, and the estimator $\hat{\theta}_1$ is asymptotically independent of the assumption about the initial condition [cf. Pierce (1971, pp. 301, 304–305)].

However, these large sample properties do not extend to the case of strict non-invertibility. First, the asymptotic properties of the different estimators discussed above are not identical, in general, since the effect of the initial condition on $\tilde{z}_t$ does not diminish as t gets large when $|\theta_1| = 1$. Second, it is easy to see that the large sample distribution of $g(T)\cdot(\hat{\theta}_1 - \theta_1)$ could not have a variance proportional to $(1-\theta_1^2)$ if $|\theta_1| = 1$. This problem occurs in part because we are trying to estimate a parameter which is on the boundary of the admissible parameter space.[7] Third, any estimator of θ_1 will never be asymptotically independent of the initial condition $\tilde{a}_0$, so the assumptions about the initial condition become much more important if $|\theta_1| = 1$.

Nelson (1974), Kang (1975), and Osborn (1976) have compared the small sample properties of different estimators of θ_1 for invertible MA processes ($|\theta_1| < 1$). There is no consistent ranking of the various least squares estimators and the maximum likelihood estimator for small sample sizes (20–30 observations), and almost all differences among these estimators disappear as the sample size increases to 100 observations. Of course, if $|\theta_1|$ is close to one the differences among estimators diminish slowly as the sample size increases. Thus, we would expect that the behavior of the various estimators would be different for all

[7] Chernoff (1954) discusses a case which has some similarities to this problem; however, the fact that $|\theta_1| = 1$ is the boundary of the admissible (restricted) parameter space creates problems which go beyond Chernoff's example.

sample sizes if $|\theta_1| = 1$ because of the non-trivial effect of the assumption about the initial condition $\tilde{a}_0$.

Since the conditional least squares estimator, where $\tilde{a}_0$ is assumed equal to zero, is the easiest to implement, and because the other estimators can involve severe numerical problems in trying to estimate $\tilde{a}_0$ when $|\theta_1| = 1$, we concentrate on the conditional least squares estimator in our Monte Carlo experiments. While this does not allow us to compare the distributional properties of the various estimators, we feel that the results of these experiments should be indicative of the results that would be obtained from other techniques because: (a) setting $\tilde{a}_0$ equal to its marginal expectation of zero is no more (or less) arbitrary than the other techniques for estimating the initial condition in this case, and (b) all of the estimators suffer from the problem that we are trying to estimate a parameter on the boundary of the admissible parameter space. In addition, since conditional least squares is the technique most widely used in applied work the small sample properties of this estimator in the strictly non-invertible case are intrinsically interesting in their own right.

3.3. *Inference about MA parameters*

If an estimate of a MA parameter is close to strict non-invertibility in an applied situation a logical question would be whether $|\hat{\theta}_1|$ is significantly different from 1. Since the large sample distribution of $\hat{\theta}_1$ is not known if $|\theta_1| = 1$, it is not clear what the appropriate test statistic would be for this situation. Nevertheless, in our Monte Carlo experiments we examine the properties of the 't-ratio', $t \equiv (\hat{\theta}_1 - 1)/\text{s.e.}(\hat{\theta}_1)$, where the standard error of $\hat{\theta}_1$, s.e.$(\hat{\theta}_1)$, is estimated from a linearization of the sum-of-squares function by analogy with linear regression techniques [cf. Nelson (1974, pp. 127–128) or Box and Jenkins (1970, pp. 226–228)].

The t-ratios are not likely to conform to either a standard Normal or Student-t distribution because $|\hat{\theta}_1|$ is almost always less than one (since the admissible range of $\hat{\theta}_1$ is restricted); therefore, the t-ratio is likely to have a negative mean and may be negatively skewed. In addition, Nelson (1974, p. 132) has noted that estimated standard errors, s.e.$(\hat{\theta}_1)$, are often substantially smaller than the standard deviation of the sampling distribution of $\hat{\theta}_1$ for invertible MA parameters. Thus, since the t-ratio cannot be positive and the denominator s.e.$(\hat{\theta}_1)$ may severely understate the sampling variability of $\hat{\theta}_1$, we expect a relatively large number of negative t-ratios which are large in absolute value.

In addition to the t-ratio, we examine the estimate of the standard error for the differences regression, $\hat{\sigma}(a_d)$, relative to the standard error for the levels regression (1.3), $\hat{\sigma}(a_L)$. If $\hat{\theta}_1 = 1$ in eq. (1.4) the disturbances from the MA process should be the same as the disturbances from the correctly specified time trend model in eq. (1.3), while if $\hat{\theta}_1 < 1$, $\hat{\sigma}(a_d)$ should be greater than

$\hat{\sigma}(a_L)$. Thus, the ratio $\hat{\sigma}(a_d)/\hat{\sigma}(a_L)$ should provide a measure of predictive efficiency which can be used to compare the differenced and undifferenced forms of the model.[8]

4. Monte Carlo experiments

4.1. Linear time trend example

In order to analyze the effects of differencing on deterministic time trends we analyze the models

$$\tilde{z}_t = \alpha + \beta(t/T) + \tilde{a}_t, \qquad t = 1, \ldots, T, \tag{4.1a}$$

$$\Delta\tilde{z}_t = \beta(1/T) + (1-\theta_1 L)\tilde{a}_t, \qquad t = 2, \ldots, T, \tag{4.1b}$$

for samples of 51, 101 and 201 observations. The samples are created by generating sequences of i.i.d. Normal random variates $\{\tilde{a}_t\}$ with mean zero and unit variance.[9] A new sample is generated for each of 1000 replications and diagnostic checks on the serial independence and Normality of $\{\tilde{a}_t\}$ do not reveal any irregularities in the artificial data. The time trend variables are normalized so that $E(\tilde{z}_0) = \alpha$ and $E(\tilde{z}_T) = \alpha+\beta$, which is analogous to increasing the sample size T by observing the process (1.3) at more frequent intervals. For these experiments $\alpha = \beta = 1$, the time trend model (4.1a) is estimated using linear least squares, and the moving average model (4.1b) is estimated using a modified Gauss–Newton iterative procedure[10] which constrains the MA parameter so that $|\hat{\theta}_1| \leqq 1$.

Table 1 contains summary statistics for estimates of the MA parameter θ_1 in eq. (4.1b) for 1000 replications of samples of 50, 100 and 200 observations. Because the range of the estimator is restricted, the estimator is downward biased, although the magnitude of the average bias decreases as the sample size increases. The sampling distributions are negatively skewed. Interestingly, the median of the sampling distributions is approximately 0.95 for all sample sizes, so the major effect of increased sample size is to reduce the number of low values of $\hat{\theta}_1$. There is no indication that the negative skewness or positive kurtosis diminishes as the sample size increases.

[8] The ratio of the estimated residual standard deviations cannot be used to construct a likelihood ratio test in this case since (a) the levels regression model has one parameter, α, and one observation that the differences regression model does not have, and (b) the hypothesized value, $\theta_1 = 1$, is on the boundary of the admissible parameter space.

[9] These pseudo-random Normal deviates are generated by Marsaglia's rectangular-wedge-tail method, incorporated in program 'RANORM', obtained from the University of Chicago Computation Center. Kinderman and Ramage (1976) discuss some attractive properties of generators like this one.

[10] Draper and Smith (1966, pp. 267–270) discuss the Gauss–Newton method.

The average standard errors of $\hat{\theta}_1$, s.e.$(\hat{\theta}_1)$, are substantially smaller than the standard deviations of the sampling distributions for all sample sizes. As a result of the bias in $\hat{\theta}_1$ and the bias in s.e.$(\hat{\theta}_1)$, the sampling distribution of the t-ratio is negatively biased and the number of values of this statistic less than -2.0 is very large. In fact, for a sample size of 200 the mean of the t-ratios is -1.9 and almost half of the t-ratios are less than -2.0. Thus, a 't-test' of whether $\theta_1 = 1$ based on estimates of $\hat{\theta}_1$ and s.e.$(\hat{\theta}_1)$ could lead to very misleading conclusions if a Student-t or Normal distribution is used naively to determine an appropriate critical value for such a test. We provide the 0.05 and 0.01 fractiles of the sampling distribution of the t-ratio in our experiments to provide a basis for such tests.

The average ratio of the standard error of the differenced model (4.1b) to the standard error of the time trend model (4.1a), $\hat{\sigma}(a_d)/\hat{\sigma}(a_L)$, indicates that the MA model has a residual standard deviation about six percent greater than the correctly specified time trend model for a sample size of 50. This difference diminishes to four and three percent for samples of 100 and 200 observations, respectively, reflecting the fact that $\hat{\theta}_1$ is closer to unity on average as the sample size increases. The fractiles of the sampling distribution of this ratio indicate that the loss in prediction efficiency from incorrectly choosing the differenced model as the correct specification would not be great for moderate sample sizes.

Table 2 contains summary statistics for estimators of the time trend coefficient β in eq. (4.1). The true value is $\beta = 1$, so the process $\{\tilde{z}_t\}$ has a mean which increases over time, while the changes $\Delta\tilde{z}_t$ have a constant mean equal to β/T. Both estimators are unbiased and their sampling distributions conform well to a Normal (or Student-t) distribution. There is very little skewness or excess kurtosis, and the studentized range statistics[11] are consistent with the hypothesis of Normality, except for the differences regression with 200 observations which has a large studentized range statistic. However, the standard deviation of the sampling distribution, which is a measure of the efficiency of the estimators of β, is substantially lower for the levels regression (4.1a) than for the differences regression (4.1b). This is because the sampling variance of $\hat{\beta}$ for the levels is inversely proportional to $\sum_{t=1}^{T} t^2$, while the sampling variance of $\hat{\beta}$ for the differences is inversely proportional to $T-1$. Thus, ignoring problems of estimating θ_1, the variance of the sampling distribution of $\hat{\beta}$ decreases with T^3 for the levels and with T for the differences.[12]

The average value of the computed standard error for $\hat{\beta}$ is reported in parentheses below the mean values of $\hat{\beta}$. The estimates of the standard error of $\hat{\beta}$ for the differences regression, the MA process, are generally larger than the

[11] David, Hartley, and Pearson (1954) derive the sampling distribution for this statistic if the sample is drawn from a Normal population.

[12] Jacobs (1976) notes that a linear time trend model where the time variable is measured with a stationary error can be estimated consistently using least squares. This occurs because t^2 becomes large relative to the variance of the measurement error as $T \rightarrow \infty$.

Table 1

Estimator of the moving average parameter, θ_1.[a]

$$\Delta z_t = \hat{\beta}(1/T+1)+(1-\hat{\theta}_1 L)\hat{a}_t.$$

(1) *Sample moments*	$T = 50$	$T = 100$	$T = 200$
Mean[b]	0.9205	0.9353	0.9508
(Avg. s.e.)	(0.0751)	(0.0412)	(0.0233)
Standard deviation	0.0885	0.0610	0.0422
Skewness	−1.143	−0.9656	−1.041
Excess kurtosis	0.7009	0.8484	1.340
Selected fractiles: 0.50	0.9527	0.9425	0.9566
0.20	0.8449	0.8857	0.9200
0.10	0.7843	0.8517	0.8948
0.05	0.7481	0.8229	0.8727
0.01	0.6554	0.7487	0.8104
(2) *t-ratios* $(\hat{\theta}_1-1)/\text{s.e.}(\hat{\theta}_1)$			
Mean	−0.9947	−1.406	−1.858
Standard deviation	0.9697	1.108	1.224
Percent $t < -2.0$	21.0	36.8	47.7
Fractiles: 0.05	−2.638	−3.106	−3.765
0.01	−3.196	−3.819	−4.695
(3) *Ratio of regression standard errors*[c] $\hat{\sigma}(a_d)/\hat{\sigma}(a_L)$			
Mean	1.063	1.044	1.031
Standard deviation	0.0433	0.0299	0.0209
Fractiles: 0.50	1.055	1.040	1.027
0.95	1.144	1.100	1.069
0.99	1.187	1.126	1.092

[a] 1000 replications of samples of size 50, 100 and 200 for eq. (4.1b).

[b] Average value of $\hat{\theta}_1$; average standard error, s.e. $(\hat{\theta}_1)$, in parentheses.

[c] Ratio of the estimate of the residual standard deviation in eq. (4.1b) to the estimate of the residual standard deviation in eq. (4.1a).

standard deviations of the sampling distributions. This is opposite from the relationship which was observed for the estimate of the standard error of the MA parameter in table 1.

The t-ratios associated with $\hat{\beta}$, $t = (\hat{\beta}-1)/\text{s.e.}(\hat{\beta})$, are also reported. The t-ratios for the differences regression are slightly fat-tailed for a sample size of 50, since over eight percent of the t-ratios are greater than two in absolute value. However, this problem disappears for larger sample sizes.

Thus, even though the estimator of the MA parameter is downward biased, the estimator of the time trend parameter β is unbiased and conforms well to a Normal distribution. The levels estimator is more efficient than the differences estimator because the sample variability of the regressor is greater in the levels.[13]

[13] In the context of testing hypotheses about β, the probability of Type I error is similar for the levels and the differences, but the power of a test is greater in the levels than the differences.

Table 2

Estimators of the time trend parameter, β.[a]

$$z_t = \hat{\alpha}+\hat{\beta}(t/T)+\hat{a}_t, \qquad \Delta z_t = \hat{\beta}(1/T)+(1-\hat{\theta}_1 L)\hat{a}_t.$$

	Levels regression	Differences regression	Levels regression	Differences regression	Levels regression	Differences regression
(1) *Sample moments*	$T = 51$		$T = 101$		$T = 201$	
Mean	1.013	1.012	1.005	1.043	0.9967	1.018
(Avg. s.e.)	(0.4830)	(0.9254)	(0.3439)	(0.8827)	(0.2442)	(0.8524)
Standard deviation	0.4947	0.8235	0.3450	0.7213	0.2427	0.6416
Skewness	−0.0161	0.0280	0.0707	−0.2438	−0.1068	−0.0207
Excess kurtosis	0.1604	−0.4938	−0.1347	−0.2003	0.0929	0.6483
Studentized range[b]	6.567	5.830	6.103	6.323	6.606	9.556[c]
(2) *t-ratios* $(\hat{\beta}-1)/\text{s.e.}(\hat{\beta})$						
Mean	0.0265	0.0258	0.0134	0.0813	−0.0147	0.0405
Standard deviation	1.043	1.219	1.007	1.109	0.9942	0.9478
Percent $t > 2.0$	2.6	4.5	2.4	3.7	2.0	2.3
$t < -2.0$	3.3	4.0	2.1	3.2	3.4	2.0

[a] 1000 replications of samples of size 51, 101 and 201 for eq. (4.1a) with $\beta = 1$. The differences regression (4.1b) has samples of size 50, 100 and 200.

[b] Studentized range = $[\max(\hat{\beta})-\min(\hat{\beta})]$/standard deviation of $\hat{\beta}$.

[c] Exceeds 0.99 fractile of the sampling distribution when sampling from a Normal population.

4.2. Stochastic regression model example

Our second example where use of the difference transformation could induce a strictly non-invertible moving average process is

$$\tilde{y}_t = \alpha+\beta\tilde{x}_t+\tilde{a}_t, \qquad t = 1,\dots,T, \tag{4.2a}$$

$$\tilde{x}_t = \tilde{x}_{t-1}+\tilde{u}_t$$

$$\Delta\tilde{y}_t = \alpha'+\beta\Delta\tilde{x}_t+(1-\theta_1 L)\tilde{a}_t, \quad t = 2,\dots,T. \tag{4.2b}$$

where $\{\tilde{a}_t\}$ and $\{\tilde{u}_t\}$ are independent sequences of i.i.d. Normal deviates with mean zero and unit variance. For the first set of experiments with eq. (4.2), one vector of x is constructed and held constant across replications, while the regression disturbances $\{\tilde{a}_t\}$ are the same as those used in the time trend example reported above. Thus, we are analyzing the conditional regression model given $\tilde{x} = x$. For the second set of experiments with eq. (4.2), a new regressor vector is generated for each replication, so we are analyzing the stochastic (unconditional) regression model in this case. In all experiments $\alpha = \beta = 1$, and $\alpha' = 0$.

Table 3 contains summary statistics for estimates of the MA parameter θ_1 in eq. (4.2b) for 1000 replications of samples of 50, 100 and 200 observations. As in table 1, the estimator $\hat{\theta}_1$ is downward biased, although the bias decreases as the sample size increases. The sampling distributions are negatively skewed with the median being about 0.95 for all sample sizes. The only apparent difference between the fixed regressor case and the stochastic regressor case is the larger mean and median for the stochastic regressor model with a sample size of 50.

The average standard errors of $\hat{\theta}_1$ understate the sampling variability of $\hat{\theta}_1$ for all sample sizes, with the possible exception of the stochastic regression model for a sample size of 50. As in table 1, the bias and skewness in $\hat{\theta}_1$ and the bias in s.e.$(\hat{\theta}_1)$ cause the sampling distribution of the t-ratio to be negatively biased with too many extreme negative values. The 0.05 and 0.01 fractiles of the sampling distributions for the t-ratio are very close to the values in table 1, which suggests that these might be used as critical values for tests of hypotheses about θ_1 no matter what the form of the regression function.

The ratio of the standard error of the differences regression to the standard error of the levels regression behaves almost identically to the results reported in table 1. Thus, the difference between $\hat{\sigma}(a_d)$ and $\hat{\sigma}(a_L)$ seems to be attributable to the fact that $\hat{\theta}_1$ is less than unity, and does not seem to be a function of the behavior of the regressor.

Table 3

Estimator of moving average parameter θ_1 in regression model (4.2b).[a]

$$\Delta y_t = \hat{\alpha} + \hat{\beta}\Delta x_t + (1 - \hat{\theta}_1 L)\hat{a}_t.$$

	Fixed regressor	Stochastic regressor	Fixed regressor	Stochastic regressor	Fixed regressor	Stochastic regressor
(1) *Sample moments*	$T = 50$		$T = 100$		$T = 200$	
Mean[b]	0.9220	0.9300	0.9403	0.9391	0.9515	0.9544
(Avg. s.e.)	(0.0795)	(0.0869)	(0.0462)	(0.0452)	(0.0248)	(0.0250)
Standard deviation	0.0864	0.0896	0.0632	0.0652	0.0426	0.0419
Skewness	−1.184	−1.422	−1.067	−1.224	−1.097	−0.8969
Excess kurtosis	0.9054	1.481	0.7687	1.452	1.409	0.4481
Selected fractiles: 0.50	0.9540	0.9816	0.9569	0.9572	0.9575	0.9613
0.20	0.8488	0.8522	0.8861	0.8886	0.9186	0.9168
0.10	0.7890	0.7932	0.8513	0.8443	0.8934	0.8976
0.05	0.7532	0.7426	0.8199	0.8126	0.8721	0.8782
0.01	0.6532	0.6547	0.7425	0.7378	0.8110	0.8320
(2) *t-ratios* $(\hat{\theta}_1 - 1)/\text{s.e.}(\hat{\theta}_1)$						
Mean	−0.9427	−0.8010	−1.184	−1.233	−1.765	−1.644
Standard deviation	0.9238	0.9359	1.106	1.128	1.209	1.250
Percent $t < -2.0$	18.1	15.8	27.2	27.8	47.2	44.2
Fractiles: 0.05	−2.577	−2.586	−3.047	−3.170	−3.725	−3.604
0.01	−3.105	−3.019	−3.777	−3.841	−4.692	−4.242
(3) *Ratio of regression standard errors*[c] $\hat{\sigma}(a_d)/\hat{\sigma}(a_L)$						
Mean	1.065	1.066	1.044	1.045	1.031	1.030
Standard deviation	0.0441	0.0461	0.0312	0.0316	0.0213	0.0211
Fractiles: 0.50	1.056	1.056	1.039	1.040	1.028	1.027
0.95	1.149	1.147	1.102	1.104	1.070	1.069
0.99	1.194	1.199	1.130	1.131	1.094	1.089

[a] 1000 replications of samples of size 50, 100 and 200 for eq. (4.2b). In fixed regressor case Δx_t is held constant across replications, while in stochastic regressor case a new sample of Δx_t is generated for each replication.

[b] Average value of $\hat{\theta}_1$; average standard error, s.e. $(\hat{\theta}_1)$, in parentheses.

[c] Ratio of the estimate of the residual standard deviation for the differences regression (4.2b) to the estimate of the residual standard deviation for the levels regression (4.2a).

Table 4 contains summary statistics for estimators of the regression coefficient β in eq. (4.2). The differences regression is estimated in two ways: (a) the regression parameters are estimated jointly with the MA disturbance process using nonlinear least squares, and (b) the regression parameters are estimated using linear least squares ignoring the autocorrelation in the disturbances. The latter case is included to illustrate the benefits available from estimating the MA parameter. All of the estimators are unbiased and seem to have Normal sampling distributions.

In this example the average standard errors of the regression coefficient are relatively close to the standard deviations of the sampling distributions for all three estimators.[14] An exception is the differences/MA regression model with 200 observations which has an average estimated standard error approximately ten percent greater than the standard deviation of the sampling distribution. The levels regression estimator is the most efficient since its sampling variance is inversely proportional to $\sum_{t=1}^{T} \tilde{x}_t^2$, which has an expected value of $(T+1)\cdot T\cdot \sigma_a^2/2$, conditional on x_0. The expected value of $\sum_{t=2}^{T} \Delta\tilde{x}_t^2$ is just $(T-1)\cdot\sigma_a^2$, so the differences regressions are less efficient. In addition, the variance of the disturbances for the differences regression which ignores the MA parameter is $(1+\theta_1^2)\sigma_a^2 = 2\sigma_a^2$, so estimating the MA parameter has a substantial effect on the efficiency of the regression coefficient estimator for the differences regression.[15]

The t-ratios associated with $\hat{\beta}$ seem to conform reasonably well to a Normal or Student-t distribution for all estimators, except possibly the differences regression model with the MA parameter for a sample size of 200 which may have too small a standard deviation and too few values greater than 2.0 in absolute value. This is probably due to the bias in the estimated standard error of $\hat{\beta}$ in this case. Thus, the probability of committing a Type I error in testing hypotheses about β would be similar whether the levels or the differences are used to compute $\hat{\beta}$.

Table 5 contains summary statistics for estimators of the regression coefficient β in the unconditional stochastic regression model, where a new realization of the random walk $\{\tilde{x}_t\}$ is generated for each replication. The sampling distributions of $\hat{\beta}$ are unbiased and symmetric, but the levels estimator and the differences regression estimator which is estimated jointly with the MA parameter have leptokurtic sampling distributions as indicated by large kurtosis and studentized range statistics. Thus, there is an interesting difference between the fixed and stochastic regressor cases.[16] The sampling distribution of $\hat{\beta}$ conditional on the

[14] Even though the disturbances are autocorrelated in the differences regression without the MA parameter, it can be shown that the usual formula for the standard error of a regression coefficient is appropriate because $\Delta\tilde{x}_t$ is serially uncorrelated in this example.

[15] In fact, the average estimated residual variance for the differences regression without the MA parameter is slightly greater than 2.0 for all sample sizes.

[16] Ken Gaver, Martin Geisel, and Harry Roberts have aided our understanding of this phenomenon.

Table 4

Estimators of regression parameter, β: Fixed regressor case.[a]

$$y_t = \hat{\alpha}+\hat{\beta}x_t+\hat{a}_t, \qquad \Delta y_t = \hat{\alpha}+\hat{\beta}\Delta x_t+(1-\theta_1 L)\hat{a}_t.$$

	Levels regression	Differences/MA[b] regression	Differences[c] regression	Levels regression	Differences/MA[b] regression	Differences[c] regression	Levels regression	Differences/MA[b] regression	Differences[c] regression
(1) *Sample moments*		$T = 51$			$T = 101$			$T = 201$	
Mean	1.001	0.9980	1.006	1.000	0.9998	0.9880	1.000	0.9998	0.9993
(Avg. s.e.)	(0.0588)	(0.1296)	(0.2269)	(0.0497)	(0.0639)	(0.1631)	(0.0218)	(0.0330)	(0.1087)
Standard deviation	0.0615	0.1266	0.2271	0.0488	0.0645	0.1599	0.0215	0.0296	0.1059
Skewness	0.0693	−0.0464	−0.0342	−0.0059	−0.0910	−0.0979	−0.0950	0.1205	−0.0233
Excess kurtosis	0.0787	0.0299	−0.0201	−0.1999	−0.2504	−0.2642	0.0724	0.4889	0.0670
Studentized range	6.547	6.301	6.541	5.919	5.945	6.004	6.430	6.783	6.827
(2) *t-ratios* $(\hat{\beta}-1)/\text{s.e.}(\hat{\beta})$									
Mean	0.0142	−0.0155	0.0282	0.0036	0.0025	−0.0726	0.0214	−0.0069	−0.0045
Standard deviation	1.060	0.9787	1.006	0.9905	1.071	0.9855	0.9917	0.8955	0.9764
Percent $t > 2.0$	3.0	1.3	2.6	2.5	2.9	1.2	2.2	1.3	1.6
$t < -2.0$	2.9	2.5	2.8	1.9	3.4	2.8	2.1	1.8	2.2

[a] 1000 replications of samples of size 51, 101 and 201 for eq. (4.2). Regressor $\tilde{x}_t$ follows a random walk, $\tilde{x}_t = \tilde{x}_{t-1}+\tilde{u}_t$ and is the same for each replication, $\tilde{x} = x$.

[b] Regression model in the differences estimated jointly with the MA disturbance process using nonlinear least squares.

[c] Regression model in the differences estimated using linear least squares, ignoring the autocorrelated disturbances.

Table 5

Estimators of regression parameter, β: Stochastic regressor case.[a]

$$y_t = \hat{\alpha}+\hat{\beta}x_t+\hat{a}_t, \qquad \Delta y_t = \hat{\alpha}+\hat{\beta}\Delta x_t+(1-\hat{\theta}_1 L)\hat{a}_t.$$

	Levels regression	Differ-ences/MA[b] regression	Differ-ences[c] regression	Levels regression	Differ-ences/MA[b] regression	Differ-ences[c] regression	Levels regression	Differ-ences/MA[b] regression	Differ-ences[c] regression
(1) *Sample moments*		$T = 51$			$T = 101$			$T = 201$	
Mean	0.9984	0.9998	1.003	0.9993	1.002	0.9948	1.000	1.007	0.9991
(Avg. s.e.)	(0.0597)	(0.0956)	(0.2050)	(0.0304)	(0.0513)	(0.1432)	(0.0152)	(0.0276)	(0.1004)
Standard deviation	0.0631	0.1129	0.2041	0.0306	0.0562	0.1382	0.0165	0.0309	0.0970
Skewness	0.1560	0.0998	0.0006	−0.1055	0.0825	−0.0071	−0.2081	−0.0620	−0.0466
Excess kurtosis	1.045	1.042	0.2718	1.538	0.9550	−0.1021	2.265	0.7497	−0.0997
Studentized range	7.777[d]	8.337[d]	6.862	9.296[d]	8.477[d]	6.856	8.995[d]	6.860	6.008
(2) *t-ratios* $(\hat{\beta}-1)/\text{s.e.}\ (\hat{\beta})$									
Mean	−0.0372	−0.0102	0.0067	−0.0275	0.0426	−0.0322	0.0135	0.0473	−0.0068
Standard deviation	1.023	1.222	0.9988	0.9872	1.195	0.9761	0.9902	1.189	0.9672
Percent $t > 2.0$	2.5	4.3	2.1	2.0	4.8	1.8	2.8	4.4	2.1
$t < -2.0$	3.0	5.2	2.6	2.2	3.3	2.3	2.6	3.6	1.9

[a] 1000 replications of samples of size 51, 101 and 201 for eq. (4.2) Regressor $\tilde{x}_t$ follows a random walk, $\tilde{x}_t = \tilde{x}_{t-1}+\tilde{u}_t$ and a new realization is generated for each replication.

[b] Regression model in the differences estimated jointly with the MA disturbance process using nonlinear least squares.

[c] Regression model in the differences estimated using linear least squares, ignoring the autocorrelated disturbances.

[d] S.R. exceeds the 0.95 fractile of the sampling distribution when sampling from a Normal population.

sample realization $\tilde{x} = x$, $p(\hat{\beta}|x)$, is Normal. However, the sampling variance of $\hat{\beta}$ is proportional to $(x'x)^{-1}$, so it is necessary to integrate the joint density, $p(\hat{\beta}, \tilde{x}) = p(\hat{\beta}|x) \cdot p(\tilde{x})$, with respect to $\tilde{x}$ in order to derive the marginal distribution of $\hat{\beta}$, $p(\hat{\beta})$. For example, if the regressor $\{\tilde{x}_t\}$ is a sequence of i.i.d. Normal variates distributed independently of the regression disturbances (which is the case in the differences regression in our example), the marginal distribution of $\hat{\beta}$, $p(\hat{\beta})$, is Student-t [cf. Zellner (1971, pp. 388–389)]. This is in accord with the results for the differences regressions which ignore the autocorrelated disturbances in table 5. On the other hand, the distribution of $(x'x)^{-1}$ is much more complicated for the levels regression where the regressor follows a random walk. Thus, the fat-tailed sampling distributions for $\hat{\beta}$ from the levels regressions in table 5 are due to the behavior of the regressor sequence $\{\tilde{x}_t\}$. It is not clear why the sampling distribution of $\hat{\beta}$ from the differences/MA regressions are fat-tailed, but this problem seems to decrease as the sample size increases.

Except for the leptokurtosis problem, the sampling distributions of $\hat{\beta}$ in table 5 are very similar to the sampling distributions conditional on $\tilde{x} = x$ in table 4. The means and standard deviations of $\hat{\beta}$ are similar whether the regressor is fixed or random. Again, the levels estimator is more efficient than the differences estimators because sample variation in $\{\tilde{x}_t\}$ is greater than sample variation in $\{\Delta\tilde{x}_t\}$, and the differences estimator with a moving average parameter is more efficient than the differences estimator which ignores the autocorrelated disturbances because the residual variance is smaller if the MA parameter is estimated.

The average standard errors of $\hat{\beta}$ are smaller than the standard deviations of the sampling distributions for all sample sizes for the differences/MA estimator. This causes the standard deviation and the number of extreme values of the t-ratio to be too large for this estimator. The linear least squares estimators for the levels and the differences have t-ratios which conform quite closely to a Student-t distribution with a mean of zero and a variance of unity. This may be somewhat surprising since the standard errors of $\hat{\beta}$ are computed using conventional formulas which assume that the regressor is fixed in repeated samples, not stochastic.

We feel it is important to illustrate the distinction between the fixed and stochastic regressor models through our sampling experiments. Many econometric models contain exogenous variables which are stochastic, so it is necessary to be clear about the distinction between the distribution of $\hat{\beta}$ conditional on $\tilde{x} = x$ (in table 4), and the unconditional distribution $p(\hat{\beta})$ (in table 5). The sampling behavior of the regressor $\{\tilde{x}_t\}$ can influence the form of the unconditional distribution for $\hat{\beta}$. Nevertheless, the distribution of the estimator of the moving average parameter, $\hat{\theta}_1$, for the disturbances of the overdifferenced regression model is not affected by this distinction.

5. Conclusions and suggestions for future research

5.1. Summary of Monte Carlo results

From the point of view of empirical investigators, the question of how to detect overdifferencing if it occurs is very important. One way to approach this issue is to perform tests on the estimated moving average parameter for the disturbances of the differences regression. If the MA parameter is not significantly different from unity one might suspect overdifferencing and choose the specification of the model in the levels of the variables.

Unfortunately, the distribution of estimators of non-invertible MA processes is not known. A moving average process with a unitary root has no autoregressive representation, and such a root is on the boundary of the region of the restricted parameter space ($|\theta_1| \leqq 1$) which is used to identify the parameters of the likelihood function. Therefore, most of the usual properties of likelihood estimators do not apply.

We use a nonlinear least squares (or conditional maximum likelihood) estimator of the MA parameter which is restricted to the region $|\hat{\theta}_1| \leqq 1$. Naturally, this estimator is downward biased and negatively skewed. Standard errors computed using usual techniques understate the sampling variability of the estimator (a fact which has been found by others for invertible MA processes).

Despite these shortcomings, it is important to note that the estimator of the MA parameter for the regression disturbances seems to have the same properties as the estimator for the univariate MA(1) process in the strictly non-invertible case. The presence of the stochastic regressor does not seem to affect the properties of the estimator of the MA parameter. This parallels Pierce's (1971) asymptotic result for invertible MA processes. Thus, even though the *t*-ratio for the MA parameter is shown to depart substantially from a Normal or Student-*t* distribution, the fractiles of the sampling distributions for $\hat{\theta}_1$ and its *t*-ratio which are derived in this paper should be useful to data analysts who are interested in testing whether overdifferencing is a problem.

We have not attempted to derive small sample or asymptotic distributions for $\hat{\theta}_1$ because of the complex analytical problems associated with the likelihood function for the non-invertible MA process. However, we have computed the mean and variance of $T^{\frac{1}{2}}(\hat{\theta}_1 - \theta_1)$ from our Monte Carlo experiments in table 6. Although the mean and variance of $T^{\frac{1}{2}}(\hat{\theta}_1 - \theta_1)$ decrease slightly as the sample size increases, the values in table 6 might be used to extrapolate the results of our sampling experiments to sample sizes we do not consider.

As expected, least squares estimators of regression parameters for overdifferenced models are unbiased and seem to have approximately Normal (or Student-*t*) distributions. If the model is correctly specified in the levels, differencing reduces the efficiency of estimators of the regression parameters since the sample dispersion of the nonstationary regressor is greater in the levels than the differences. However, the efficiency of the differences estimator is substantially

Table 6

Mean and variance of $T^{\frac{1}{2}}(\hat{\theta}_1-\theta_1)$.

	Mean	Variance
(1) *Time trend model*[a]		
$T = 50$	−0.562	0.392
$T = 100$	−0.647	0.372
$T = 200$	−0.696	0.356
(2) *Regression model, x fixed*[b]		
$T = 50$	−0.552	0.373
$T = 100$	−0.597	0.399
$T = 200$	−0.686	0.363
(3) *Regression model,* $\tilde{x}$ *random*[b]		
$T = 50$	−0.495	0.401
$T = 100$	−0.609	0.425
$T = 200$	−0.645	0.351

[a] Derived from table 1.
[b] Derived from table 3.

improved if a moving average process is estimated for the disturbances of the differences regression using nonlinear least squares. For all estimators, a *t*-ratio based on a correctly estimated standard error[17] has a distribution which approximates the Student-*t* distribution, so the probability of committing a Type I error in tests about regression parameters does not seem to be affected by differencing.

Finally, we note that the residual variance (or the standard error of the regression) is not substantially affected by overdifferencing if an MA parameter is estimated for the differences regression. Although the standard error of the overdifferenced regression is higher than the standard error of the correctly specified levels regression (because $\hat{\theta}_1$ is usually less than 1), the difference is small. Thus, the loss in prediction efficiency due to overdifferencing is not substantial. The fractiles of the sampling distribution of the ratio $\hat{\sigma}(a_d)/\hat{\sigma}(a_L)$ could provide another basis for testing whether overdifferencing is a problem, since this ratio does not seem to depend on the form of the regression model.

[17] In general, the formula: $(x'x)^{-1}x'\Sigma x(x'x)^{-1}$ should be used to compute the covariance matrix for the regression parameters, where Σ is the estimated covariance matrix of the regression disturbances. In all of our examples the usual formula: $\hat{\sigma}_a{}^2(x'x)^{-1}$ yields correct standard errors, but that certainly is not correct in general. In fact, the examples of spurious regressions given by Granger and Newbold (1974) illustrate the dangers of using the latter formula when it is not appropriate.

5.2. *Implications for empirical work*

Our experiments indicate that the Box, Jenkins, Granger, Newbold strategy of differencing all variables until the transformed variables have constant unconditional means over the sample period, prior to estimating regression parameters, should not lead to serious errors of inference, even if differencing is not necessary to induce a stationary disturbance term for the regression equation. The only apparent cost of overdifferencing in this situation is reduced efficiency for estimators of the regression parameters due to the reduced sample variability of the regressor.

Of course, the question that must be answered by an empirical investigator is whether it is plausible that non-stationary variables are related through a constant parameter regression relationship with a stationary error term. In some instances, such as the relationship between interest rates and subsequently observed inflation rates analyzed by Fama (1975) or Nelson and Schwert (1977), this may be an appropriate specification of a model. However, in many instances regression equations which relate the levels of non-stationary variables may capture spurious relationships due to deterministic or stochastic trends in all of the variables. This is the phenomenon which is illustrated by Granger and Newbold (1974). We believe that the dangers inherent in such situations are substantially greater than those associated with overdifferencing, as we argue in Plosser and Schwert (1977).

5.3. *Relationships to other work*

The scope of these experiments is necessarily limited; however, the general principles involved should be applicable to more complicated models. For example, if the disturbances of the levels regression follow an autoregressive-moving average process the disturbances of the differences regression follow an ARMA process with a unitary root in the MA polynomial. In this case, a test for overdifferencing could be based on the distribution of the roots of the MA polynomial. We conjecture that such an extension would not involve significant new difficulties.

We have not explicitly considered the effect of overdifferencing on a model where the regressor is measured with error. However, in Plosser and Schwert (1977) we use a textbook example of the errors-in-variables problem to illustrate the effect of overdifferencing. Differencing can cause the inconsistency of the least squares estimator of the regression parameter to be increased if the levels of the regressor are positively autocorrelated. (In this analysis we assume that the variables are stationary in the levels as well as the differences.) On the other hand, the inconsistency is reduced by differencing if the levels of the regressor are negatively autocorrelated. If the measurement errors are autocorrelated the results are even more ambiguous. The important thing to remember is that least squares estimators are inconsistent for *both* the levels regression and the differences regression in this case. The real solution to this problem is not related to differencing, it lies in finding a consistent, efficient estimator for both the levels and the differences.

A related problem treated by Sims (1972) is the effect of differencing on distributed lag models, where rational lag functions are used to approximate infinite distributed lag relationships. We have not considered this problem of misspecification, or any other problems of specification analysis. We always assume that the model is correctly specified as a linear regression model in the levels of the variables.

Finally, other tests of overdifferencing could be analyzed. For example, a Bayesian posterior odds ratio, where the prior distribution on θ_1 is restricted to the range $-1 \leqq \theta_1 \leqq 1$, provides an alternative means of deciding whether overdifferencing has occurred.

References

Aigner, Dennis, 1971, A compendium on estimation of the autoregressive-moving average model, International Economic Review 12, 348–371.

Ali, Mukhtar, 1976, Analysis of ARMA models — Estimation and prediction, unpublished Working Paper (University of Chicago, Chicago, IL).

Anderson, T.W., 1959, On asymptotic distributions of estimates of parameters of stochastic difference equations, Annals of Mathematical Statistics 30, 676–687.

Anderson, T.W., 1971, The statistical analysis of time series (Wiley, New York).

Box, G.E.P. and D.R. Cox, 1964, An analysis of transformations, Journal of the Royal Statistical Society Series B 26, 211–243.

Box, G.E.P. and G.M. Jenkins, 1970, Time series analysis (Holden-Day, San Francisco, CA).

Box, G.E.P. and M.E. Muller, 1958, A note on the generation of random normal deviates. Annals of Mathematical Statistics 29, 610–611.

Chambers, John M., 1973, Fitting nonlinear models: Numerical techniques, Biometrika 60. 1–13.

Chernoff, Herman, 1954, On the distribution of the likelihood ratio, Annals of Mathematical Statistics 25, 573–578.

David, H.A., H.O. Hartley and E.S. Pearson 1954, The distribution of the ratio, in a single normal sample, of range to standard deviation, Biometrika 61, 483–491.

Draper, N. and H. Smith, 1966, Applied regression analysis (Wiley, New York).

Fama, Eugene F., 1975, Short term interest rates as predictors of inflation, American Economic Review 65, 269–282.

Feller, William, 1968, An introduction to probability theory and its applications, vol. I, 3rd ed. (Wiley, New York).

Galbraith, R.F. and J.I. Galbraith, 1974, On the inverses of some patterned matrices arising in the theory of stationary time series, Journal of Applied Probability 11, 63–71.

Gonedes, Nicholas J. and Harry V. Roberts, 1976, Statistical analysis of random walks and near random walks, Center for Mathematical Studies in Business and Economics Report no. 7606 (University of Chicago, Chicago, IL).

Granger, C.W.J. and P. Newbold, 1974, Spurious regressions in econometrics, Journal of Econometrics 2, 111–120.

Hendry, D.F. and P.K. Trivedi, 1972, Maximum likelihood estimation of difference equations with moving average errors: A simulation study, Review of Economic Studies 39, 117–145.

Jacobs, Rodney, 1976, Data errors and data differences, Department of Economics Working Paper no. 82 (University of California, Los Angeles, CA).

Kang, K.M., 1975, A comparison of estimators for moving average processes, unpublished Working Paper (Australian Bureau of Statistics,).

Kendall, Maurice G. and Alan Stuart, 1968, The advanced theory of statistics, vol. 3, 2nd ed. (Hafner, New York).

Kinderman, A.J. and J.G. Ramage, 1976, Computer generation of normal random variables, Journal of the American Statistical Association 71, 893–896.

Nelson, Charles R., 1974, The first-order moving average process, Journal of Econometrics 2, 121–141.

Nelson, Charles R. and G. William Schwert, 1977, On testing the hypothesis that the real rate of interest is constant, American Economic Review 67, June.

Newbold, P., 1973, Bayesian estimation of Box–Jenkins transfer function – noise models, Journal of the Royal Statistical Society Series B 35, 323–336.

Nicholls, D.F., A.R. Pagan and R.D. Terrell, 1975, The estimation and use of models with moving average disturbance terms: A survey, International Economic Review 16, 113–134.

Nicholls, D.F. and A.R. Pagan, 1977, Specification of the disturbance for efficient estimation – An extended analysis, Econometrica 45, 211–217.

Osborn, Denise R., 1976, Maximum likelihood estimation of moving average processes, Annals of Economic and Social Measurement 5, 75–87.

Pesaran, M.H., 1973, Exact maximum likelihood estimation of a regression equation with first-order moving average error, Review of Economic Studies 40, 529–535.

Pierce, David, 1971, Least squares estimation in the regression model with autoregressive–moving average errors, Biometrika 58, 299–312.

Pierce, David, 1975, On trend and autocorrelation, Communications in Statistics 4, 163–175.

Plosser, Charles I. and G. William Schwert, 1977, Money, income, and sunspots: Measuring economic relationships and the effects of differencing, unpublished Working Paper.

Quenouille, M.H., 1968, The analysis of multiple time-series (Hafner, New York).

Rao, J.N.K. and G. Tintner, 1962, The distribution of the ratio of the variances of variate differences in the circular case, Sankhya Series B 24, 385–394.

Shaman, P., 1969, On the inverse of the covariance matrix of a first order moving average, Biometrika 56, 595–600.

Sims, Christopher A., 1972, The role of approximate prior restrictions in distributed lag estimation, Journal of the American Statistical Association 67, 169–175.

Tintner, Gerhard, 1940, The variate difference method, Cowles Commission Monograph no. 5 (Cowles, Bloomington, IN).

Tintner, Gerhard, 1955, The distribution of the variances of variate differences in the circular case, Metron 17, 43–52.

Trivedi, P.K., 1970, Inventory behaviour in U.K. manufacturing, Review of Economic Studies 37, 517–536.

Vinod, Hirshikesh D., 1976, Effect of ARMA errors on the significance tests for regression coefficients, Journal of the American Statistical Association 71, 929–933.

Wecker, W.E., 1974, The prediction of backward and noninvertible time series, unpublished Working Paper (University of Chicago, Chicago, IL).

Whittle, P., 1963, Prediction and regulation (Van Nostrand, Princeton, NJ).

Wichern, D.W., 1973, The behavior of the sample autocorrelation function for an integrated moving average process, Biometrika 60, 235.

Wold, H.O., 1938, A study in the analysis of stationary time series (Almquist and Wicksell, Uppsala).

Yaglom, A.M., 1955, The correlation of processes whose *n*th differences constitute a stationary process, Matem. Sbornik 37, 141–196.

Yule, G. Udny, 1926, Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series, Journal of Royal Statistical Society 89, 1–69.

Zellner, Arnold, 1971, Introduction to bayesian inference in econometrics (Wiley, New York).

Zellner, Arnold and Charles Plosser, 1977, On discriminating between random walk and closely related stochastic processes using bayesian and non-bayesian procedures, unpublished Working Paper (University of Chicago, Chicago, IL).